Determining the number of components in mixture models for hierarchical data

Olga Lukočienė[1] and Jeroen K. Vermunt[2]

[1] Department of Methodology and Statistics, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands O.Lukociene@uvt.nl
[2] Department of Methodology and Statistics, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands J.K.Vermunt@uvt.nl

Summary. Recently, various types of mixture models have been developed for data sets having a hierarchical or multilevel structure (see, e,g., [9, 12]). Most of these models include finite mixture distributions at multiple levels of a hierarchical structure. In these multilevel mixture models, selection of the number of mixture component is more complex than in standard mixture models because one has to determine the number of mixture components at multiple levels.

In this study the performance of various model selection methods was investigated in the context of multilevel mixture models. We focus on determining the number of mixture components at the higher-level. We consider the information criteria BIC, AIC, and AIC3, and CAIC, as well as ICOMP and the validation log-likelihood. A specific difficulty that occurs in the application of BIC and CAIC in the context of multilevel models is that they contain the sample size as one of their terms and it is not clear which sample size should be used in their formula. This could be the number of groups, the number of individuals, or either the number of groups or number of individuals depending on whether one wishes to determine the number of components at the higher or at the lower level.

Our simulation study showed that when one wishes to determine the number of mixture components at the higher level, the most appropriate sample size for BIC and CAIC is the number of groups (higher-level units). Moreover, we found that BIC, CAIC and ICOMP detect very well the true number of mixture components when both the components' separation and the group-level sample size are large enough. AIC performs best with low separation levels and small sizes at the group-level.

Key words: Multilevel mixture model; Multilevel latent class analysis; Mixture components; BIC; AIC; ICOMP; Validation log-likelihood

1 Introduction

Vermunt [9, 11, 12, 13] proposed several types of latent class (LC) and mixture models for multilevel data sets with applications in sociological, behavioral, and medical research. Examples of two-level data sets include data

from individuals (lower-level units) nested within families (higher-level units), pupils nested within schools, patients nested within primary care centers, and repeated measurements nested within individuals. A multilevel latent class model can be applied when in addition multiple responses are recorded for the lower-level units, and is thus, in fact, a model for three-level data sets. The multilevel LC models dealt with in this paper assume that lower-level units (say individuals) belong to LCs at the lower level and that higher-level units (say groups) belong to LCs at the higher level. In other words, the models contain mixture distributions at two levels.

There is wide variety of literature available on the performance of model selection statistics for determining the number of mixture components in mixture models. The Bayesian (also known as Schwarz's) information criterion (BIC) is the most popular measure for determining the number of mixture components and it is generally considered to be a good measure [5, 7]. Other authors, however, prefer the Akaike information criterion (AIC) [6]. While deciding about the number of mixture components is already a complicated task in standard mixture models, it is even more complex for multilevel mixture models. One of the difficulties consists in choosing the appropriate sample size in the BIC and CAIC formulae:

$$BIC = -2\ln L + k\ln(n) \tag{1}$$

and

$$CAIC = -2\ln L + k(1 + \ln(n)). \tag{2}$$

Here, L is the maximized value of the likelihood function for the estimated model, k is the number of free parameters to be estimated, and n is the number of observations, or equivalently, the sample size. There are several options for defining the sample size in the multilevel context, including the number of groups, the number of individuals, or either the number of groups or number of individuals depending on whether one wishes to determine the number of components at the higher or at the lower level. Neither the literature on mixture models nor the literature on multilevel analysis give hints on what sample size to use in the computation of BIC and CAIC in multilevel mixture models.

This article presents the results of a simulation study in which we compared the performance of several methods for determining the number of mixture components in the multilevel LC models. We investigated the performance of BIC and CAIC using different sample size definitions, as well as compare BIC and CAIC with other model selection measures, such as AIC, AIC3, ICOMP [2], and the validation log-likelihood [8]. Our focus is on deciding about the number of mixture components at the higher level.

The next section describes the multilevel LC model. The design of the simulation study is explained in Section 3. The obtained result are presented in Section 4. The main conclusions are highlighted in the last section.

2 Multilevel latent class model

Let $\boldsymbol{y}_j = (y_{j1}, \ldots, y_{ji}, \ldots, y_{jI})$ denote the vector with the I responses of individual j, $(j = 1, \ldots, n)$. A discrete LC variable is denoted by x_j, a particular LC by l_2, and the number of classes by L_2 $(l_2 = 1, \ldots, L_2)$. The basic assumptions of the LC model are: 1) that each individual belongs to (no more than) one latent class, 2) that the responses of individuals belonging to the same LC are generated by the same (probability) density, and 3) that the I responses of individual j are conditionally independent of one another given his/her class membership. Under these assumptions, the traditional LC model is defined by the following formula:

$$f(\boldsymbol{y}_j) = \sum_{l_2=1}^{L_2} P(x_j = l_2) \prod_{i=1}^{I} f(y_{ji}|x_j = l_2), \tag{3}$$

where $f(\boldsymbol{y}_j)$ is the marginal density of the responses of individual j, $P(x_j = l_2)$ is the unconditional probability of belonging to LC l_2, and $f(y_{ji}|x_j = l_2)$ is the conditional density for response variable i given that one belongs to LC l_2.

A multilevel LC model differs from a standard LC model in that the parameters of interest are allowed to differ randomly across groups (across higher-level units). It should be noted that the multilevel LC model is actually a model for three-level data sets; that is, for multiple responses (level-1 units) nested within individuals (level-2 units) and individuals (level-2 units) are nested within groups (level-3 units). The random variation of LC parameters across groups can be modelled using continuous or discrete group-level latent variables, or by a combination of these two. It should be noted that using the discrete latent variable approach, where parameters are allowed to differ across latent classes of groups, is similar to using a nonparametric random effects approach [1, 10]. In this article we focus on this discrete approach which makes use of group-level latent classes.

Let $\boldsymbol{y}_{kj} = (y_{kj1}, \ldots, y_{kji}, \ldots, y_{kjI})$ denote the I responses of individual j $(j = 1, \ldots, n_k)$ from group k $(k = 1, \ldots, K)$, and $\boldsymbol{y}_k = (\boldsymbol{y}_{k1}, \ldots, \boldsymbol{y}_{kj}, \ldots, \boldsymbol{y}_{kn_j})$ the full response vector of group k. The class membership of individual j from group k is now denoted by x_{kj}. In the discrete random-effects approach it is assumed that every group belongs to one of the L_3 group-level LCs or mixture components. Let w_k denote the class membership of group k and l_3 denote a particular group-level LC $(l_3 = 1, \ldots, L_3)$. The multilevel LC model can then be described by the following two equations:

$$f(\boldsymbol{y}_k) = \sum_{l_3=1}^{L_3} P(w_k = l_3) \prod_{j=1}^{n_k} f(\boldsymbol{y}_{kj}|w_k = l_3) \tag{4}$$

and

$$f(\boldsymbol{y}_{kj}|w_k = l_3) = \sum_{l_2=1}^{L_2} P(x_{kj} = l_2|w_k = l_3) \prod_{i=1}^{I} f(y_{kji}|x_{kj} = l_2, w_k = l_3). \quad (5)$$

Equation (4) shows how the responses of the n_k individuals belonging to group k are linked to obtain the density for the full response vector of group k, $f(\boldsymbol{y}_k)$. More precisely, it shows that the group members' responses are assumed to be mutually independent conditional on the group-level class membership. Furthermore, from Equation (5) it can be seen that both the lower-level mixture proportions – $P(x_{kj} = l_2|w_k = l_3)$ – and the parameters defining the response densities – $f(y_{kji}|x_{kj} = l_2, w_k = l_3)$ – may differ across higher-level mixture components.

Two interesting special cases of the multilevel LC model are obtained by constraining the terms appearing in Equation (5) [10, 13]. The first special case, which is the one we will use in our simulation study, is a model in which the individual-level class membership probabilities differ across group-level classes, but in which the parameters defining the conditional distributions for the response variables do not vary across group-level classes. The latter implies that $f(y_{kji}|x_{kj} = l_2, w_k = l_3) = f(y_{kji}|x_{kj} = l_2)$. The second special case is a model in which the parameters defining the conditional distributions for the response variables differ across group-level classes, but in which individual-level class membership probabilities do not vary across group-level classes. The latter restriction implies that $P(x_{kj} = l_2|w_k = l_3) = P(x_{kj} = l_2)$. The first special case is the most natural specification if one uses the multilevel LC models a multiple-group LC model for a large number of groups. The second one is more similar to three-level random-effects regression analysis.

The unknown parameters of a multilevel LC model can be estimated by means of Maximum Likelihood (ML). For this purpose one can use the Expectation-Maximization (EM) algorithm [3] – the most popular algorithm for obtaining ML estimates in the context of mixture modeling – which in the context a multilevel LC model requires a specific implementation of the E step. As shown by Vermunt [9, 12], the relevant marginal posterior probabilities can be computed in an efficient way by making use of the conditional independence assumptions implied by the multilevel LC model. This special version of the EM algorithm, as well as a Newton-Raphson algorithm with analytic first-order derivatives and numerical second-order derivatives are implemented in the Latent GOLD software package [14]. The last version of the Latent GOLD software package was used for the realization of the simulation study reported below.

3 Design of the simulation study

The purpose of the simulation study was to compare the performance of different model selection indices for determining the number of mixture components at the higher-level in the multilevel LC model. These indices are BIC, AIC,

AIC3, CAIC, ICOMP, and the validation log-likelihood. For BIC and CAIC we use two versions, one with the number of groups and one with the total number of individuals as the sample size.

Because we focus on detecting the correct number of group-level classes rather than on detecting the correct number of individual-level classes, we decided to keep the LC structure at the individual level fixed in our simulation design. More specifically, we used a three-class model ($L_2 = 3$) for six binary responses ($I = 6$). The class-specific "positive" response probabilities – $P(y_{kji} = 1|x_{kj} = l_2)$ – for the six items were set to $\{0.8, 0.8, 0.8, 0.8, 0.8, 0.8\}$, $\{0.8, 0.8, 0.8, 0.2, 0.2, 0.2\}$, and $\{0.2, 0.2, 0.2, 0.2, 0.2, 0.2\}$ for LCs 1, 2, and 3, respectively. So LC 1 has a high probability of giving the positive response for all items, LC 3 a low probability for all items, and LC 2 a low probability for 3 items and a high probability for the other 3 items. Our choice of number of items, number of classes, and response probabilities is such that we obtain a condition with moderately separated classes. To give an impression of the level of the separation, our setting corresponds to an entropy based R-squared – a measure indicating how well one can predict the class memberships based on the observed responses – of about 0.63. By using moderately separated classes at the lower level, we make sure that detection of the group-level classes is neither made too easy nor too difficult as far as this part of the model is concerned.

So far we have discussed the factors that were fixed in the simulation study. The three factors which were varied are the lower-level sample size, the higher-level sample size, and the number of LCs at the higher-level. Previous simulation studies have shown that the sample size, the number of classes, and the level of separation between the classes are the most important factors affecting the performance of model selection measures in the context mixture models [4]. It should be noted that the separation between the higher-level classes can be manipulated in several ways; that is, by increasing the level of separation of the lower-level classes, by increasing the number of individuals per group (the lower-level sample size n_k), and by making the $P(x_{kj}|w_k)$ more different across values of w_k. We used only the lower-level sample size n_k to manipulate the level of separation. More specifically by using $n_k = 5, 10, 15, 20$ and 30 for the number of the lower-level units per higher-level unit, we created conditions ranging from very low to very high separation. The corresponding entropy-based R-squared values are given below after discussing the other design factors.

The other two factors that were varied are the higher-level sample size, for which we used $K = 50$ and 500, and the number of classes at the higher level, for which we used $L_3 = 2$ and 3. In the condition with two higher-level classes, the model probabilities were set to $P(w_k = \{1, 2\}) = \{0.5, 0.5\}$, $P(x_{kj} = \{1, 2, 3\}|w_k = 1) = \{0.2, 0.2, 0.6\}$, and $P(x_{kj} = \{1, 2, 3\}|w_k = 2) = \{0.4, 0.4, 0.2\}$. These probabilities are such that the two LCs are moderately distinguishable. The condition with three LCs at the higher-level was created by splitting the above second class into two new classes. For this condition,

the model probabilities were $P(w_k = \{1, 2, 3\}) = \{0.5, 0.25, 0.25\}$, $P(x_{kj} = \{1, 2, 3\}|w_k = 1) = \{0.2, 0.2, 0.6\}$, $P(x_{kj} = \{1, 2, 3\}|w_k = 2) = \{0.2, 0.6, 0.2\}$, and $P(x_{kj} = \{1, 2, 3\}|w_k = 3) = \{0.6, 0.2, 0.2\}$. Also here we have moderately different group-level classes. The five different n_k settings yielded entropy-based R-squared values of 0.35, 0.57, 0.71, 0.80, and 0.90 for the 2 class condition, and 0.36, 0.58, 0.73, 0.82, and 0.92 for the 3 class condition. This shows that in our settings separation was very much affected by n_k but not so much by L_3.

In total the simulation study design contained $5 \times 2 \times 2 = 20$ cells which are all possible combinations of the three design factors. For each of these cells we generate 100 data sets. With each data set we estimated multilevel LC models with a fixed number of LCs at the lower-level ($L_2 = 3$) and with different numbers of LCs at the higher-level.

4 Results of the simulation study

As was indicated above, the main goal of the simulation study was to determine which of the investigated model selection measures is preferable for the deciding about the number of higher-level mixture components in multilevel mixture models. For BIC and CAIC, which both have the sample size in their formula, we used two versions, one based on the number of higher-level observations (K) and one based on the total number of lower-level observations (Kn_k).

Table 1 reports the results of our simulation study per design factor aggregated over the other two design factors. For each level of the three design factors and for each investigated fit measure, we indicate the number of simulation replications in which the true number higher-level latent classes was underestimated ($\hat{L}_3 < L_3$), estimated correctly ($\hat{L}_3 = L_3$), and overestimated ($\hat{L}_3 > L_3$).

Let us first have a look at the results for BIC and CAIC using the two different definitions for the sample size. From the results in Table 1, one can easily see that both for BIC and CAIC using the number of groups as sample size is the best option. Underestimation of the number of mixture components it is much more likely with BIC(Kn_k) or CAIC(Kn_k) than with BIC(K) or CAIC(K). This is especially true for the conditions corresponding to low or moderate levels of separation (small or middle n_k values), as well as for the smaller higher-level sample size.

Comparison of the results of all eight investigated fit measures shows that overall AIC3 performs best. The results for BIC(K), CAIC(K), ICOMP are similar in the sense that they perform best when the number of individuals per group (the level of separation) is large enough ($n_k \geq 15$). AIC, on the other hand, performs best when separation is weak ($n_k = 5$) and when the sample size is small. As was found in other studies, AIC3 seems to provide

		n_k					K		L_3		
		5	10	15	20	30	50	500	2	3	*Total*
BIC(Kn_k)	$\hat{L}_3 < L_3$	233	131	67	18	1	400	50	136	314	450
	$\hat{L}_3 = L_3$	167	269	333	382	399	600	950	864	686	1550
	$\hat{L}_3 > L_3$	0	0	0	0	0	0	0	0	0	0
BIC(K)	$\hat{L}_3 < L_3$	199	83	26	6	0	286	28	93	221	314
	$\hat{L}_3 = L_3$	201	317	374	394	399	713	972	907	778	1685
	$\hat{L}_3 > L_3$	0	0	0	0	1	1	0	0	1	1
CAIC(Kn_k)	$\hat{L}_3 < L_3$	253	146	81	33	5	456	62	153	365	518
	$\hat{L}_3 = L_3$	147	254	319	367	395	544	938	847	635	1482
	$\hat{L}_3 > L_3$	0	0	0	0	0	0	0	0	0	0
CAIC(K)	$\hat{L}_3 < L_3$	228	101	46	9	0	337	47	114	270	384
	$\hat{L}_3 = L_3$	172	299	354	391	400	663	953	886	730	1616
	$\hat{L}_3 > L_3$	0	0	0	0	0	0	0	0	0	0
AIC	$\hat{L}_3 < L_3$	103	45	9	3	0	158	5	41	122	163
	$\hat{L}_3 = L_3$	278	320	344	349	355	766	880	853	793	1646
	$\hat{L}_3 > L_3$	16	35	47	48	45	76	115	106	85	191
AIC3	$\hat{L}_3 < L_3$	155	68	13	5	0	236	5	70	171	241
	$\hat{L}_3 = L_3$	245	323	375	389	385	745	972	904	813	1717
	$\hat{L}_3 > L_3$	0	9	12	6	15	19	23	26	16	42
ICOMP	$\hat{L}_3 < L_3$	208	85	20	4	0	274	43	91	226	317
	$\hat{L}_3 = L_3$	191	310	380	392	398	714	957	900	771	1671
	$\hat{L}_3 > L_3$	1	5	0	4	2	12	0	9	3	12
Validation	$\hat{L}_3 < L_3$	78	37	9	1	0	121	4	46	79	125
log-likelihood	$\hat{L}_3 = L_3$	215	239	272	286	291	582	721	691	612	1303
	$\hat{L}_3 > L_3$	107	124	119	113	109	297	275	263	309	572

Table 1. The number of simulation replicates in which the investigated fit measure underestimated, correctly estimated, and overestimated the number of group-level mixture components for each of the three conditions.

a compromise between these two sets of measures [4]. In contrast to our expectations, the validation log-likelihood did not perform very well: it tends to overestimate the number of mixture components under all conditions.

5 Conclusions

Based on the simulation study we can draw two important conclusions. The first concerns the preferred sample size definition in the BIC and CAIC formulae. Our results show clearly that it is much better to use the number of higher-level units as the sample size instead of the total number of lower-level unit. Using the latter makes it much more likely that the number of

mixture components is underestimated, especially if the separation between components is weak or moderate.

The second set of conclusions concern the comparison of all investigated measures. These results are very much in agreement with what is known from simulation studies on standard mixture models. BIC, CAIC, and ICOMP perform very well when the level of separation and the sample size are large enough. In contrast, AIC seems to be the preferable method when the sample size is small and when the level of separation is low. AIC3 offers a good compromise between the tendency of BIC, CAIC, and ICOMP to underestimate the number of mixture components with low separation and small sample sizes and the tendency of AIC to overestimate the number of mixture components with higher separation and large sample sizes.

As in any simulation study, we had to make various choices which limit the range of our conclusions. First of all, we concentrated on selecting the number of classes at the higher level assuming that the number of classes at the lower level is known. Further research is needed to determine whether the same conclusions apply for selecting the number of lower-level classes, or for selecting simultaneously the number of lower- and higher-level classes. Second, we used a classical LC model for binary responses whereas multilevel mixture models can also be used with other types of response variables. Finally, we concentrated on the variant of the multilevel LC model in which only the lower-level class proportions differ across higher-level classes. As was shown when introducing the model, other multilevel LC models may assume that response variables are directly related to the group-level class membership. It seems to be useful to replicate our simulation study for other types of multilevel mixture models, as well as for response variables of other scale types.

References

1. Aitkin, M., 1999. A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **55**, pages 218–234.
2. Bozdogan, H., 1993. *Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-fisher information matrix.* In: Opitz, O., Lausen, B., Klar, R. (Eds.), Information and Classification. Springer, Heidelberg, pages 218–234.
3. Dempster, A.P., Laird, N.M., and Rubin, D.B., 1977. Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* **39**(1), pages 1–38.
4. Dias, J.G., 2006. *Model selection for the binary latent class model: A Monte Carlo simulation.* In: V. Batagelj, H.-H. Bock, A. Ferligoj, A. Žiberna (Eds.), Data science and classification, Springer, Berlin, pages 91–99.
5. Hagenaars, J.A., and McCutcheon, A.L., 2002. *Applied latent class analysis models.* Cambridge University Press.
6. Leroux, B.G., 1992. Consistent estimation of a mixing distribution. *Annals of Statistics* **20**, pages 1350–1360.

7. Nylund, K.L., Muthen, B.O., and Asparouhov, T., 2003. Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal* **14**, pages 535–569.
8. Smyth, D., 2000. Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing* **9**, pages 63–72.
9. Vermunt, J.K., 2003. Multilevel latent class models. *Sociological Methodology* **33**, pages 213–239.
10. Vermunt, J.K., 2004. An EM algorithm for the estimation of parametric and nonparametric hierarchical nonlinear models. *Statistical Neerlandica* **58**, pages 220–233.
11. Vermunt, J.K., 2005. Mixed-effects logistic regression models for indirectly observed outcome variables. *Multilevel Behavioral Research* **40**, pages 281–301.
12. Vermunt, J.K., 2007. A hierarchical mixture model for clustering three-way data sets. *Computational Statistics and Data Analysis* **51**, pages 5368–5376.
13. Vermunt, J.K., 2008. Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research* **17**, pages 33–51.
14. Vermunt, J.K., and Magidson, J., 2008. *LG-Syntax User's Guide: Manual for Latent GOLD 4.5 Syntax Module.* Belmont, MA: Statistical Innovations Inc.