Proceedings of the

38th International Workshop on Statistical Modelling

Durham, UK
14-19 July, 2024

Editors: Jochen Einbeck, Reza Drikvandi, Georgios Karagiannis, Konstantinos Perrakis, Qing Zhang

Proceedings of the 38th International Workshop on Statistical Modelling Durham, July 14-19, 2024, Jochen Einbeck, Reza Drikvandi, Georgios Karagiannis, Konstantinos Perrakis, Qing Zhang (Editors),
Durham, 2024.
ISBN 978-0-907552-44-4

Editors:

Jochen Einbeck, jochen.einbeck@durham.ac.uk
Reza Drikvandi, reza.drikvandi@durham.ac.uk
Georgios Karagiannis, georgios.karagiannis@durham.ac.uk
Konstantinos Perrakis, konstantinos.perrakis@durham.ac.uk
Qing Zhang, qing.zhang@durham.ac.uk

Durham University
Mathematical Sciences & Computer Science Building
Durham University
Upper Mountjoy Campus
Stockton Road
Durham University
DH1 3LE

Model selection procedure in multilevel cross-classified latent class models

Nicola Piras[1], Silvia Columbu[1], Jeroen K. Vermunt[2]

[1] Department of Mathematics and Computer Science, University of Cagliari, Italy
[2] Department of Methodology and Statistic, Tilburg University, The Netherlands

E-mail for correspondence: nicola.piras97@unica.it

Abstract: The availability of proper selection criteria is of fundamental importance in the definition of latent class analysis models. When the data structure is multilevel the selection procedure must be applied to each level of the model. In the case of the multilevel cross-classified extension, we propose to apply a three step procedure that takes into account the mutually dependence between the two levels of the structure in the selection. The performances of the method are investigated through simulation studies in which different information criteria are considered. The definition of these criteria are based on approximations of the log-likelihood, which is intractable in such a cross-classified structure.

Keywords: Information Criteria; Multilevel Cross-Classified; Latent Class.

1 Introduction

When performing a latent class (LC) analysis it is crucial to apply model selection criteria for determining the number of latent classes. The decision on that number is affected by different factors (like the sample size and the level of separation between the classes) that must be taken into account in the definition of selection measures. In case of multilevel data, the task of selection gets even more complex as it concerns latent classes at all levels of the data structure. Indeed, the extension of latent class models to multilevel data can be defined taking a set of membership variables for each level. The problem of selection in the hierarchical model has been exhaustively discussed in (Lukociene, Varriale, Vermunt 2010). In this work we focus instead on the cross-classified case, introduced in (Columbu, Piras, Vermunt 2023), in which units are simultaneously nested within two higher level groups (for instance, children belonging to both schools and neighborhoods). Therefore, the selection applies to the number of level-1 latent classes (L) and the number of two separate sets of level-2 latent classes (H and R).

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024.

We propose to consider a three step approach similar to what done in the hierarchical frame. That is, in a first step we determine the number of lower level classes as in a standard LC, ignoring the multilevel structure. Then we determine the number of higher level classes, taking fixed the number of level-1 latent classes selected at the previous step. Finally, we determine again the number of lower level classes fixing the number of higher level classes at the values we obtained from the second step.
The implementation of a three step procedure enables to take under consideration the mutually dependence between the levels of the data structure. The selection of latent classes at level-1 influences that of level-2, the reverse is also true, however in the second direction the dependency is typically weaker.
In the LC framework, and more in general in mixture modeling, the problem of choosing the right number of classes corresponds to a model selection problem. In this context, information criteria (IC) based on some form of log-likelihood penalization are usually considered. Such statistics are built as a weighted combination between the fit of the model and its complexity, measured in terms of number of free parameters involved. The missing data structure typical of LC, with the presence of unknown membership variables to be estimated, requires the introduction of a complete version of the likelihood. Therefore, the goodness of fit can be measured taking either the log-likelihood (logL) value or the complete log-likelihood (CL). In the extension to the multilevel cross-classified version the definition of IC gets more complex as both levels of the structure must be considered and no finite expression of the data likelihood is available. Indeed, the double missing data structure at the higher level makes the likelihood untractable, implying that approximated versions of it can be computed. We will consider multiple IC statistics by letting vary the penalization terms that are computed considering the number of latent classes and/or the sample size (in a BIC fashion). In addition, for each of them we take two versions, one based on the logL and a second based on the CL. Given the presence of two levels in the data, in the level-2 of the structure the sample size can be intended as the total number of individuals (n) or the number of combinations of level-2 cross classified units (K and Q) present in the data.

2 Three-step procedure

The procedure we consider consists of three steps, and is based on three criteria for level-1 and their correspondent definition for the cross-classified level-2. For each IC we consider also a version with CL, generally denoted as IC_e, and for what concerns the BIC we consider alternatives penalized by the number of level-2 groups within which the observations are nested (BIC_g and BIC_{eg}). The steps taken are:

1) Determine the number of lower level classes L fitting standard LC, without taking into account the multilevel structure. Level-1 information criteria used are:

$$BIC = -2 \cdot logL + ((L-1) + npar) \cdot log(n)$$
$$AIC = -2 \cdot logL + 2 \cdot ((L-1) + npar)$$
$$AIC3 = -2 \cdot logL + 3 \cdot ((L-1) + npar)$$

$npar = \sum_{i=1}^{I} L \cdot (C_i - 1)$ is the number of free distribution parameters with I categorical indicators of C_i categories. The lower the value of an IC, the better the model.

2) Determine the number of higher level cross-classified classes H and R fixing the number of lower level classes to that selected in step 1. Level-2 information criteria used are defined as:

$$BIC = -2 \cdot logL + ((R-1) + (H-1) + (H \cdot R \cdot (L-1)) + npar) \cdot log(n)$$

$$AIC = -2 \cdot logL + 2 \cdot ((R-1) + (H-1) + (H \cdot R \cdot (L-1)) + npar)$$

$$AIC3 = -2 \cdot logL + 3 \cdot ((R-1) + (H-1) + (H \cdot R \cdot (L-1)) + npar)$$

the analogous definition taking CL instead of logL and substituting n with the number of group combinations available in the data can be also considered.

3) Determine again the number of level-1 classes L, setting this time the number of level-2 cross-classified classes to the value selected in step 2. This third step allows to evaluate if the suitable number of lower level classes can change after taking into account the multilevel structure. The same criteria as in step 2 are applied.

TABLE 1. Number of replicates in which the investigated criteria, at step 2 and 3, estimated various combinations of level-1 and level-2 cross-classified latent classes. The correct combination is $(4, 3, 2)$.

		Step 2				Step 3		
IC	n_{kq}	4 2 2	**4 3 2**	4 2 3	4 3 3	3 3 2	**4 3 2**	5 3 2
BIC	6	5	33	50	12	0	100	0
	8	0	20	55	24	0	100	0
BIC_g	6	1	32	50	17	0	100	0
	8	0	14	56	30	0	100	0
AIC	6	0	23	47	30	0	100	0
	8	0	4	48	48	0	100	0
$AIC3$	6	26	47	27	1	0	100	0
	8	0	4	48	48	0	100	0
BIC_e	6	37	60	1	2	86	14	0
	8	16	84	0	0	100	0	0
BIC_{eg}	6	29	66	1	4	78	22	0
	8	8	92	0	0	100	0	0
AIC_e	6	13	72	1	14	78	22	0
	8	0	88	0	12	100	0	0
$AIC3_e$	6	14	73	1	12	78	22	0
	8	0	92	0	8	100	0	0

2.1 Simulation results

To evaluate the performances of the criteria proposed and the three step procedure, we carried out simulations taking dataset with two different separation

conditions, obtained setting different values of level-1 units within each k, q group ($n_{kq} = \{6, 8\}$). For each scenario we generated 100 datasets with six binary indicators, and set $K = 30$, $Q = 12$, $L = 4$, $H = 3$ and $R = 2$.
The results of simulations are summarized in Table 1. We observe that at step 2 the versions of the IC based on the CL are those selecting in higher proportion the right number of level-2 classes, keeping fixed the level-1 classes to the correct one $L = 4$. For step 3, once fixed the number of cross-classified classes to the real one ($H = 3$ and $R = 2$) the measures based on logL give always the correct selection. In particular we highlight that at step 2 $AIC3$ performs better with lower separation, while increasing the separation BIC_{eg} provides also good results. These preliminary results are in line with what found in multilevel hierarchical latent class models.

References

Columbu, S., Piras, N. and Vermunt, J. K. (2023). Multilevel Cross-Classified Latent Class Models. In: *CLADAG 2023 Book of Abstracts and Short Papers, ISBN: 9788891935632*, 390 – 393.

Lukociene, O., Varriale, R. and Vermunt, J., K. (2010). The simultaneous decision(s) about the number of lower- and higher-level classes in multilevel latent class analysis. *Sociological Methodology*, **40**, 247 – 283.