

Model selection for categorical data with factor level merging

Scelta del modello per dati categoriali attraverso l'aggregazione di livelli di fattori

Claudia Tarantola ¹
Dip. di Ec. Pol. e Met. Quant.
Università di Pavia
ctaranto@unipv.it

Petros Dellaportas
Department of Statistics
Athens University of Econ. and Business
petros@aueb.gr

Riassunto: La struttura di indipendenza tra un insieme di variabili categoriche, rappresentate da una tabella di contingenza, viene descritta attraverso un modello grafico non orientato. Se la dimensione della tabella esaminata è elevata può essere desiderabile ridurla combinando tra loro alcune categorie. Si introducono condizioni in base alla quali questa operazione non altera le proprietà markoviane del grafo associato alla tabella in esame. Queste condizioni vengono utilizzate per sviluppare una metodologia Bayesiana per la scelta del modello.

Keywords: Bayesian inference, Contingency table, Graphical models, Log-linear models, Reversible jump MCMC.

1. Introduction

We consider the problem of reducing the dimension of a given contingency table by merging adjacent categories, or levels, of variables. Since we do not deal with the problem of collapsing a contingency table by marginalizing one variable, we assume that at least one of the factors in the contingency table has more than two levels.

This problem has attracted the interest of many authors in the past. In particular, Wermuth and Cox (1998) have extensively illustrated, through a series of real data examples, the resulting simplified interpretation of the association between variables when level merging is possible. Apart from these obvious advantages, level aggregation may be useful in designing level borders in future studies when, for example, precise measurement is expensive. It can also be a powerful tool in data squashing where data is compressed in such a way that a statistical analysis carried out on the squashed data provides the same outputs that would have resulted from analysing the entire dataset, DuMouchel (2001). Finally, level merging is the common data disclosure technique called ‘table redesign’, Willenborg and de Waal (2000). It may replace other strategies such as cell suppression, interval protection or perturbation, Fienberg (1997) and it is especially useful when marginal cells themselves might obtain sensitive cells.

In this paper, we use a particular log linear parameterisation presented by Wermuth and Cox (1998) that can be useful when level merging is of interest. Differently from Wermuth and Cox (1998), we propose an aggregation rule that allows merging of variable levels without changing the undirected graphical model that fits the data before the data aggregation. We present a RJMCM algorithm, based on this rule, that allows us to efficiently explore the space of all graphical models and aggregations. See Dellaportas and

¹Ricerca parzialmente finanziata dal progetto PRIN 2001

Tarantola (2003) for more details. In the following, we assume that the reader is familiar with the graphical models terminology, for more details see e.g. Lauritzen (1996) and Whittaker (1990)

2. A suitable log-linear parameterisation

We now briefly describe the log-linear parameterisation by Wermuth and Cox (1998). This is particularly useful since the contrast matrix, the inverse of the design matrix, identifies parameters that are interpreted as functions of odds ratios or conditional odds ratios.

Let $X_1, X_2, \dots, X_N \in \mathcal{V}$ be a set of categorical variables whose outcomes are labelled by the integers in label sets $I_i, i = 1, \dots, N$, with $I_i = \{1, 2, \dots, |I_i|\}$. The joint distribution of the variables is represented by a $(|I_1||I_2| \cdots |I_N| \times 1)$ vector

$$\pi^{X_1 X_2 \dots X_N} = \{\pi_{i_1 i_2 \dots i_N}^{X_1 X_2 \dots X_N}, i_j \in I_j, j = 1, \dots, N\}.$$

A log-linear parameterisation in terms of a suitable matrix C of dimension $|I_1||I_2| \cdots |I_N| \times |I_1||I_2| \cdots |I_N|$ is given by

$$\gamma^{X_1 X_2 \dots X_N} = C_{|I_1||I_2| \dots |I_N|} \log \pi^{X_1 X_2 \dots X_N} \quad (1)$$

where γ is a $(|I_1||I_2| \cdots |I_N| \times 1)$ vector of identifiable log-linear parameters. The matrix $C_{|I_1||I_2| \dots |I_N|}$ is called contrast matrix because it contains weights that define contrasts of log probabilities. The contrast matrix can be computed via the Kronecker product (in the inverse order) of matrices $C_{|I_i|}, i = 1, \dots, N$. The generic matrix $C_{|I|}$ has dimension $|I| \times |I|$ with elements C^{ij} given by

$$C^{ij} = \begin{cases} 1 & \text{if } i = 1 \text{ or } i = j \\ -1 & \text{if } i = j + 1 \\ 0 & \text{otherwise.} \end{cases}$$

3. Conditions for level mergings

The γ -parameters correspond to main effects and interactions of different orders for specific adjacent factor levels. They are interpreted as functions of odd ratios and conditional odd ratios. In particular the second order interaction parameters are function of conditional odds ratios.

For simplicity, consider first three generic variables X_1, X_2 and X_3 . In this case, the second order interaction term, $\gamma_{i(i+1).j(j+1)}^{X_1 X_2}$, is the sum of $|I_3|$ log-conditional-odds ratios, one for each value of X_3 . The conditional odds ratio of X_1 versus X_2 given $X_3 = k$ is given by

$$Pr(X_1 = i, X_2 = j | X_3 = k) = \frac{\pi_{ijk}^{X_1 X_2 X_3} \pi_{(i+1)(j+1)k}^{X_1 X_2 X_3}}{\pi_{j(j+1)k}^{X_1 X_2 X_3} \pi_{(i+1)jk}^{X_1 X_2 X_3}}. \quad (2)$$

If this ratio is one, there is independence between X_1 and X_2 in the 2×2 sub-table specified by levels i and $i + 1$ of X_1 , j and $j + 1$ of X_2 and k of X_3 . If the ratio in (2) is

one for all values of k , then $\gamma_{i(i+1),j(j+1)}^{X_1 \cdot X_2} = 0$. Thus, absence of a second order interaction term is related to a conditional independence. Following Wermuth and Cox (1998), there is conditional independence for the levels i and $i + 1$ of X_1 and j and $j + 1$ of X_2 if and only if $\gamma_{i(i+1),j(j+1)}^{X_1 \cdot X_2} = 0$ and all higher order terms involving levels i and $i + 1$ of X_1 , and j and $j + 1$ of X_2 , are also zero. According to these authors, such conditional independence is an adequate criterion to impose merging of levels.

We diversify from this rule since we require both $\gamma_{i(i+1)}^{X_1} = 0$ and that all higher order parameters involving variable X_1 at levels $i, i + 1$ are also zero (*level-hierarchy*, Dellaportas and Tarantola (2003), p 5). Notice that our extra requirement seems natural: if levels i and $i + 1$ of variable X_1 are merged together, there seems to be no need to exist any γ -parameter involving these two levels, so $\gamma_{i(i+1)}^{X_1}$ should be omitted.

This rule can also be justified considering the properties of the graphical model associated to the examined data. Suppose that $\gamma_{i(i+1)}^{X_1} = 0$ and that the level hierarchical principle holds. Dellaportas and Tarantola (2003) have shown that in this case the dimensionality of the contingency table can be reduced by merging levels i and $i + 1$ of variable X_1 without modifying the conditional independence structure of the associated model (Propositions 1 and 2). The original data and the one obtained after the merging can be represented by the same graphical model. This rule is the basic of the development of the Bayesian model selection procedure described in the following section.

4. Bayesian model selection procedure

Our target is to develop an efficient methodology to deal simultaneously with uncertainty in both the model and the aggregation level. The exploration of the space of all models under all possible data aggregations is an extremely demanding computational task. In fact for every possible graphical models $M_1 = 2^{\binom{N}{2}}$ we have $M_2 = \prod_{k=1}^N (2^{|I_k|-1} - 1)$ possible level aggregations. The possible models under consideration are $M = M_1 M_2$, hence a full comparison of their posterior probabilities becomes infeasible. A possible solution to this problem is the use of reversible Jump MCMC techniques, Green (1995).

In order to develop our model selection procedure it is convenient to transform (1) to the standard log-linear parameterisation. The design matrix for the full (saturated) model is given by $Z = C_{|I_1||I_2|\dots|I_N|}^{-1}$ so we can re-write (1) as

$$\log \mu^{X_1 X_2 \dots X_N} = \log \mu = Z \gamma^{X_1 X_2 \dots X_N} + \log n.$$

where $\mu^{X_1 X_2 \dots X_N}$ represents the corresponding cell mean vector and n is a vector with elements the total sample size. For a given Z , the design matrix for any aggregation can be obtained by applying a sequence of rules based on the condition on the gamma terms presented in the previous section.

Our Bayesian model determination MCMC procedure resembles closely that of Dellaportas and Forster (1999). On the model space we assign a discrete uniform prior, whereas the prior density on the parameter space is obtained by adapting the default prior proposed by Dellaportas and Forster (1999) for the analysis of standard log-linear models.

We now briefly describe our reversible jump MCMC algorithm. Suppose that the current state of the Markov chain is model m , with corresponding parameters γ_m of dimension $d(\gamma_m)$, likelihood function $f(y|m, \gamma_m)$ and prior density $f(\gamma_m|m)$. The algorithm proceeds as follows:

- (i) Propose a new model m' with probability $j(m, m')$.
- (ii) Generate $\gamma'_{m'}$ from a specified proposal density $q(\gamma'_{m'})$.
- (iii) Accept the proposed move to model m' with probability

$$\alpha = \min \left(1, \frac{f(y|m', \gamma'_{m'})f(\gamma'_{m'}|m')f(m')j(m', m)q(\gamma'_{m'})}{f(y|m, \gamma_m)f(\gamma_m|m)f(m)j(m, m')q(\gamma_m)} \right),$$

since the jacobian is one.

At each step we either propose to remain to the same model, or jump to another graphical model, or jump to another aggregation. A new graphical model is proposed by adding or deleting an edge from the current model. In order to update the aggregation we have experimented different strategies. On the one extreme, a ‘long jump’ is proposed by randomly choosing, with equal probability, one of all possible aggregations. On the other extreme, a ‘local jump’ picks up a factor with equal probability, and then proposes a level-merge or level-split move with equal probabilities. To update γ_m within the same model, we use a block random walk Metropolis-Hastings move with proposal density $q(\gamma_m)$ a multivariate normal with mean equal to the current point and variance equal to the (properly scaled) maximum likelihood estimated covariance matrix. Our proposal densities $q(\gamma'_{m'})$ for jump moves are normal densities based on maximum likelihood estimates that are calculated with Newton-Raphson algorithm. This procedure is not performed for all models but only for the models that are proposed to be visited by the MCMC algorithm. Note if the proposed model is accepted, these estimates are used for within-model parameter updates as described above.

During the talk the performance of our methodology will be illustrated with reference to real sets of data. Note that, although we have proposed a Bayesian approach, our condition can be adopted in a classical statistical models choice setting.

References

- Dellaportas P. and Forster J. (1999) Markov chain monte carlo model determination for hierarchical and graphical log-linear models, *Biometrika*, 86, 615–633.
- Dellaportas P. and Tarantola C. (2003) Categorical data squashing by combining factor levels, *Technical report, University of Pavia*, 153.
- DuMouchel W. (2001) Data squashing: constructing summary data sets, in *Handbook of Massive Data Sets*, J. Abello editor, Kluwer Academic Publishers, USA.
- Fienberg S.. (1997) Confidentiality and disclosure limitation methodology: Challenges for national statistics and statistical research, *Technical report, Carnegie Mellon University*, 668, 197–205.
- Green P.J. (1995) Reversible jump markov chain monte carlo computation and bayesian model determination, *Biometrika*, 82, 711–32.
- Lauritzen S. (1996) *Graphical Models*, Oxford University Press, Oxford.
- Wermuth N. and Cox D. (1998) On the application of conditional independence to ordinal data, *Inter. Statist. Review*, 66, 181–199.
- Whittaker J. (1990) *Graphical models in applied multivariate statistics*, John Wiley and Sons, New York; Chichester.
- Willenborg L. and de Waal T. (2000) *Elements of Statistical Disclosure Control*, Springer-Verlag, New York.