Journal of Data Science **7**(2009), 497-511

An Application of Bayesian Model Averaging Approach to Traffic Accidents Data Over Hierarchical Log-Linear Models

Haydar Demirhan and Canan Hamurkaroglu
Hacettepe University

Abstract: In this article, a Bayesian model averaging approach for hierarchical log-linear models is considered. Posterior model probabilities are approximately calculated for hierarchical log-linear models. Dimension of interested model space is reduced by using Occam's window and Occam's razor approaches. 2002 road traffic accident data of Turkey is analyzed by using the considered approach.

Key words: Bayesian estimation, Bayesian model averaging, Gibbs sampling, log-linear model, model selection, traffic accidents.

1. Introduction

Many fields of scientific investigation include analysis of qualitative data. It is important to discover association structure of the interested categorical variables. On this point, log-linear models are widely used and very flexible. Bayesian methods give the opportunity of combining sample information with expert information. By this way, all available information is included in the conducted analysis. The association structure can be displayed by a Bayesian approach of log-linear modelling. The Bayesian approach is more advantageous than the classical setting because the inference is exact rather than asymptotic. Bayesian setting gives an entire posterior distribution for each element of the model; however the classical setting yields a point estimate and a precision estimated via an asymptotic method. In addition, the Bayesian approach would give better estimates of variability than the likelihood analysis (Gelfand and Mallick, 1995). All of these advantages are also valid for the Bayesian approaches to the log-linear modelling.

Estimation of the model parameters does not complete the analysis process. Model selection should be considered to obtain the most parsimonious model. When the Bayesian approach is used along with the log-linear modelling, Bayesian

model selection procedures take place instead of the classical ones. General advantages of the Bayesian approach are also valid here.

In this sense, prior information can be induced on the log-linear parameters, expected cell counts or cell probabilities. In each case, induced prior information should be consistent with the other two cases. Posterior inferences are also drawn by using different algorithms for each case. In this article, we will put the prior information on log-linear parameters.

There is a huge literature on the model selection. Most of the methods can be used for large class of models, including log-linear models. Some of the methods are based on a series of significance tests and some include prior information, use Markov chain Monte Carlo (MCMC) methods and some are based on the Bayes factors. Almost each method has its own problems. Most important problem is on the model uncertainty. When a single model is selected and inferences are conditionally based on the selected model, model uncertainty is ignored (Raftery, 1996). This case is especially seen in the classical setting.

This difficulty can be overcome by including the information provided by all suitable models into the analysis process. 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 such that posterior distribution of considered quantity is obtained over the set of suitable models, then they are weighted by their posterior model probabilities (Leamer, 1978; Raftery, 1996). Leamer (1978) extended the model averaging idea of Roberts (1965). However, computational difficulties were handicapped for progress of model averaging idea. Draper (1995) and Raftery (1995) reviewed the Bayesian model averaging (BMA) and the cost of ignoring model uncertainty. Madigan and Raftery (1994) are also considered the BMA that they give Occam's razor and Occam's window approaches to reduce the number of candidate models. Work of Hoeting *et al.* (1999) is a good tutorial for BMA. They discuss implementation of BMA for graphical, regression and generalized linear models, survival analysis, software for BMA, prior model probabilities and predictive performance of BMA, and give several examples.

In this article, we aim to present a BMA approach for model selection in hierarchical log-linear (HLL) models and analyze the road traffic accidents data by using the BMA approach. We use the normal likelihood, which is used by Leighty and Johnson (1990), and normal prior distribution, following to the approach given in Leighty and Johnson (1990) and Demirhan and Hamurkaroglu (2006). We calculate a required integral, which is mentioned in detail in Section 4, to develop a BMA approach for HLL models. Calculation of the integral is one of main difficulties of general BMA approach.

Our approach is useful and more advantageous than classical setting. In

the classical setting, the model that gives the best fit to the data is determined by using a criterion or a model selection procedure, and classical estimates of model parameters are obtained conditionally on the predetermined model. In addition, estimation of another quantity such as quartiles of parameters is not possible in the classical setting. However, one can estimate other quantities and model parameters simultaneously, obtain entire posterior distributions of them in the Bayesian setting. Moreover, model uncertainty is also included in the resulting estimates at the same time. Because of this, standard error estimates of the parameters and the considered quantities will be smaller, and one can draw inferences on the distribution of model parameters. These are significant gains of the BMA approach in general. Specifically, for the road traffic accidents data, we are able to determine the best fitting model more confidentially in the Bayesian setting. We obtain estimates of model parameters with smaller standard errors. This is important, because our inferences are mainly based on these estimates. In addition, we are able to see the posterior distributions of model parameters. Details of the analysis of the data set are given in Section 5.

Section 2 mentions used log-linear model notations and hierarchy principle. BMA approach is outlined in Section 3. Section 4 is on the BMA for HLL models. In Section 5, 2002 road traffic data of Turkey is analyzed by using the given BMA approach.

2. Hierarchical Log-linear Models and Notation

Number of terms of a log-linear model increases by the increase in the number of categorical variables. Then standard notations become cumbersome. Instead, King and Brooks (2001a) give very flexible and practical notations, which are also used in this work.

Set of sources, 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 $S = \{S_\zeta : \zeta = 1, ..., |S|\}$. Set of levels for source S_ζ is K_ζ, for $\zeta = 1, ..., |S|$. Cells of a contingency table can be represented by the set $K = K_1 \times \cdots \times K_{|S|}$, so the cells are indexed by $\boldsymbol{k} \in$K. Expected and observed cell counts are denoted by $\mathrm{n}_{\boldsymbol{k}}$ and $y_{\boldsymbol{k}}$ for $\boldsymbol{k} \in$K, respectively. The set of subsets of S is defined by $\wp(S) = \{s : s \subseteq S\}$. Then $m \subseteq \wp(S)$ 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 $c \in m \subseteq \wp(S)$. Constant term of the log-linear model is represented by $\emptyset \in \wp(S)$. $\boldsymbol{M}^c$ contains all possible combinations of the levels of sources included in c. In general, the highest level is not included by the elements of $\boldsymbol{M}^c$. Thus the set $\boldsymbol{M}^c$ is $\{\boldsymbol{m}_1^c, \ldots, \boldsymbol{m}_{|\boldsymbol{M}^c|}^c\}$. Then the log-linear model vector for each $c \in m \subseteq \wp(S)$ is $(\boldsymbol{\beta}^c)^T = \{\beta_{\boldsymbol{m}_1}^c, \beta_{\boldsymbol{m}_2}^c, \ldots, \beta_{\boldsymbol{m}_{|\boldsymbol{M}^c|}}^c\}$. Thus the log-linear parameter

vector for the model m is $\boldsymbol{\beta}_m = \{(\boldsymbol{\beta}^{c_1})^T, (\boldsymbol{\beta}^{c_2})^T, \ldots, (\boldsymbol{\beta}^{c_{|m|}})^T\}$. Design matrix or model matrix corresponding to the model $m \subseteq \wp(S)$ is denoted by $\boldsymbol{X}_m$. Using the design matrix and the parameter vector, the log-linear model is represented as follows:

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

More detailed notations for the elements of design matrix, order of parameters and cells, and examples are given in King and Brooks (2001a, 2001b).

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, and all of its lower order relatives must be in the model at the same time (Bishop, Fienberg and Holland, 1975, chap. 2). Hierarchy principle helps us to decrease the number of models of interest in model selection.

3. Bayesian Model Averaging

Underestimation due to the model uncertainty can lead very risky false decisions (Hodges, 1987). BMA provides a way to deal with model uncertainty. Because, all elements of the model space $\wp(S)$ is considered for the estimation of a quantity of interest. Here, quantity of interest can be a group of parameters, effect size, odds ratio, etc. Let the quantity of interest be Δ, and D be data then

$$P(\Delta|D) = \sum_{m \in \wp(S)} P(\Delta|m, D) P(m|D). \tag{3.1}$$

In (3.1), posterior distribution of the quantity of interest, under each model in $\wp(S)$, is weighted by the posterior model probability of corresponding model. Posterior model probability of m is obtained as follows:

$$P(m|D) = \frac{P(D|m)P(m)}{\sum_{m' \in \wp(S)} P(D|m')P(m')},$$

where $P(m)$ is the prior model probability, and $P(D|m)$ is the likelihood under model m and obtained as follows:

$$P(D|m) = \int P(D|\boldsymbol{\beta}_m, m) P(\boldsymbol{\beta}_m|m) d\boldsymbol{\beta}_m. \tag{3.2}$$

Posterior mean and variance of Δ are obtained by averaging as follows:

$$E(\Delta|D) = \sum_{m \in \wp(S)} \hat{\Delta}_m P(m|D), \tag{3.3}$$

$$V(\Delta|D) = \sum_{m \in \wp(S)} (V(\Delta|D, m) + \hat{\Delta}_m^2) P(m|D) - E(\Delta|D)^2, \tag{3.4}$$

where $\hat{\Delta}_m = E(\Delta|D, m)$ (Hoeting *et al.*, 1999).

Although BMA is seen as a solution to the model uncertainty problem, it is not used as a standard analysis toolkit. Because it has some difficulties:

1. The integral in (3.2) is in general hard to compute.
2. Dimension of the model space can be enormous, which prevents considering whole model space.
3. Specification of $P(m)$ is not clear, especially if there is prior information on some of the models of $\wp(S)$.
4. Let $\Im(S) \subseteq \wp(S)$ be the model class over which to average, then choosing $\Im(S)$ is also problematic (Madigan and Raftery, 1994; Draper, 1995; Hoeting *et al.*, 1999).

To relax the effect of the difficulty mentioned in 1, dimension of the considered model space can be reduced. Madigan and Raftery (1994) proposed an approach that is working on a subset of the whole model space, namely $\Im(S) \subseteq \wp(S)$. $\Im(S)$ includes models that do not predict data far less well than the best model (Madigan and Raftery, 1994). We should find a subset of $\wp(S)$, over which to apply (3.1). Let $\Im'(S)$ be the mentioned subset of $\wp(S)$, and it is defined as follows:

$$\Im'(S) = \left\{ m : \frac{\max[P(m'|D)]}{P(m|D)} \leq C \right\},$$

where $m \in \wp(S)$, m' is the model that has the maximum posterior model probability and C is an arbitrary constant. Models not belonging to $\Im'(S)$ are excluded from $\wp(S)$. This is the first principle of Madigan and Raftery (1994), called Occam's window. According to their second principle, which is called as Occam's Razor, complex models receiving less support from the data than their counterparts are excluded from $\Im'(S)$ by the following equation:

$$\Re(S) = \left\{ m : m' \in \Im'(S), m' \subset m, \frac{P(m'|D)}{P(m|D)} > 1 \right\}. \tag{3.5}$$

If a model m has a simpler sub-model m' and the posterior model probability of the sub-model is higher then the considered model is excluded from $\Im'(S)$ by the equation (3.5). Finally, $\wp(S)$s seen in (3.1), (3.3) and (3.4) are replaced by $\Im(S) = \Im'(S) \setminus \Re(S)$.

4. BMA for HLL Models

One of the difficulties of the BMA, which is also valid for HLL models, is the computation of the integral (3.2). We calculated the required integral.

When (3.2) is rewritten as follows

$$P(D|m) = \int_{R_{\boldsymbol{\beta}_m}} P(D, \boldsymbol{\beta}_m|m) d\boldsymbol{\beta}_m = \int_{R_{\boldsymbol{\beta}_m}} P(D|\boldsymbol{\beta}_m, m) P(\boldsymbol{\beta}_m|m) d\boldsymbol{\beta}_m, \quad (4.1)$$

computation of the integral becomes simpler. In (4.1), $P(D|\boldsymbol{\beta}_m, m)$ is the likelihood of data, and $P(\boldsymbol{\beta}_m|m)$ is the prior distribution of log-linear parameter vector under the model m.

Leighty and Johnson (1990) used

$$P(D|\boldsymbol{\beta}_m, m) \approx P(\boldsymbol{b}_m|\boldsymbol{\beta}_m, m) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{b}_m - \boldsymbol{\beta}_m)^T \boldsymbol{V}_{\boldsymbol{b}_m}^{-1}(\boldsymbol{b}_m - \boldsymbol{\beta}_m)\right\}$$

to approximate the likelihood of data. Here $\boldsymbol{b}_m$ is maximum likelihood estimate (MLE) of $\boldsymbol{\beta}_m$ and $\boldsymbol{V}_{\boldsymbol{b}_m}$ is the corresponding covariance matrix. Approach of Leighty and Johnson (1990) is followed for the specification of $P(\boldsymbol{\beta}_m|m)$. (4.1) is rewritten as follows:

$$\begin{aligned} P(D|m) &= \int_{R_{\boldsymbol{\beta}_m}} P(\boldsymbol{b}_m|\boldsymbol{\beta}_m, m) P(\boldsymbol{\beta}_m|m) d\boldsymbol{\beta}_m \\ &= \int_{R_{\boldsymbol{\beta}_m}} \frac{1}{(2\pi)^{p/2}\det(\boldsymbol{V}_{\boldsymbol{b}_m})^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{b}_m - \boldsymbol{\beta}_m)^T \boldsymbol{V}_{\boldsymbol{b}_m}^{-1}(\boldsymbol{b}_m - \boldsymbol{\beta}_m)\right\} \\ &\quad \times \frac{1}{(2\pi)^{p/2}\det(\boldsymbol{\Sigma}_m)^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_m - \boldsymbol{\mu}_m)^T \boldsymbol{\Sigma}_m^{-1}(\boldsymbol{\beta}_m - \boldsymbol{\mu}_m)\right\} d\boldsymbol{\beta}_m, \end{aligned}$$

where $\det(\cdot)$ denotes the determinant of inner matrix, p is the dimension of $\boldsymbol{\beta}_m$, $\boldsymbol{V}_{\boldsymbol{b}_m}^{-1}$ 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. After some algebra, it is obtained that

$$\begin{aligned} P(D|m) = \int_{R_{\boldsymbol{\beta}_m}} \frac{\det(\boldsymbol{\Sigma}_m)^{-1/2}}{(2\pi)^p \det(\boldsymbol{V}_{\boldsymbol{b}_m})^{1/2}} \exp\Big\{ &-\frac{1}{2}\big[\boldsymbol{\beta}_m^T(\boldsymbol{V}_{\boldsymbol{b}_m}^{-1} + \boldsymbol{\Sigma}_m^{-1})\boldsymbol{\beta}_m \\ &-2\boldsymbol{\beta}_m^T(\boldsymbol{V}_{\boldsymbol{b}_m}^{-1}\boldsymbol{b}_m + \boldsymbol{\Sigma}_m^{-1}\boldsymbol{\mu}_m)\big]\Big\} \\ \times \exp\Big\{ &-\tfrac{1}{2}\big[\boldsymbol{b}_m^T \boldsymbol{V}_{\boldsymbol{b}_m}^{-1}\boldsymbol{b}_m + \boldsymbol{\mu}_m^T \boldsymbol{\Sigma}_m^{-1}\boldsymbol{\mu}_m\big]\Big\} d\boldsymbol{\beta}_m. \end{aligned} \quad (4.2)$$

Let $\boldsymbol{A} = [\boldsymbol{V}_{\boldsymbol{b}_m}^{-1} + \boldsymbol{\Sigma}_m^{-1}]$ and $\boldsymbol{Az} = [\boldsymbol{V}_{\boldsymbol{b}_m}^{-1}\boldsymbol{b}_m + \boldsymbol{\Sigma}_m^{-1}\boldsymbol{\mu}_m]$, then it is obtained from

(4.2) that

$$P(D|m) = \frac{\det(\boldsymbol{A}^{-1})^{1/2}(2\pi)^{-p/2}}{\det(\boldsymbol{V}_{\boldsymbol{b}_m})^{1/2}\det(\boldsymbol{\Sigma}_m)^{1/2}} \exp\left\{-\frac{1}{2}\left[\boldsymbol{b}_m^T\boldsymbol{V}_{\boldsymbol{b}_m}^{-1}\boldsymbol{b}_m + \boldsymbol{\mu}_m^T\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\mu}_m - \boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z}\right]\right\} \times \int_{R_{\boldsymbol{\beta}_m}} \frac{1}{(2\pi)^{p/2}\det(\boldsymbol{A}^{-1})^{1/2}} \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}_m^T\boldsymbol{A}\boldsymbol{\beta}_m - 2\boldsymbol{\beta}_m^T\boldsymbol{A}\boldsymbol{z} + \boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z}\right]\right\} d\boldsymbol{\beta}_m. \tag{4.3}$$

Integral part of (4.3) is the integral of multivariate normally distributed random vector $\boldsymbol{\beta}_m$ with mean vector $\boldsymbol{z}$ and covariance matrix $\boldsymbol{A}$ and equals to 1. Thus, $P(D|m)$ is obtained as follows:

$$P(D|m) = h \cdot \exp\left\{-\tfrac{1}{2}\left[\boldsymbol{b}_m^T\boldsymbol{V}_{\boldsymbol{b}_m}^{-1}\boldsymbol{b}_m + \boldsymbol{\mu}_m^T\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\mu}_m - \boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z}\right]\right\}, \tag{4.4}$$

where $\boldsymbol{z} = [\boldsymbol{V}_{\boldsymbol{b}_m}^{-1} + \boldsymbol{\Sigma}_m^{-1}]^{-1}[\boldsymbol{V}_{\boldsymbol{b}_m}^{-1}\boldsymbol{b}_m + \boldsymbol{\Sigma}_m^{-1}\boldsymbol{\mu}_m]$ and

$$h = \frac{\det(\boldsymbol{A}^{-1})^{1/2}(2\pi)^{-p/2}}{\det(\boldsymbol{V}_{\boldsymbol{b}_m})^{1/2}\det(\boldsymbol{\Sigma}_m)^{1/2}}.$$

The approach of Leighty and Johnson (1990) is used to represent degree of belief in the prior information. In the approach, 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, $\boldsymbol{\Sigma}_m = \alpha\boldsymbol{C}_m = \alpha c\boldsymbol{I}_m$, where $\boldsymbol{I}_m$ is the identity matrix dimension of p, and $c = p/tr(\boldsymbol{V}_{\boldsymbol{b}_m}^{-1})$ (Leighty and Johnson, 1990). Distribution of the general precision parameter α is given by the second stage prior. It is taken that $\tau = 1/(1+\alpha)$ and $\tau \sim uniform(0,1)$ to make calculations easier. Values of τ represent the degree of our belief in prior. Leonard (1975) and Leighty and Johnson (1990) state that close to zero values of this precision parameter represent disbelief. When these definitions are applied to h, the following is obtained:

$$h = \frac{\det\left(\left[\boldsymbol{V}_{\boldsymbol{b}_m}^{-1} + (\alpha c)^{-1}\boldsymbol{I}_m\right]^{-1}\right)^{1/2}(2\pi)^{-p/2}}{\det(\boldsymbol{V}_{\boldsymbol{b}_m})^{1/2}\det(\alpha c\boldsymbol{I}_m)^{1/2}}.$$

If a noninformative prior distribution is defined, $\tau \to 0$. This implies that $\alpha \to \infty$, and hence $\lim_{\alpha\to\infty} h = 0$, then $P(D|m,\tau) \simeq 0, \forall m \in \Im(S)$. When equation (4.4) is used to find $P(D|m,\tau)$, models that have greater dimensions are penalized by the term $(2\pi)^{p/2}$. As the result of this penalization, simpler models have greater model probability than complex models, even if they fit the data better. For these reasons, it is more appropriate using the result of (4.4) up to a proportionality constant such that

$$P(D|m,\tau) \propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{b}_m^T\boldsymbol{V}_{\boldsymbol{b}_m}^{-1}\boldsymbol{b}_m + \boldsymbol{\mu}_m^T[(1-\tau)/\tau]\boldsymbol{C}_m^{-1}\boldsymbol{\mu}_m - \boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z}\right]\right\}. \tag{4.5}$$

After the integration problem, we deal with the huge number of HLL models. Number of log-linear models grows rapidly by the increase in the number of considered categorical variables. Although this is slower for HLL models, the problem still remains.

Occam's Window approach of Madigan and Raftery (1994) provides a solution to this problem. We adapted the approach for HLL models. Let $\Omega(S)$ and $\Upsilon(S)$ be subsets of $\wp(S)$, set $\Omega(S) = \emptyset$ and $\Upsilon(S)$ is the set of starting models. Then following down algorithm and up algorithm are applied successively.

Down Algorithm

1. Select a model m from $\Upsilon(S)$.
2. $\Upsilon(S) \leftarrow \Upsilon(S) \setminus \{m\}$ and $\Omega(S) \leftarrow \Omega(S) \cup \{m\}$.
3. Select a hierarchical submodel m_0 of m such that $dim(\boldsymbol{\beta}_{m_0}) = dim(\boldsymbol{\beta}_m) - 1$, where $dim(\cdot)$ denotes the dimension of the inner vector.
4. Compute $B = \log[P(m_0|D)/P(m|D)]$.
5. If $B > O_R$, then $\Omega(S) \leftarrow \Omega(S) \setminus \{m\}$ and if $m_0 \notin \Upsilon(S)$, then $\Upsilon(S) \leftarrow \Upsilon(S) \cup \{m_0\}$.
6. If $O_L \leq B \leq O_R$, then if $m_0 \notin \Upsilon(S)$, $\Upsilon(S) \leftarrow \Upsilon(S) \cup \{m_0\}$.
7. If there are more submodels of m then go to 3.
8. If $\Upsilon(S) = \emptyset$ then go to 1.

Up Algorithm

1. Select a model m from $\Upsilon(S)$.
2. $\Upsilon(S) \leftarrow \Upsilon(S) \setminus \{m\}$ and $\Omega(S) \leftarrow \Omega(S) \cup \{m\}$.
3. Select a hierarchical supermodel m_1 of m such that $dim(\boldsymbol{\beta}_{m_1}) = dim(\boldsymbol{\beta}_m) + 1$.
4. Compute $B = \log[P(m|D)/P(m_1|D)]$.
5. If $B < O_L$, then $\Omega(S) \leftarrow \Omega(S) \setminus \{m\}$; if $m_1 \notin \Upsilon(S)$, then $\Upsilon(S) \leftarrow \Upsilon(S) \cup \{m_1\}$.
6. If $O_L \leq B \leq O_R$, then if $m_1 \notin \Upsilon(S)$, $\Upsilon(S) \leftarrow \Upsilon(S) \cup \{m_1\}$.
7. If there are more supermodels of m then go to 3.
8. If $\Upsilon(S) = \emptyset$ then go to 1.

Here, choice of O_L and O_R is based on the considered data set. Raftery, Madigan and Volinsky (1994) suggest taking 1/20 and 20 for O_L and O_R, respectively.

After the application of above algorithms, (3.5) is applied by replacing 1 with $\exp(O_R)$. In addition, models satisfying

$$\frac{\max\{P(m'|D)\}}{P(m|D)} > \exp(-O_L), \tag{4.6}$$

are excluded (Madigan and Raftery, 1994). After all, $\Omega(S)$ contains acceptable models, and number of the possible models is noticeably reduced. Then the set of models, over which to average, is $\Im(S) = \Omega(S)$.

$P(\boldsymbol{\beta}_m|m, D, \tau)$ and posterior estimates of the log-linear parameters, given the data and model, should be obtained to complete the BMA procedure for HLL models. Gibbs sampling is employed for this purpose. Implementation of Gibbs sampling with these prior and likelihood settings requires finding full conditional distribution of each log-linear parameter given the other parameters, model, the data and τ. These full conditionals are derived in Demirhan and Hamurkaroglu (2006) and will also be used here.

5. Analysis of Road Traffic Accidents Data

Road traffic accidents of Turkey in 2002 are taken into consideration. Considered factors are overlay type of the road (OT), place of the accident (AP), state of the included people (SP), and the result of the accident (RA). Considered overlay types are concrete, asphalt, parquet, stabilized and dirt. Considered places of accident are city and inter-city roads. Included people are classified as driver, passenger and pedestrian; and results of an accident are taken as killed or injured. The data set is recorded by The Department of Traffic Training and Research of the General Directorate of Security Affairs of Turkey in 2002 and taken from Traffic Statistics Annual - 2002 [1].

Following the notation given in Section 2, $|S| = 4$, S_1 is OT, S_2 is AP, S_3 is SP and S_4 is RA, then $K_1 = \{1, 2, 3, 4, 5\}$, $K_2 = K_4 = \{1, 2\}$, $K_3 = \{1, 2, 3\}$,

$$\begin{aligned} K = \{&(1,1,1,1),(2,1,1,1),(3,1,1,1),(4,1,1,1),(5,1,1,1),(1,2,1,1),(2,2,1,1),\\ &(3,2,1,1),(4,2,1,1),(5,2,1,1),(1,1,2,1),(2,1,2,1),(3,1,2,1),(4,1,2,1),\\ &(5,1,2,1),(1,2,2,1),(2,2,2,1),(3,2,2,1),(4,2,2,1),(5,2,2,1),(1,1,3,1),\\ &(2,1,3,1),(3,1,3,1),(4,1,3,1),(5,1,3,1),(1,2,3,1),(2,2,3,1),(3,2,3,1),\\ &(4,2,3,1),(5,2,3,1),(1,1,1,2),(2,1,1,2),(3,1,1,2),(4,1,1,2),(5,1,1,2),\\ &(1,2,1,2),(2,2,1,2),(3,2,1,2),(4,2,1,2),(5,2,1,2),(1,1,2,2),(2,1,2,2),\\ &(3,1,2,2),(4,1,2,2),(5,1,2,2),(1,2,2,2),(2,2,2,2),(3,2,2,2),(4,2,2,2),\\ &(5,2,2,2),(1,1,3,2),(2,1,3,2),(3,1,3,2),(4,1,3,2),(5,1,3,2),(1,2,3,2),\\ &(2,2,3,2),(3,2,3,2),(4,2,3,2),(5,2,3,2)\}, \end{aligned}$$

and

$$\begin{aligned} \wp(S) = \{&\emptyset, \{S_1\}, \{S_2\}, \{S_3\}, \{S_4\}, \{S_1, S_2\}, \{S_1, S_3\}, \{S_1, S_4\}, \{S_2, S_3\},\\ &\{S_2, S_4\}, \{S_3, S_4\}, \{S_1, S_2, S_3\}, \{S_1, S_2, S_4\}, \{S_1, S_3, S_4\}, \{S_2, S_3, S_4\},\\ &\{S_1, S_2, S_3, S_4\}\}. \end{aligned}$$

Log-linear parameter vector of the saturated model is $\boldsymbol{\beta}_m = \left(\emptyset, (\boldsymbol{\beta}^{c_1})^T, \ldots, (\boldsymbol{\beta}^{c_{14}})^T\right)^T$.

After these definitions, BMA is applied over the given down and up algorithms. For these algorithms O_R and O_L are taken as 20 and 0.00001, respectively. The Newton-Raphson algorithm is used to obtain $\boldsymbol{V}_{\boldsymbol{b}_m}$ and $\boldsymbol{b}_m$. $P(m|D, \tau)$ is obtained by using (4.5) and $P(m)$.

[1] http://www.egm.gov.tr/teadb/02yillik/02fihrist.htm, p. 47.

Dimension of model space is 168. Prior model probabilities were equal to 0.00479 for the models other than

$$m' = \{\emptyset, \{S_1\}, \{S_2\}, \{S_3\}, \{S_4\}, \{S_1, S_2\}, \{S_1, S_3\}, \{S_1, S_4\}, \{S_2, S_3\}, \{S_2, S_4\}, \{S_3, S_4\}, \{S_1, S_2, S_4\}, \{S_1, S_3, S_4\}, \{S_2, S_3, S_4\}\},$$

and it was 0.2 for m'. The reason of this choice is the prior information that the result of an accident should be related with OT, AP and SP; however state of the person does not have an association with overlay type or place of accident. Prior distribution of the log-linear parameter vector is defined by using the approach of Leigthy and Johnson (1990), τ of which is taken as 10^{-7}. This setting induces a diffuse prior on the log-linear parameters. Gibbs sampling is employed to obtain posterior estimate of each model parameter ($\widetilde{\beta}^c_{\boldsymbol{m}}$) for a given model. Total number of iterations were 30000, 10000 of which were discarded as burn-in. A record has been made at the end of each 200 cycles to reduce the autocorrelation of the Gibbs sequence. Full conditional distributions given in Demirhan and Hamurkaroglu (2006) are used in the mentioned Gibbs sampling scheme.

At the end of the down and up algorithms, elements of $\Im'(S)$ and corresponding posterior model probabilities were obtained as given in Table 1.

Table 1: Elements of $\Im'(S)$ and corresponding posterior model probabilities.

Model	Posterior model probability
$m_1 = \{\emptyset, \{S_1\}, \{S_2\}, \{S_3\}, \{S_4\}, \{S_1, S_2\}, \{S_1, S_3\}, \{S_1, S_4\}, \{S_2, S_3\}, \{S_2, S_4\}, \{S_3, S_4\}, \{S_1, S_2, S_3\}, \{S_1, S_2, S_4\}, \{S_1, S_3, S_4\}, \{S_2, S_3, S_4\}\}$	0.02298
$m_2 = \{\emptyset, \{S_1\}, \{S_2\}, \{S_3\}, \{S_4\}, \{S_1, S_2\}, \{S_1, S_3\}, \{S_1, S_4\}, \{S_2, S_3\}, \{S_2, S_4\}, \{S_3, S_4\}, \{S_1, S_2, S_4\}, \{S_1, S_3, S_4\}, \{S_2, S_3, S_4\}\}$	0.95399
$m_3 = \{\emptyset, \{S_1\}, \{S_2\}, \{S_3\}, \{S_4\}, \{S_1, S_2\}, \{S_1, S_3\}, \{S_1, S_4\}, \{S_2, S_3\}, \{S_2, S_4\}, \{S_3, S_4\}, \{S_1, S_2, S_3\}, \{S_1, S_2, S_4\}, \{S_1, S_3, S_4\}\}$	0.02303

As seen in the Table 1, there are three models remained as possibly acceptable models in $\Im'(S)$. Therefore it is not necessary to go on with the application of (3.5), and $\Im(S) = \Im'(S)$.

After the determination of $\Im(S)$, Δ of (3.1) is taken as each element of $\boldsymbol{\beta}_{m_i}$, $i = 1, 2, 3$; and $\widetilde{\beta}^c_{\boldsymbol{m}}$, $c \in m_i$, $\boldsymbol{m} \in \boldsymbol{M}^c$, are obtained. Also, it is taken as 5, 25, 50, 75 and 95th percentiles of the distribution of each parameter of each m_i. Results, obtained over the Gibbs sampling, are combined to reach BMA estimates by using (3.1), and presented in Table 2. MLEs of parameters of m_2 ($\widehat{\boldsymbol{\beta}}_{m_2}$) and their standard errors are also presented in the Table 2 to see what we gain over the classical setting by using the presented approach.

Table 2: Bayesian and classical parameter estimates

	Percentiles						St.Dev.		St.Dev.
Par.	5%	25%	50%	75%	95%	$\widetilde{\boldsymbol{\beta}}_m$	$\widetilde{\boldsymbol{\beta}}_m$	$\widehat{\boldsymbol{\beta}}_{m_2}$	$\widehat{\boldsymbol{\beta}}_{m_2}$
$\emptyset$	3.4973	3.4974	3.4974	3.4975	3.4980	3.4976	0.0002	-0.9463	0.3062
$\beta_1^{c_1}$	-0.2802	-0.2777	-0.2769	-0.2761	-0.2745	-0.2772	0.0018	2.8302	0.3177
$\beta_2^{c_1}$	4.1340	4.1345	4.1346	4.1346	4.1347	4.1344	0.0002	7.6170	0.3043
$\beta_3^{c_1}$	-1.3599	-1.3585	-1.3578	-1.3571	-1.3555	-1.3578	0.0012	1.1406	0.3601
$\beta_4^{c_1}$	-0.5068	-0.5049	-0.5040	-0.5029	-0.5008	-0.5038	0.0018	1.8216	0.3308
$\beta_1^{c_2}$	0.9187	0.9192	0.9193	0.9193	0.9194	0.9192	0.0002	3.8429	0.2036
$\beta_1^{c_3}$	0.2790	0.2791	0.2792	0.2792	0.2793	0.2791	0.0001	3.3685	0.2635
$\beta_2^{c_3}$	0.4029	0.4029	0.4030	0.4030	0.4033	0.4030	0.0001	3.9340	0.2641
$\beta_1^{c_4}$	-1.5349	-1.5348	-1.5347	-1.5347	-1.5342	-1.5346	0.0002	0.1214	0.9417
$\beta_{11}^{c_5}$	-0.1595	-0.1591	-0.1588	-0.1583	-0.1574	-0.1587	0.0007	-0.2612	0.2113
$\beta_{21}^{c_5}$	-0.6635	-0.6635	-0.6634	-0.6634	-0.6630	-0.6633	0.0002	-0.9886	0.2004
$\beta_{31}^{c_5}$	0.7173	0.7182	0.7187	0.7192	0.7200	0.7186	0.0008	2.2007	0.2748
$\beta_{41}^{c_5}$	-0.0839	-0.0832	-0.0827	-0.0823	-0.0812	-0.0829	0.0007	-0.2802	0.2141
$\beta_{11}^{c_6}$	-0.1744	-0.1725	-0.1715	-0.1707	-0.1689	-0.1717	0.0017	-0.6463	0.2723
$\beta_{12}^{c_6}$	0.2068	0.2076	0.2079	0.2085	0.2095	0.2079	0.0008	-0.5852	0.2728
$\beta_{21}^{c_6}$	0.0087	0.0087	0.0088	0.0088	0.0090	0.0088	0.0001	-0.8078	0.2611
$\beta_{22}^{c_6}$	0.0739	0.0742	0.0742	0.0743	0.0743	0.0742	0.0001	-0.6799	0.2618
$\beta_{31}^{c_6}$	0.1742	0.1760	0.1768	0.1774	0.1791	0.1766	0.0014	-0.9758	0.2673
$\beta_{32}^{c_6}$	-0.3719	-0.3708	-0.3700	-0.3692	-0.3682	-0.3701	0.0012	-1.4546	0.2698
$\beta_{41}^{c_6}$	-0.0927	-0.0909	-0.0901	-0.0891	-0.0878	-0.0903	0.0015	0.0259	0.2870
$\beta_{42}^{c_6}$	0.2290	0.2301	0.2307	0.2316	0.2331	0.2305	0.0013	0.2851	0.2867
$\beta_{11}^{c_7}$	-0.1459	-0.1435	-0.1426	-0.1416	-0.1400	-0.1429	0.0018	-1.7472	1.0467
$\beta_{21}^{c_7}$	-0.0656	-0.0651	-0.0650	-0.0649	-0.0649	-0.0651	0.0002	-1.5588	0.9403
$\beta_{31}^{c_7}$	-0.3627	-0.3613	-0.3607	-0.3598	-0.3581	-0.3607	0.0014	-1.8043	1.3481
$\beta_{41}^{c_7}$	0.1371	0.1391	0.1400	0.1409	0.1424	0.1397	0.0016	-0.5059	1.0447
$\beta_{11}^{c_8}$	-0.2032	-0.2031	-0.2031	-0.2031	-0.2030	-0.2031	0.0001	-2.1105	0.0383
$\beta_{12}^{c_8}$	-0.4582	-0.4581	-0.4581	-0.4581	-0.4581	-0.4581	0.0000	-2.7171	0.0376
$\beta_{11}^{c_9}$	-0.2640	-0.2635	-0.2635	-0.2635	-0.2634	-0.2636	0.0002	-0.6145	0.8970
$\beta_{11}^{c_{10}}$	-0.2440	-0.2438	-0.2438	-0.2437	-0.2436	-0.2438	0.0001	-3.0593	0.6286
$\beta_{21}^{c_{10}}$	-0.3785	-0.3784	-0.3784	-0.3784	-0.3781	-0.3783	0.0001	-3.8756	0.7633
$\beta_{111}^{c_{11}}$	0.0017	0.0018	0.0018	0.0018	0.0019	0.0018	0.0007		
$\beta_{112}^{c_{11}}$	0.0045	0.0045	0.0045	0.0045	0.0046	0.0045	0.0002		
$\beta_{211}^{c_{11}}$	-0.0022	-0.0022	-0.0022	-0.0022	-0.0022	-0.0022	0.0008		
$\beta_{212}^{c_{11}}$	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001	0.0007		
$\beta_{311}^{c_{11}}$	0.0063	0.0063	0.0064	0.0064	0.0065	0.0064	0.0017		
$\beta_{312}^{c_{11}}$	-0.0072	-0.0071	-0.0071	-0.0071	-0.0070	-0.0071	0.0008		
$\beta_{411}^{c_{11}}$	-0.0052	-0.0051	-0.0051	-0.0050	-0.0049	-0.0051	0.0002		
$\beta_{412}^{c_{11}}$	-0.0014	-0.0013	-0.0013	-0.0012	-0.0012	-0.0013	0.0002		
$\beta_{111}^{c_{12}}$	0.0411	0.0416	0.0419	0.0424	0.0432	0.0420	0.0012	-0.8546	0.9784
$\beta_{211}^{c_{12}}$	-0.1030	-0.1029	-0.1029	-0.1028	-0.1024	-0.1028	0.0011	-1.4291	0.8948
$\beta_{311}^{c_{12}}$	-0.3175	-0.3168	-0.3162	-0.3156	-0.3148	-0.3162	0.0013	-2.2776	1.2704
$\beta_{411}^{c_{12}}$	0.1209	0.1214	0.1220	0.1224	0.1234	0.1219	0.0012	-0.5284	0.9870
$\beta_{111}^{c_{13}}$	-0.0949	-0.0932	-0.0923	-0.0914	-0.0892	-0.0925	0.0001	1.0136	0.7781
$\beta_{121}^{c_{13}}$	0.2212	0.2221	0.2226	0.2230	0.2243	0.2225	0.0001	2.0866	0.8614
$\beta_{211}^{c_{13}}$	0.1601	0.1601	0.1602	0.1602	0.1604	0.1602	0.0000	1.7722	0.6231
$\beta_{221}^{c_{13}}$	0.0937	0.0940	0.0940	0.0940	0.0941	0.0939	0.0000	2.0850	0.7592
$\beta_{311}^{c_{13}}$	0.1880	0.1896	0.1902	0.1908	0.1922	0.1900	0.0001	1.9065	0.8712
$\beta_{321}^{c_{13}}$	0.1018	0.1030	0.1037	0.1044	0.1057	0.1036	0.0001	2.1805	1.0850
$\beta_{411}^{c_{13}}$	-0.1774	-0.1758	-0.1752	-0.1741	-0.1732	-0.1754	0.0001	0.0073	0.7514
$\beta_{421}^{c_{13}}$	-0.1184	-0.1169	-0.1163	-0.1157	-0.1143	-0.1165	0.0000	0.5684	0.8546
$\beta_{111}^{c_{14}}$	0.0450	0.0450	0.0450	0.0451	0.0451	0.0450	0.0000	0.7595	0.1139
$\beta_{121}^{c_{14}}$	0.0954	0.0954	0.0954	0.0955	0.0955	0.0954	0.0000	0.9659	0.1149

As seen in the Table 1, the model m_2, which is a conditional association model, has the greatest posterior model probability. Thus it is the most appropriate model for the data. According to m_2, two-way interactions between overlay type and place of accident, overlay type and state of people, and place of accident and state of people are independent given the result of accident. In addition, associations between the levels of overlay type and place of accident, overlay type and state of people, and place of accident and state of people are homogeneous given the result of an accident. In the classical approach, m_2 appears as the best fitting model, but there are two more models (m_1 and the model including $\{S_1, S_2, S_3\}, \{S_1, S_3, S_4\}, \{S_2, S_3, S_4\}$ interactions) that fit well also. When our approach is compared to the classical setting, we can more confidentially conclude that m_2 gives the best fit and there is no other candidate model, because of the very high posterior model probability.

It is drawn from the Table 2 that all posterior distributions of the log-linear parameters, reached using BMA, are narrow, and approximately symmetric. This inference cannot be drawn from the classical setting. The gain of drawing it is that if posterior distribution of one of the model parameters is skewed then we may decide to use the median of the resulting Gibbs sequences as the posterior estimate instead of the mean. Although $\{S_1, S_2, S_3\}$ is not included in the model m_2, associated parameters are given in the Table 2. Due to the low posterior model probabilities of the models m_1 and m_3, estimates of the elements of $\boldsymbol{\beta}^{\mathbf{c_{11}}}$ are close to zero. Contrary to the classical setting, standard error estimates of the model parameters are very smaller in the Bayesian setting. Because, inclusion of the model uncertainty decreases the uncertainty on the model parameters and hence the standard errors. Comparison of values of the parameter estimates of the Bayesian and classical settings is not appropriate. But similar inferences should be drawn in both settings when a diffuse prior distribution is used as in our case. Inferences drawn from the Bayesian and classical estimates are different from each other. The reason of this is also the inclusion of the model uncertainty in the posterior estimates. Consequently, BMA approach to the road accidents data set is more reliable and provide better estimates and inferences than its classical counterparts.

Following inferences are drawn from the Bayesian estimates of the model parameters. Weak negative association exists between asphalt roads and being killed in an accident for all places and all levels of SP. There is positive association between having an accident on inter-city roads and being killed in an accident for all overlay types and all levels of SP. Association of being driver and being killed in an accident for all overlay types and all places is negative. Positive association exists between asphalt roads, inter-city roads and being killed in an accident for all the levels of SP. Positive association exists between asphalt roads, being driver

and being killed in an accident for both city and inter-city roads. There is weak negative association between inter-city roads, being driver and being killed in an accident for all of the overlay types.

All calculations, required for the application, were done on a computer program that is written by the authors in the Delphi 6 application development environment.

5. Disscussion

When HLL models are used to discover association structure of categorical variables, Bayesian model averaging provides a suitable way. General difficulties of it are also valid for HLL models. For HLL models, number of considered models are reduced by the algorithm given by Madigan and Raftery (1994), and the integral used to find posterior model probabilities is obtained up to a proportionality constant. Although, an approximation is used to find posterior model probabilities, reasonable results were obtained in the application. Approach of Leighty and Johnson (1990) is used to obtain likelihood function and to determine the prior distribution of the log-linear parameters.

The algorithm of Madigan and Raftery (1994) works well for the HLL models, for instance, number of possible models reduced to 3 from 168 in the road traffic data application. In addition, the whole procedure is applied for various τ values. It is seen that the results are not sensitive to choice of the τ. However, results are sensitive to choice of O_R and O_L. For the choice of them, we suggest running the whole procedure for several times and careful investigation of the B values of the "down" and "up" algorithms. Because smaller values of the O_R causes the exclusion of appropriate complex models in the "down" algorithm, and greater values of the O_L causes the exclusion parsimonious appropriate models in the "up" algorithm. Application of equation (4.6) is optional. If the number of possibly acceptable models is small enough, it may not be applied.

In conclusion, BMA is an effective way of including model uncertainty in the analysis. As seen in the application, some parameters are not included in the best model but they have positive effect on the estimation of expected cell counts, even if they are small. In addition, it is not only used for the estimation of log-linear parameters but also for the estimation of various percentiles of the posterior distributions of them. As a future work, results of this work can be extended to other likelihood-prior settings.

References

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

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

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

Gelfand, A.E. and Mallick, B.K. (1995). Bayesian analysis of proportional hazards built from monotone functions. *Biometrics* **51**, 843-852.

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

Hoeting, J.A., *et al.* (1999). Bayesian model averaging: A tutorial (with discussion). *Statistical Science* **14**, 382-417.

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

King, R. and Brooks, S.P. (2001b). On the analysis of population size. *Biometrika* **88**, 317-336.

Leamer, E.E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimetal Data*. Wiley.

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

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

Madigan, D. and 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.

Raftery, A.E. (1995). Bayesian model selection in social research (with discussion). *Social Methodology* **25**, 111-163, 1995.

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

Raftery, A.E., Madigan, D. and Volinsky, C.T. (1994). Accounting for model uncertainty in survival analysis improves predictive performance (with discussion). *In Bayesian Statistics 5* (Edited by J. Bernardo, J. Berger, A. Dawid and A. Smith), 323-329, Oxford University Press.

Roberts, H.V. (1965). Probabilistic prediction. *Journal of the American Statistical Association* **60**, 50-62.

Received October 23, 2007; accepted January 21, 2008.

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

Canan Hamurkaroglu
Department of Statistics
Hacettepe University
Beytepe, Ankara, 06800, TURKEY
caca@hacettepe.edu.tr