Robust maximum likelihood estimation of latent class models

Francesco Bartolucci[1], Brian Francis[2], Silvia Pandolfi[1], Fulvia Pennoni[3]

[1] Department of Economics, University of Perugia, Italy
[2] Lancaster University, United Kingdom
[3] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Italy

E-mail for correspondence: fulvia.pennoni@unimib.it

Abstract: We develop a suitable reweighting approach to deal with outliers when maximum-likelihood estimation is used to estimate latent class models. In such a context, the EM algorithm is used and the presence of singularities and spurious local maxima is common. The proposed method is motivated by an application aimed at finding clusters of offending behaviours.

Keywords: Categorical data; Expectation-Maximization algorithm; Local maxima; Outliers; Trimmed log-likelihood.

1 Introduction

We address the problem of the outliers detection and robust estimation in the context of latent class (LC) models for categorical data. These models, introduced by Lanzarsfeld and Henry (1968), represent a valid tool to explain the association between the categorical variables by assuming the existence of a finite set of latent classes. Maximum likelihood estimates of the model parameters are found by using the Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977) and inference is based on the solution corresponding to the largest value of the log-likelihood at convergence. As for other finite mixture models (McLachlan and Peel, 2000), strategies to single out the global maximum (e.g. Aitkin et al., 1981; McCutheon, 2002) still need improvements. The likelihood may be multimodal and to deal with this problem, a random rule may be applied for the initialization of the EM algorithm. This method, when repeated a suitable number of

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

times, allows us to explore the parameter space adequately, provided that the number of parameters is reduced.
We propose a joint use of the trimmed maximum likelihood approach as that developed by Garcia-Escudero et al. (2014) and of appropriate constraints on the parameters of the latent variable and on the parameters of the conditional distribution of the response variables given the latent variable (Bartolucci et al., 2007; Pennoni, 2014). The aim is to obtain a potential improved fit of the LC model. This may also allows us to cope with the problem of the multimodality of the likelihood function.
We illustrate the proposed method by considering the sample data from 1 in 13 sample of all England and Wales offenders born in 1953 related to the dates of conviction and type of offenses from age 16 up to age 20.

2 Proposal

We denote by $Y_1, \ldots, Y_r$ the categorical response variables with categories labeled from 0 to $c-1$. It includes the case of binary responses when $c = 2$. We suppose the existence of a latent variable U with k levels, $u = 1, \ldots, k$. The model parameters are the conditional probabilities of a single response variable y_j given the latent variable denoted by $\phi_{jy|u}$ and the weights π_u for each latent class. Obviously, we have $\sum_{u=1}^{k} \pi_u = 1$ and $\sum_{y=0}^{c-1} \phi_{jy|u} = 1$. Given a sequence of responses $\boldsymbol{y} = (y_1, \ldots, y_r)$, the conditional distribution of all responses given the latent variable is given by

$$p(\boldsymbol{y}|u) = \prod_{j=1}^{r} \phi_{jy_j|u},$$

and, then, the manifest probability of this sequence is equal to

$$p(\boldsymbol{y}) = \sum_{u} \pi_u p(\boldsymbol{y}|u).$$

The posterior probability that an individual with response vector $\boldsymbol{y}$ belongs to the latent class u is used to construct the allocation rule for each individual to a latent class.
In order to estimate the model parameters on the basis of the observations $\boldsymbol{y}_i$, $i = 1, \ldots, n$, we maximize the weighted log-likelihood

$$\ell(\boldsymbol{\theta}) = \sum_{i} w_i \log p(\boldsymbol{y}_i),$$

where the weight w_i is close to 0 for outliers and θ denotes the vector of the model parameters. This maximization is based on the EM algorithm and uses the weighted complete log-likelihood, which is equal to

$$\ell^*(\boldsymbol{\theta}) = \sum_{i} w_i \left[\sum_{u} z_{iu} \left(\sum_{j} \sum_{y} I(y_{ij} = y) \log \phi_{jy|u} + \log \pi_u \right) \right],$$

where $I(\cdot)$ is the indicator function and z_{iu} is an indicator variable equal to 1 if subject i belongs to latent class u and to 0 otherwise.
The estimation is carried out by using a modified version of the EM algorithm. In the standard case, the EM is performed in the following way. At the E-step, we consider the conditional expected value of the frequency of subjects in each latent class u having value y for the j-th response variable. These are computed at the current value of the parameters. At the M-step, the complete data log-likelihood is maximized by using exact solutions for π_u and for $\phi_{jy|u}$.
The optimal number of latent classes is selected by considering the Bayesian information criterion (BIC, Schwarz, 1978) which involves a penalty for the number of parameters:

$$BIC = -2\hat{\ell} + \log(n)\#par,$$

where, for a given model, $\hat{\ell}$ is the maximum of the log-likelihood and $\#par$ is the number of free model parameters. According to this criterion, the number of classes corresponding to the minimum of the index has to be selected. The estimated proportion of classification error is also considered for each latent class which states how well the latent classes are separated.
Here, we propose to use a similar strategy as that developed by Garcia-Escudero et al. (2014); see also Neykov et al. (2007). First, we identify the number of latent classes according to the BIC index. Then, we perform the modified version of the EM algorithm for a large number of random starting values. In the E-step, the observations with the smallest likelihood contribution are tentatively discarded, by setting the corresponding posterior probabilities z_{iu} equal to zero for all $u = 1, \ldots, k$. In the M-step the model parameters are updated on the basis of the selected subsample of observations. After applying the trimmed EM steps, the associated weighted likelihood is evaluated, by setting $w_i = 0$ for the discarded observations and $w_i = 1$ for the selected subsample, until convergence of the algorithm.

3 Application

In criminology research a common task is that of clustering criminal behaviors accounting for their evolution in time. For this aim, LC models may be effectively used. We apply the proposed methodology for the analysis of criminal data referred to males and females offenders in England and Wales. More precisely, the data refer to the 1953 cohort (Francis et al., 2010) and concern 38 binary different indicators of criminal activity. We consider only the young males and females within the cohort age range 16-20 years old, which covers $n = 4558$ cases.
We first fit the LC model for an increasing number of latent classes k from 1 to 11. Table 1 shows the results of this preliminary fitting, in terms of maximum log-likelihood and the corresponding values of the BIC index.

According to this criterion, we select 8 latent classes. The latent classes identified according to the estimated conditional probabilities are the following:

1. shoplifting (9,8%);
2. criminal damage (7,7%);
3. theft by employee: with some fraud and forgery (3,6%);
4. theft (17,7%);
5. versatile of type 2: theft from vehicles, handling and receiving stolen goods, burglary and going equipped (24,9%);
6. versatile of type 1: burglary, commercial burglary theft, criminal damage with some burglary, theft, violence, shoplifting (8,8%);
7. violence (13,4%);
8. fraud and forgery with some theft, handling and receiving stolen goods (14,1%).

We then apply the proposed trimmed estimation strategy, assuming the selected number of latent classes $k = 8$, for different trimming levels, 0.25%, 0.9%, and 2%. Preliminary results are reported in Table 2 for the 0.25% trimming level, which leads to discard 12 observations. For these cases (outlier), the table reports the corresponding residual deviance.
In particular, the first case refers to a subject convicted for violence, burglary, going equipped and robbery. Case 1884 was convicted for violence, robbery and theft. Case 181 was convicted for violence, sexual offenses, handling and receiving stolen goods, criminal damage and perjury/attempting to pervert course of justice. The group cases made of id's 1581, 3255, 4264, 4265 refer to subjects convicted for violence, sexual consensual, burglary (dwelling), going equipped, theft, theft from vehicles, shoplifting.
The group cases made of id's 1566, 1576, 2564 for violence, sexual with above 16 years old, burglary (dwelling), theft, theft from person, theft by employee and criminal damage. Case 3719 was convicted for violence, sexual with above 16 years old, sexual under 16 years old and sexual consensual, burglary (dwelling), burglary (other) and theft from vehicles. Case 1361 was convicted for lethal violence, sexual under 16 years old, going equipped, shoplifting and criminal damage. These are cases mostly characterized by sexual offenses with violence and property offenses. This cluster collects the most dangerous individuals, which were not properly identified in the standard LC model.
Finally, it is worth noting that the results of the trimmed estimation strategy may also be used to define sensible starting values for the EM algorithm in the standard approach, so as to prevent the problem of the multimodality of the model log-likelihood.

TABLE 1. Selection of the number of latent classes for the LC model; k is the number of latent classes, $\hat{\ell}$ is the corresponding maximum log-likelihood, $\#par$ is the corresponding number of parameters, and index BIC is defined in Section 2.

k	$\hat{\ell}$	$\#par$	BIC
1	-23315,36	34	46916,41
2	-23006,21	69	46592,18
3	-22671,11	104	46216,07
4	-22498,11	139	46164,17
5	-22294,99	174	46052,00
6	-22162,30	209	46080,72
7	-22089,01	244	46228,23
8	-21748,02	279	45840,32
9	-21775,68	314	46189,74
10	-21588,08	349	46108,62
11	-21539,43	384	46305,41

TABLE 2. Cases with high log-likelihood contribution.

ID numbe r	$-2\ell^*$(contribution)
707	68,0585
1884	52,8958
181	49,2723
1581	48,3980
3255	48,3979
4264	48,3979
4265	48,3979
3719	46,9193
1566	45,1571
1576	45,1571
2564	45,1571
1361	42,3028

Acknowledgments: We acknowledge the financial support from the grant RBFR12SHVV of the Italian Government (FIRB project "Mixture and latent variable models for causal inference and analysis of socio-economic data"). F. Pennoni also thanks the financial support of the STAR project *"Statistical models for human perception and evaluation"* funded by the University of Naples Federico II.

References

Aitkin M., Anderson, D., and Hinde, J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society, Series A*, **144**, 419 - 448.

Bartolucci, F., Pennoni, F., and Francis, B. (2007). A latent Markov model for detecting patterns of criminal activity. *Journal of the Royal Statistical Society, Series A*, **170**, 115 – 132.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society. Series B*, **39**, 1 – 38.

Francis, B., Liu, J., and Soothill, K. (2010). Criminal lifestyle specialization: female offending in England and Wales. *International criminal justice review*, **20**, 188 – 204.

Garcia-Escudero, L., Gordaliza, A., Mayo-Iscar, A. (2014). A constrained robust proposal for mixture modeling avoiding spurious solutions. *Advances in Data Analysis and Classification*, **8**, 27 – 43.

Lazarsfeld, P.F. and Henry, N.W. (1968). *Latent Structure Analysis*. Boston: Houghton Mifflin.

Neykov, N., Filzmoser, P., Dimova, R., and Neytchev, P. (2007). Robust fitting of mixtures using the trimmed likelihood estimator. *Computational Statistics & Data Analysis*, **52**, 299 – 308.

McLachlan, G.J. and Peel, D. (2000). *Finite Mixture Models*. Wiley Series: New York.

McCutheon, A.L. (2002). Basic concepts and procedures in single and multiple group latent class analysis. In: *Applied Latent Class Analysis*, Hagenaars, J.A., McCutheon, A.L. (eds.), pp. 56 – 85. Cambridge University Press.

Pennoni, F. (2014). *Issues on the estimation of latent variable and latent class models*. Scholars' Press, Saarbücken.

Schwarz, G. (1978). Estimation the dimension of a model. *The Annals of Statistics*, **6**, 461 – 464.