Noname manuscript No.
(will be inserted by the editor)

Beyond the number of classes: separating substantive from non-substantive dependence in latent class analysis

DL Oberski

the date of receipt and acceptance should be inserted later

Abstract Latent class analysis (LCA) for categorical data is a model-based clustering and classification technique applied in a wide range of fields including the social sciences, machine learning, psychiatry, public health, and epidemiology. Its central assumption is conditional independence of the indicators given the latent class, i.e. "local independence"; violations can appear as model misfit, often leading LCA practitioners to increase the number of classes. However, when not all of the local dependence is of substantive scientific interest this leads to two options, that are both problematic: modeling uninterpretable classes, or retaining a lower number of substantive classes but incurring bias in the final results and classifications of interest due to remaining assumption violations.

This paper suggests an alternative procedure, applicable in cases when the number of substantive classes is known in advance, or when substantive interest is otherwise well-defined. I suggest, in such cases, to model substantive local dependencies as additional discrete latent variables, while absorbing nuisance dependencies in additional parameters. An example application to the estimation of misclassification and turnover rates of the decision to vote in elections of 9510 Dutch residents demonstrates the advantages of this procedure relative to increasing the number of classes.

Keywords latent class analysis, local dependence, bivariate residual, score test, information criteria, vote misclassification.

1 Introduction

Latent class (finite mixture) models for categorical variables are applied in a broad range of fields including stratification research in the social sciences (Savage et al., 2013), document classification in machine learning (Hastie et al., 2009), psycho-

DL Oberski
Department of Methodology and Statistics, Tilburg University
Room P1105, PO Box 90153, 5000 LE Tilburg, The Netherlands
E-mail: doberski@uvt.nl

logical measurement (Heinen, 1996), and public health and epidemiology (Collins and Lanza, 2010).

The key assumption of such models is conditional independence of the observed variables given the latent class (mixture component). Violations of this assumption may occur when there are unmodeled latent classes, and a common reaction to detected misfit is therefore to increase the number of classes based on criteria such as L^2, χ^2, (C)AIC, BIC, CVIC or ICL (McLachlan and Peel, 2000, Ch. 6). In response, a literature has developed to aid the researcher in finding the "correct" number of classes (see Nylund et al., 2007; Tofighi and Enders, 2008). As an alternative to deciding the "correct" number of classes, Hennig and Liao (2013) suggested distance-based clustering methods as an exploratory method to find clusters that bring similar observations together, while Anderlucci and Hennig (2014) compared this approach to latent class analysis.

However, not all of the local dependence and pursuant additional latent classes may be of substantive interest to the researcher. For example, Hagenaars and McCutcheon (2002) suggested that local dependence between items in a questionnaire or psychological test can occur because respondents attempt to make their responses consistent; and Oberski and Vermunt (2014) found that ethnicity measurements discussed by Johnson (1990) were locally dependent due to the fact that some were measured on the same occasion. In these instances, additional classes do not yield substantively useful results.

To deal with the problem of non-substantive classes, one might select the number of classes based on their relationship with external, substantively meaningful, variables (Baudry et al., 2014). While this approach does prevent the modeling of non-substantive classes, the nuisance local dependence within the substantive classes remains. Local dependence is still problematic in such cases, because unmodeled local dependencies may bias model parameters of interest as well as posterior classifications (Vacek, 1985; Qu et al., 1996; Hadgu et al., 2005).

This paper demonstrates another alternative to increasing the number of classes: modeling additional discrete latent variables when the dependence between items is substantively interesting, while modeling local dependencies directly when dependence is considered a nuisance. Local fit measures are used to detect conditional dependence, and substantive considerations are then used to decide how detected dependencies should be modeled. The goal of this approach is to deal with the problem of non-substantive classes in latent class analysis while avoiding the bias associated with ignoring nuisance dependencies.

Multiple discrete latent variable models and latent class models with local dependencies have a long history (Harper, 1972; Clogg, 1981; Hagenaars, 1988a,b; Skrondal and Rabe-Hesketh, 2004; Vermunt and Magidson, 2013). However, the circumstances under which it may be preferable to use these techniques rather than increasing the number of classes have never been clarified. This has led the model-based clustering literature to develop methods for selecting the number of classes somewhat separately from the consideration of these alternatives (Hennig and Liao, 2013). This paper therefore aims to reconnect the fields by clarifying the connection between these models and demonstrating the use in data analysis of recently developed local fit measures.

Section 2 introduces the data used to illustrate the suggested approach to modeling local dependence. The latent class model with several discrete variables

and local dependence for binary data is presented in Section 3, together with the use of the "bivariate residual" (BVR) for detecting local dependence. Subsequently, Section 4 demonstrates the advantages of the approach introduced here over simply increasing the number of classes, after which Section 5 concludes.

2 Example application data

Why citizens vote in elections is studied intensively in political science (e.g. Campbell et al., 1960; Franklin, 2004; Gallego and Oberski, 2012). Even so, instead of citizens' actual turnout decisions, the answer to the survey question "did you vote in the last election?" is usually observed. The conclusions of such studies are therefore potentially threatened by misclassification in the answers to this question, and indeed validation studies (see Ansolabehere and Hersh, 2012) have found that respondents are reluctant to admit not having voted. This means that estimating this misclassification so that parameter estimates of substantive interest to political science may be corrected for its biasing effects (Vermunt, 2010) is an important endeavour for the field.

In this application, the goal was therefore to estimate misclassification in voting and turnover of vote decisions between elections by applying latent class analysis to repeated survey measurements. Latent class analysis has the advantage that it can be applied to existing panel surveys in which respondents are asked about their turnout decisions, without requiring difficult-to-obtain administrative data on voting. The disadvantage of latent class models is, however, that they make assumptions of local independence that may be incorrect. We demonstrate how this issue may be dealt with by applying latent class analysis to the LISS panel, a Dutch probability sample of 9510 voters. For more information on the design of the study, response rates, and recruitment efforts, please see Scherpenzeel (2011). All data used in this application are publicly available online (`http://lissdata.nl/`).

The 9510 participants who are eligible to vote were asked whether they had voted in the Parliamentary elections held in the Netherlands in November 2006 (official turnout 80.4%) and June 2010 (turnout 75.4%). Participants were asked whether they had voted in these two elections on five occasions: in 2008, 2009, and 2010 (for the 2006 election), and 2011 and 2012 (for the 2010 election). The percentages of respondents who claimed to have voted were 87%, 84%, 81%, 87%, and 84% respectively. Strikingly, this means that initially reported turnout exceeded actual turnout, possibly due to nonresponse error. But even though the same respondents were asked whether they had voted in the same elections, over time the claimed turnout rate declined toward the actual turnout rates. We therefore suspect that misclassification plays a role that may change over time.

3 Model

3.1 Latent class model with possible local dependencies

Suppose an i.i.d. sample of size N is obtained on J observed binary variables, aggregated by the R response patterns into $\mathbf{Y}$. Let $\mathbf{n}$ be the R-vector of observed response pattern counts. We also postulate K discrete latent variables ξ_k, collected

in a vector $\boldsymbol{\xi}$, whose distribution is to be estimated. The K-way cross-table of $\boldsymbol{\xi}$ yields T unobserved patterns. In the case of latent structure analysis, there is only one discrete latent variable and T will equal the number of latent classes. The log-likelihood for the latent class model is then the discrete mixture (e.g. Formann, 1992)

$$\ell(\boldsymbol{\theta}) = \mathbf{n}' \log \Pr(\mathbf{Y}) = \mathbf{n}' \log \left[\sum_T \Pr(\mathbf{Y}|\xi) \Pr(\boldsymbol{\xi}) \right], \tag{1}$$

where log and exp denote elementwise operations,

$$\Pr(\mathbf{Y}|\boldsymbol{\xi}) = \frac{\exp(\boldsymbol{\eta}_{\mathbf{Y}|\boldsymbol{\xi}})}{\mathbf{1}_R' \exp(\boldsymbol{\eta}_{\mathbf{Y}|\boldsymbol{\xi}})}, \qquad \text{and} \qquad \Pr(\boldsymbol{\xi}) = \frac{\exp(\boldsymbol{\eta}_{\boldsymbol{\xi}})}{\mathbf{1}_T' \exp(\boldsymbol{\eta}_{\boldsymbol{\xi}})}. \tag{2}$$

The GLM linear predictors $\boldsymbol{\eta}_{\mathbf{Y}|\boldsymbol{\xi}}$ and $\boldsymbol{\eta}_{\boldsymbol{\xi}}$ are parameterized using effect-coded design matrices (Evers and Namboodiri, 1979):

$$\boldsymbol{\eta}_{\mathbf{Y}|\boldsymbol{\xi}} = \mathbf{X}_{(Y)}\boldsymbol{\tau} + \mathbf{X}_{(YY)}\boldsymbol{\psi} + \mathbf{X}_{(Y\boldsymbol{\xi})}\boldsymbol{\lambda}, \qquad \text{and} \qquad \boldsymbol{\eta}_{\boldsymbol{\xi}} = \mathbf{X}_{(\boldsymbol{\xi})}\boldsymbol{\alpha} + \mathbf{X}_{(\boldsymbol{\xi}\boldsymbol{\xi})}\boldsymbol{\beta}, \tag{3}$$

where $\mathbf{X}_{(Y)}$, $\mathbf{X}_{(YY)}$ and $\mathbf{X}_{(Y\boldsymbol{\xi})}$ are design matrices for the observed variables' main effects $\boldsymbol{\tau}$, bivariate associations $\boldsymbol{\psi}$, and associations with the latent discrete variables $\boldsymbol{\lambda}$ ("slopes"), respectively. Similarly, $\mathbf{X}_{(\boldsymbol{\xi})}$ and $\mathbf{X}_{(\boldsymbol{\xi}\boldsymbol{\xi})}$ are design matrices for the discrete unobserved variables' main effects $\boldsymbol{\alpha}$ and associations $\boldsymbol{\beta}$. This parameterization of the local dependence latent class model is similar to that adopted by Hagenaars (1988b) and Formann (1992, section 4.3), except that we additionally allow for explicit modeling of multiple discrete latent variables and their interrelations (Magidson and Vermunt, 2001; Vermunt and Magidson, 2013).

For example, with two binary discrete latent variables and choosing "dummy coding", there are four unobserved patterns, $T = 4$, and the main effects and associations design matrices are

$$\mathbf{X}_{(\boldsymbol{\xi})} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} \qquad \text{and} \qquad \mathbf{X}_{(\boldsymbol{\xi}\boldsymbol{\xi})} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \tag{4}$$

The β parameter is then the log-odds ratio in the 2×2 cross-table of the two latent variables. A similar interpretation holds for the λ parameters, while the ψ parameters can be interpreted as conditional log-odds ratios in the cross-tables of the observed variables after conditioning on the latent variables.

The q-vector of parameters $\boldsymbol{\theta}$ can be defined as $\boldsymbol{\theta}' := (\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\tau}', \boldsymbol{\lambda}', \boldsymbol{\psi}')$. There are thus $q \leq T(J+1) - 1 + \binom{J}{2}$ (possible) parameters. The standard local independence latent class model, however, has as its key assumption that $\boldsymbol{\psi} = \mathbf{0}$. In addition, the slopes $\boldsymbol{\lambda}$ are typically restricted such that, given exactly one unobserved discrete variable, each indicator is conditionally independent from all other latent variables; in analogy with linear factor analysis this might be termed "simple structure".

Maximum likelihood estimates of the parameters of the model are usually obtained as $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \mathbb{R}^q} \ell(\boldsymbol{\theta})$ by expectation-maximization (see Formann, 1992), quasi-Newton methods, or a combination of both (Vermunt and Magidson, 2013).

Goodman (1974) showed that the parameters of the model are locally identifiable when the Jacobian $\mathbf{S} := \partial \Pr(\mathbf{Y})/\partial \boldsymbol{\theta}$ is of full column rank. A necessary but not sufficient condition for this is that $R > q$. In practice, local identifiability can be evaluated empirically by examining the rank of the information matrix at the maximum likelihood solution, or by randomly sampling many parameter values in the parameter space and evaluating the information matrix at each point (Forcina, 2008). For a general discussion of identification in latent class models, we refer to Huang and Bandeen-Roche (2004); for a discussion of identifiability of the local dependence parameters, see Oberski and Vermunt (2014, Appendix).

3.2 Model misfit and local dependence

After estimation, for each response pattern expected frequencies $\hat{\mu}_r := N \cdot \Pr(\mathbf{Y}_r | \boldsymbol{\theta} = \hat{\boldsymbol{\theta}})$ are obtained, which can be compared with the observed frequencies n_r. Overall goodness of fit measures based on this comparison such as the chi-square and likelihood ratio (L^2), as well as information criteria such as BIC, AIC, CAIC, CVIC, and ICL are often used to evaluate whether the latent class model adequately describes the observed data (see McLachlan and Peel, 2000, chapter 6).

Since the key assumption is that of local independence ($\boldsymbol{\psi} = \mathbf{0}$), a major source of misfit will be locally dependent item pairs. In our example, local dependence may, for instance, arise because respondents remember their answer on the first measurement occasion and try to remain consistent on later occasions (Hagenaars and McCutcheon, 2002). Assuming the model is overidentified, such local dependence will be picked up by the overall fit statistics and information criteria. When these indicate a problem, additional latent classes are then included in the model to account for the dependence. This will lead to a latent class model in which some of the classes represent, for instance, "consistent answering".

However, local dependencies and the pursuant additional classes are not necessarily of scientific interest. For theoretical reasons, one may prefer a model with fewer classes in the voting data application: we know that respondents have either voted or not and that the measurements pertain to two separate elections. Two classes are also preferred when evaluating diagnostic tests for disease/non-diseased status (Qu et al., 1996).

When a specific number of classes is desired or local dependence is not substantively meaningful, it may be preferable to model local dependencies by freeing elements of $\boldsymbol{\psi}$. Freeing all local dependencies is, however, usually not desirable for reasons of model stability and (sometimes) identifiability (Oberski and Vermunt, 2014). We therefore use the "bivariate residual" (BVR) between item pairs to monitor whether it might be necessary to free local dependencies (Vermunt and Magidson, 2013). The bivariate residual is an intuitively attractive fit index measuring the degree to which the bivariate cross-table between a pair of observed variables fits the model:

$$\text{BVR}_{jj'} := \sum_{k \in \{0,1\}} \sum_{l \in \{0,1\}} \frac{(n_{kl} - \hat{\mu}_{kl})^2}{\hat{\mu}_{kl}} = r_{11}^2 \sum_{k \in \{0,1\}} \sum_{l \in \{0,1\}} \frac{1}{\hat{\mu}_{kl}}, \tag{5}$$

where the raw residuals $r_{kl} := n_{kl} - \hat{\mu}_{kl}$, and n_{kl} and $\hat{\mu}_{kl}$ now indicate observed and expected frequencies in the bivariate 2×2 cross-table of the observed variables

y_j and $y_{j'}$ $(j \neq j')$. The last step is a simplification possible with binary indicators, for which the marginals are perfectly reproduced (Oberski et al., 2013). A BVR can be obtained for each of the $\binom{J}{2}$ pairs of observed variables; in this way, for each pair it can be investigated whether the cross-table between this pair appears to fit the hypothesis of local independence.

The BVR has the same form as a Pearson residual and is often treated in applied research as though its asymptotic distribution converged to a chi-square distribution. Oberski et al. (2013) showed that this is not a good approximation; the BVR is a score test uncorrected for cell interdependencies and far from chi-square distributed. The score test for residual dependencies, which does asymptotically follow a chi-square distribution, was introduced by Oberski and Vermunt (2014) and Oberski and Vermunt (2013) give example applications of its usage. Alternatively, p-values for the BVR very close to Rao (1948)'s classic efficient score test can be obtained by a parametric bootstrap (Efron, 1982; Langeheine et al., 1996). The software Latent Gold 5.0 (Vermunt and Magidson, 2013) implements these procedures. Since there are many item pairs, we will also adjust the obtained p-values for multiple testing using the procedure of Benjamini and Hochberg (1995)[1].

4 Example application results

To demonstrate the approach introduced here, we now follow two procedures for data analysis of the Dutch voting example. The first procedure is a standard single nominal latent class model, which is fitted to the data with an increasing number of classes. BIC and CAIC are used to select the number of classes, after which these are interpreted. We compare this standard procedure with one in which two discrete latent variables are modeled jointly, one for voting in each of the 2006 and 2010 elections, and the bivariate residuals are inspected to decide which local dependencies should be freed. The substantive interest of a typical political scientist would focus here on true voting behavior and its relationship with other variables, rather than measured voting behavior.

Figure 1 shows criteria used to select the number of classes. Both BIC and CAIC select the four-class model. When this model is fit to the five claims of having voted, the conditional probabilities shown in Figure 2 result. The left-hand side of Figure 2 shows the probability of claiming to have voted on each of the five measurement occasions given the four latent classes, indicated by the different lines (colors, point shapes). The right-hand side of Figure 2 provides a legend and shows class size estimates with 2 s.e. error bars. Figure 2 shows that class 1 is the class of people who voted in both elections, while class 3 is voting in neither election. Class 4 appears to represent voting in 2010 but not in 2006, although the probability of claiming to have voted in 2006 in this class is still around 0.25. Class 2, containing 10% of observations, is the most difficult to explain; it contains people who initially claim to have voted, but, as time goes by, become more likely to admit that they did not.

The standard latent class model procedure applied to these data is somewhat unsatisfactory. Considering that there are only two actual elections, the only latent

[1] We thank the participants of the 2013 meeting of the Italian Statistical Society in Brescia for this suggestion.

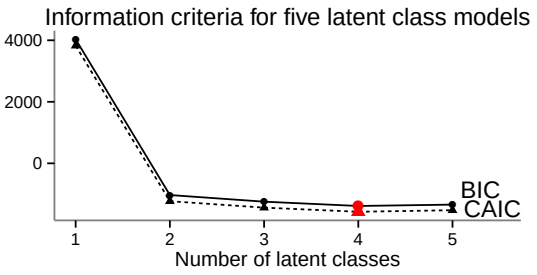


Fig. 1 Model selection increasing the number of classes. Using both BIC and CAIC the four-class solution would be selected.

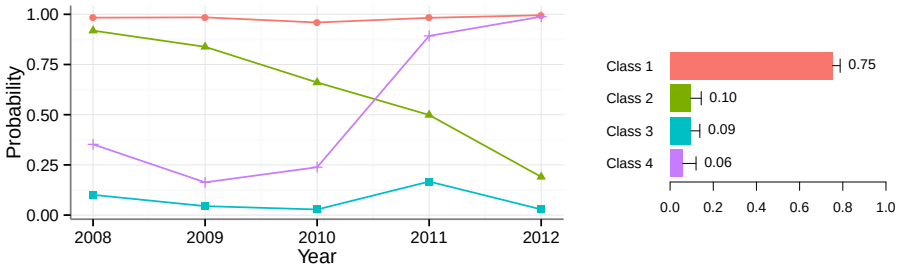
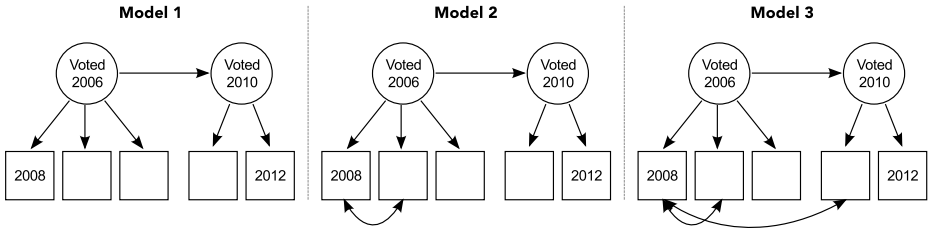


Fig. 2 Left: probability profile plot for the four-class solution. Right: legend with estimated class sizes and 2 s.e. error bars.

classes that represent the "true voting" variable of substantive interest to political scientists would correspond to the $2 \times 2 = 4$ cells in the cross-table of voting or not in 2006 and 2010. Four classes are indeed selected, but instead of a class "voting in 2006 and not in 2010", the difficult-to-interpret class 2 results, which partially also represents artefacts that are not of interest to political scientists.

An alternative procedure is to fit a model with two discrete latent variables, one for each election, each with two classes (voted/did not vote). The first three answers, being about the 2006 elections, are related to the first latent variable and the last two answers, about the 2010 elections, to the second latent variable. Conditional probabilities then represent misclassification rates with respect to true turnout in the 2006 and 2010 elections, which is the question of scientific interest.

Initially a model is fit in which all $\binom{5}{2} = 10$ possible local dependencies are set to zero. This "Model 1" is shown as a graph in Figure 3. The table under "Model 1" in Figure 3 provides p-values for the 10 bivariate residuals obtained by parametric bootstrapping. All p-values have been adjusted for multiple testing using the procedure of Benjamini and Hochberg (1995). The BVR's of the dependence between answers in 2008 and in other years correspond to Hagenaars and McCutcheon (2002)'s suggestion that respondents sometimes attempt to make their answers consistent with the first occasion. Based on these and the values of the BVR's (not shown for conciseness), we free the local dependence between the answer in 2008 and in 2009 and re-fit the model to obtain the model and BVR p-values shown under "Model 2". One adjusted p-value is then still < 0.01 and in line with the memory effect theory: the corresponding dependence is therefore freed. The final model ("Model 3") does not have any BVR with adjusted bootstrapped



'08	*'09*	*'10*	*'11*
0.00			
0.73	0.28		
0.00	0.82	0.02	
0.63	0.28	0.00	0.79

Boot'd L^2 p-value: 0.000.

'08	*'09*	*'10*	*'11*
0.79			
0.28	0.11		
0.00	0.16	0.51	
0.09	0.02	0.11	0.79

L^2 p-value: 0.000.

'08	*'09*	*'10*	*'11*
0.75			
0.28	0.43		
0.72	0.28	0.39	
0.44	0.11	0.39	0.82

L^2 p-value: 0.052.

Fig. 3 Top: three sequential models, starting from the conditional independence model (Model 1). Bottom: False Discovery Rate-adjusted bootstrapped p-values for the bivariate residuals and bootstrapped p-value for L^2 under each model.

p-value < 0.01. The overall bootstrapped likelihood ratio test L^2 indicates a good fit as well.

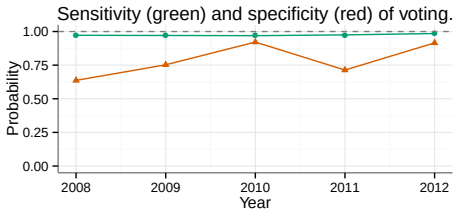
This final model has several advantages over the four-class model. First, it explicitly models true turnout in the two elections so that the conditional probabilities may be interpreted as misclassification rates ("specificity" and "sensitivity"). These misclassification rates are of interest to political scientists. Second, the two-variable classification allows researchers to relate voting in these two elections to external variables (Vermunt, 2010). Third, nuisance local dependencies such as memory effects are not part of the classification but are accounted for by local dependence parameters.

Sensitivity and specificity (misclassification rates) are shown on the left-hand side of Figure 4. The Figure shows that the probability of a respondent claiming to have voted when they have not decreases as the election period becomes more distant. This finding corresponds to the idea that false positives are due to social desirability, since the "norm" of voting will be less salient three years after the election than during election season[2]. This pattern explains the overall pattern that claimed turnout rates approached the actual turnout rates as time goes by[3].

An interesting external validation of our analysis is to compare the latent class model estimates of misclassification with those obtained using administrative data. Based on comparing US vote validation data with the National Election Study, Ansolabehere and Hersh (2012, Table 1 on p. 446) report the false negative rate as between 0.002 and 0.012, whereas the latent class model used here estimates it at 0.019 and 0.024 for 2006 and 2010 respectively. Similarly, the true negatives from validation were between 0.723 and 0.747, while we have estimated them at 0.739 and 0.780. The estimates obtained here, even though they do not use expensive validation data and are for a Dutch rather than a US population, were therefore very close to other results in the literature.

[2] Note that 2010 was not known in advance to be an election season by respondents since elections were called due to the sudden collapse of the government.

[3] As noted by a reviewer, the four-class solution will of course also fit this overall pattern. However, how this "explanation" connects to a social-scientific theory is, in my opinion, unclear.



	Voted in 2006		
	No	Yes	
2010			
No	0.713	0.051	18%
Yes	0.287	0.949	82%
	19%	81%	

Fig. 4 Left: probability profile (i.e. sensitivity and specificity) plot for the two-class, two-variable solution. Right: turnover table of "true vote" classes.

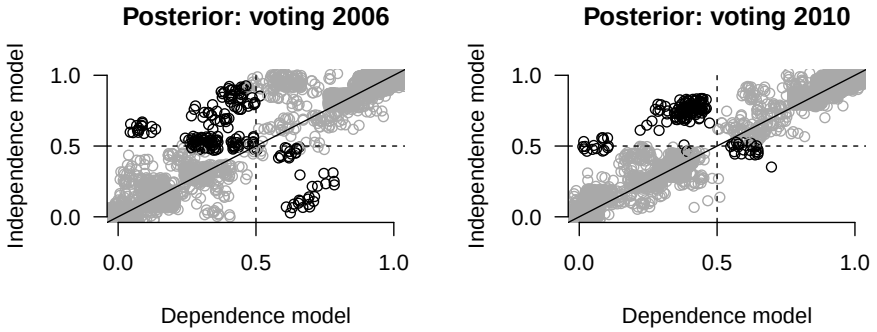


Fig. 5 Jittered posterior probabilities of voting in 2006 and voting in 2010 under the local dependence model (Model 3) versus local independence model (Model 1). The darker observations in off-diagonal quadrants are classified differently (before jittering) depending on the model.

The right-hand side of Figure 4 shows the estimated turnover table of true turnout from 2006 to 2010 with class sizes in the margins. The class prevalences of 81 and 82 percent are higher than the actual turnout rates 80.4 and 75.4 percent, although they are much closer to true turnout than the raw reported rates (around 87%). The turnover table suggests that voters mostly remained voters whereas non-voters in 2006 had a chance of 0.287 of voting in the 2010 election. If such a pattern were to be predicted for future elections, it would suggest that efforts to encourage citizens to vote would be best focused on non-voters in previous elections.

Finally, Figure 5, which shows the posterior classifications under two different models (Model 1 vs. Model 3 in Figure 3), demonstrates that ignoring the local dependencies may lead to bias. Posterior classifications shown in Figure 5 are different for the two latent class variables depending on whether the local dependencies are taken into account or not, potentially biasing subsequent analyses of the classifications. This shows that simply ignoring non-substantive local dependence is not an attractive option in this case.

5 Summary

Latent class analysis often involves selecting the number of classes. Several approaches to do this have been suggested in the literature, focusing on the purely statistical concerns of balancing model complexity and model fit (for an exception, see Hennig and Liao, 2013). Since the identifying assumption of latent class models is local dependence, this means data dependencies that fail to be predicted by the local dependence model are absorbed as additional classes. This approach can be expected to work adequately (e.g. Nylund et al., 2007) when the model is correctly specified and all latent classes are of substantive interest. It can also be useful when the analysis is of an entirely exploratory nature and the researcher simply wishes to determine groups that are maximally "different" by some criterion (Anderlucci and Hennig, 2014).

This paper demonstrated a possible problem with this approach when not all of the model fit violations are of substantive interest. An application demonstrated that it may sometimes be more advantageous to model substantively interesting local dependence as additional discrete latent variables, while modeling nuisance dependencies using additional local dependence parameters rather than additional classes. When estimating misclassification and turnover rates of the decision to vote in an election, increasing the number of classes led to better fit but uninterpretable classes, whereas retaining a lower number of classes or otherwise ignoring the non-substantive local dependence led to bias. The alternative procedure suggested here yielded a model that was better-interpretable in the sense that the results corresponded more directly to those of substantive interest to most political scientists. Moreover, the resulting estimates were close to those from independent external validation studies. Local fit measures such as the bivariate residual or the score test (Oberski and Vermunt, 2013) can be used to guide in this procedure.

Although the BVR used here to detect local dependence is closely connected to the score test for local independence, it can be seen as a general measure of model misfit. It could therefore also be used to guide decisions on the number of classes when increasing classes is of interest. In other words, the decision to increase the number of classes or follow the procedure suggested here is not a statistical issue but a theoretical one that should be based on the substantive interest of the researcher.

The approach suggested here does have several limitations. First, it is inapplicable when the substantive interest is not well defined. In such more exploratory cases, increasing the number of classes may be more attractive. Second, when allowing for local dependencies, there is a certain danger that substantively interesting dependencies are inadvertently modeled as nuisance dependencies. In the example application, a survey methodologist's interest might focus exactly on the "consistent answering", for instance. In other words, a consequence of the procedure is that the model selected depends on the goal of the analysis, which must be kept in mind when introducing local dependence parameters. Overall, however, many data analyses in model-based clustering and classification may be more amenable to the approach discussed here than to an increase in the number of classes.

Acknowledgements

The author was supported by the Netherlands Organization for Scientific Research (NWO) [Veni grant number 451-14-017].

References

Anderlucci, L. and Hennig, C. (2014). The clustering of categorical data: A comparison of a model-based and a distance-based approach. *Communications in Statistics-Theory and Methods*, 43(4):704–721.

Ansolabehere, S. and Hersh, E. (2012). Validation: What big data reveal about survey misreporting and the real electorate. *Political Analysis*, 20(4):437–459.

Baudry, J.-P., Cardoso, M., Celeux, G., Amorim, M. J., and Ferreira, A. S. (2014). Enhancing the selection of a model-based clustering with external categorical variables. *Advances in Data Analysis and Classification*. DOI 10.1007/s11634-014-0177-3.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Campbell, A., Converse, P., Miller, W., and Stokes, D. (1960). *The American Voter*. Wiley, New York.

Clogg, C. C. (1981). New developments in latent structure analysis. In Jackson, D. and Borgatta, E., editors, *Factor analysis and measurement in sociological research*, pages 215–246. Sage, Beverly Hills, CA.

Collins, L. M. and Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. Wiley, New York.

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. *CBMS-NSF Regional Conference Series in Applied Mathematics*, 38.

Evers, M. and Namboodiri, N. (1979). On the design matrix strategy in the analysis of categorical data. *Sociological Methodology*, 10:86–111.

Forcina, A. (2008). Identifiability of extended latent class models with individual covariates. *Computational Statistics & Data Analysis*, 52(12):5263–5268.

Formann, A. (1992). Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association*, 87(418):476–486.

Franklin, M. (2004). *Voter turnout and the dynamics of electoral competition in established democracies since 1945*. Cambridge University Press, New York.

Gallego, A. and Oberski, D. (2012). Personality and political participation: The mediation hypothesis. *Political Behavior*, 34:424–451.

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215.

Hadgu, A., Dendukuri, N., and Hilden, J. (2005). Evaluation of nucleic acid amplification tests in the absence of a perfect gold-standard test: a review of the statistical and epidemiologic issues. *Epidemiology*, 16(5):604–612.

Hagenaars, J. A. (1988a). *LCAG-loglinear modelling with latent variables: A modified LISREL approach*, volume 2. Sociometric research foundation, Amsterdam.

Hagenaars, J. A. P. (1988b). Latent structure models with direct effects between indicators: Local dependence models. *Sociological Methods & Research*, 16(3):379–405.

Hagenaars, J. A. P. and McCutcheon, A. (2002). *Applied latent class analysis*. Cambridge University Press, Cambridge, UK.

Harper, D. (1972). Local dependence latent structure models. *Psychometrika*, 37(1):53–59.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences*. Sage, Thousand Oaks, CA.

Hennig, C. and Liao, T. F. (2013). How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(3):309–369.

Huang, G. and Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika*, 69(1):5–32.

Johnson, R. (1990). Measurement of Hispanic ethnicity in the US census: An evaluation based on latent-class analysis. *Journal of the American Statistical Association*, 85(409):58–65.

Langeheine, R., Pannekoek, J., and Van de Pol, F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods & Research*, 24(4):492–516.

Magidson, J. and Vermunt, J. K. (2001). Latent class factor and cluster models, bi-plots, and related graphical displays. *Sociological Methodology*, pages 223–264.

McLachlan, G. and Peel, D. (2000). *Finite mixture models*. Wiley, New York.

Nylund, K. L., Asparouhov, T., and Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. *Structural equation modeling*, 14(4):535–569.

Oberski, D., Van Kollenburg, G., and Vermunt, J. (2013). A Monte Carlo evaluation of three methods to detect local dependence in binary data latent class models. *Advances in Data Analysis and Classification*, 7(3).

Oberski, D. and Vermunt, J. (2013). A model-based approach to goodness-of-fit evaluation in item response theory. *Measurement: Interdisciplinary Research & Perspectives*, 11:117–122.

Oberski, D. and Vermunt, J. (2014). The Expected Parameter Change (EPC) for local dependence assessment in binary data latent class models. *Accepted for publication in Psychometrika*. [Obtained from http://daob.nl/wp-content/uploads/2013/08/lca-epc-revision-nonblinded.pdf].

Qu, Y., Tan, M., and Kutner, M. (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics*, 52(3):797–810.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society*, 44(1):50–57.

Savage, M., Devine, F., Cunningham, N., Taylor, M., Li, Y., Hjellbrekke, J., Le Roux, B., Friedman, S., and Miles, A. (2013). A new model of social class? findings from the BBC's Great British Class Survey Experiment. *Sociology*, 47(2):219–250.

Scherpenzeel, A. (2011). Data collection in a probability-based internet panel: How the LISS panel was built and how it can be used. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 109(1):56–61.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling : multilevel, longitudinal, and structural equation models*. Interdisciplinary statistics series. Chapman & Hall/CRC, Boca Raton, FL.

Tofighi, D. and Enders, C. K. (2008). Identifying the correct number of classes in growth mixture models. In Hancock, G. R. and Samuelsen, K. M., editors, *Advances in latent variable mixture models*, pages 317–341. Information Age, Charlotte, NC.

Vacek, P. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics*, 41(4):959–968.

Vermunt, J. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, 18:450–469.

Vermunt, J. K. and Magidson, J. (2013). *Technical guide for Latent GOLD 5.0: Basic and advanced*. Statistical Innovations Inc., Belmont, MA.