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

A Monte Carlo evaluation of three methods to detect local dependence in binary data latent class models

Daniel L. Oberski · Geert H. van Kollenburg · Jeroen K. Vermunt

the date of receipt and acceptance should be inserted later

Abstract Binary data latent class analysis is a form of model-based clustering applied in a wide range of fields. A central assumption of this model is that of conditional independence of responses given latent class membership, often referred to as the "local independence" assumption. The results of latent class analysis may be severely biased when this crucial assumption is violated; investigating the degree to which bivariate relationships between observed variables fit this hypothesis therefore provides vital information. This article evaluates three methods of doing so. The first is the commonly applied method of referring the so-called "bivariate residuals" to a chi-square distribution. We also introduce two alternative methods that are novel to the investigation of local dependence in latent class analysis: bootstrapping the bivariate residuals, and the asymptotic score test or "modification index". A Monte Carlo simulation indicates that the latter two methods perform adequately, while the first method does not perform as intended.

D. L. Oberski (corresponding author) · G. H. van Kollenburg · J. K. Vermunt
Department of Methodology and Statistics, Tilburg University, The Netherlands
P.O. Box 90153, 5000 LE Tilburg, The Netherlands
Ph.: +31 13 466 2959, Fax: +31 13 466 3002
E-mail: doberski@uvt.nl

G. H. van Kollenburg
E-mail: g.h.vankollenburg@uvt.nl

J. K. Vermunt
E-mail: j.k.vermunt@uvt.nl

1 Introduction

Latent class models for binary (dichotomous) variables are discrete finite mixtures of binomials, in which a key assumption is that of conditional independence given the latent class or mixture component (Agresti, 2002). These models have found applications in a broad range of fields including the social sciences (Hagenaars and McCutcheon, 2002), machine learning (Hofmann, 2001), psychological measurement (Heinen, 1996), and the biomedical sciences (Faraone and Tsuang, 1994). Violations of the crucial conditional independence assumption are known to severely bias outcomes of interest; an issue noted particularly in the biomedical context of estimation of sensitivity and specificity of diagnostic tests by Vacek (1985), Torrance-Rynard and Walter (1998), Walter and Irwig (1988), Albert and Dodd (2004), and Hadgu et al. (2005). It is therefore important in latent class analysis to monitor model fit; particularly useful is the possibility of assessing the source of the misfit in terms of the residual dependence between observed variables.

One measure of such residual dependence between observed variables is the *bivariate residual* (BVR). The BVR for a pair of observed variables is defined as the Pearson residual in their bivariate cross-table (Vermunt and Magidson, 2005, pp. 72-3). The BVR is used as an overall measure of model fit, or as a flag of potentially problematic restrictions; see, for example, the application to item bias in psychological measurement described by Tay et al. (2011). The idea behind this measure is then that a "high" BVR value for a pair of variables indicates residual local dependency that causes model misfit. Though intuitively appealing and straightforward to compute, the BVR has the drawback that its asymptotic distribution is not known. While Vermunt and Magidson claim that BVR's are "Lagrange-type chi-squared statistics" (p. 73), the BVR will not typically follow a chi-squared distribution.

In spite of the fact that the BVR is not asymptotically distributed as chi-square, in applied research it is frequently interpreted as though it were, so that "high" BVR's are judged to be those exceeding the quantiles of a chi-square distribution.

See, for example, Nyholt et al. (2004, p. 235); Gaffikin et al. (2007, p. 3); Chen et al. (2007, p. 4); Baughman et al. (2008, p. 110); Hybels et al. (2009, p. 5); and Gallego and Oberski (2012, p. 440). To our knowledge, not much is known about the practical consequences of this practice. As a more principled alternative, one might consider the parametric bootstrap (Efron, 1982; Langeheine et al., 1996) to obtain p-values for the BVR. Alternatively, Rao (1948)'s classical score (or "Lagrange multiplier") test, sometimes called "modification index" (MI), also provides a test for local dependence that is asymptotically chi-square distributed under the null hypothesis.

This article evaluates three approaches to the evaluation of local dependencies in binary data latent class models: 1) referring the BVR to a chi-square distribution with one degree of freedom, 2) obtaining a p-value for the BVR by a parametric bootstrap, and 3) the score test (MI). The latter two methods are novel to the investigation of local dependence in latent class analysis. We evaluate the behavior of these three methods by Monte Carlo simulation under the null hypothesis and under various conditions that violate conditional independence. The results of the simulation provide guidance for applied researchers on appropriate ways of applying the BVR or MI to evaluate the model assumptions of latent class analysis.

The remainder of this article is structured as follows. The following section introduces the latent class model. Subsequently, the BVR as a measure of the fit of bivariate observed cross-tables to the hypothesis of local independence is introduced. We propose two additional methods of assessing the source of model misfit: parametric bootstrap p-values for the BVR, and the modification index (score test). The three methods are then evaluated under a range of conditions using Monte Carlo simulation. The final section provides concluding remarks.

2 The latent class model

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. The log-likelihood for the latent class model with T classes for the unobserved discrete variable ξ can then be formulated (Formann, 1992) as the discrete mixture GLM

$$\ell(\boldsymbol{\theta}) = \mathbf{n}' \log \Pr(\mathbf{Y}) = \mathbf{n}' \log \left[\sum_{t=1}^{T} \Pr(\xi = t) \left(\frac{\exp(\boldsymbol{\eta}_t)}{\mathbf{1}_R' \exp(\boldsymbol{\eta}_t)} \right) \right], \tag{1}$$

where log and exp denote elementwise operations, $\Pr(\xi = t) = \exp(\alpha_t)/\mathbf{1}_T' \exp(\boldsymbol{\alpha})$, and

$$\boldsymbol{\eta}_t = \mathbf{X}_{(Y)}\boldsymbol{\tau} + \mathbf{X}_{(YY)}\boldsymbol{\psi} + \mathbf{X}_{(Y\xi_t)}\boldsymbol{\lambda}, \tag{2}$$

where $\mathbf{X}_{(Y)}$, $\mathbf{X}_{(YY)}$ and $\mathbf{X}_{(Y\xi_t)}$ are design matrices for the observed variables' main effects $\boldsymbol{\tau}$, bivariate associations $\boldsymbol{\psi}$, and associations with the latent class variable $\boldsymbol{\lambda}$, respectively (Evers and Namboodiri, 1979). The vector $\boldsymbol{\alpha}$ contains the logistic main effect parameters for the latent class proportions. This parameterization of the local dependence latent class model is similar to that adopted by Hagenaars (1988) and Formann (1992, section 4.3). The q-vector of parameters $\boldsymbol{\theta}$ can be defined as $\boldsymbol{\theta}' := (\boldsymbol{\alpha}', \boldsymbol{\tau}', \boldsymbol{\lambda}', \boldsymbol{\psi}')$. The standard local independence latent class model, however, has as its key assumption that $\boldsymbol{\psi} = \mathbf{0}$, so that $\boldsymbol{\theta}' := (\boldsymbol{\alpha}', \boldsymbol{\tau}', \boldsymbol{\lambda}')$ constitutes the free parameter vector.

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, 2005). 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 the number of unique response patterns R at least equal the number of parameters q. Thus, for instance, the two-class model for two binary observed variables is not identified, nor is the three class model for

three binary observed variables. That the condition is not sufficient is evidenced by the three class model for four binary observed variables, which has one degree of freedom but is nevertheless not identified. 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 more detailed discussion of identification, we refer to Huang and Bandeen-Roche (2004).

3 Evaluating the source of model misfit

Goodness-of-fit of the latent class model to the data is often evaluated with statistics based on or derived from the χ^2 statistic

$$\chi^2 := \sum_{r \in 1..R} \frac{(n_r - \hat{\mu}_r)^2}{\hat{\mu}_r}, \tag{3}$$

where R is the number of unique response patterns, n_r is the number of observations for a response pattern r, and $\hat{\mu}_r := N \cdot \Pr(\mathbf{Y}_r|\boldsymbol{\theta} = \hat{\boldsymbol{\theta}})$ is the model-based expectation of n_r evaluated at the maximum likelihood solution. As suggested by the name of this statistic, when the model holds, its asymptotic distribution (as the sample size N approaches infinity) converges to a chi-square distribution with $R - q$ degrees of freedom (e.g. Maydeu-Olivares and Joe, 2005). Among other statistics in common use are the likelihood ratio, AIC, and BIC (e.g. McLachlan and Peel, 2000).

The χ^2 statistic gives an indication of overall model misfit, but it does not aid the researcher in detecting the source of the misfit. Since the key assumption is that of local independence, an intuitively attractive fit index measuring the degree to which the bivariate cross-table between a pair of observed variables fits the

model is the "bivariate residual" (BVR),

$$\text{BVR}_{jj'} := \sum_{k\in\{0,1\}} \sum_{l\in\{0,1\}} \frac{(n_{kl} - \hat{\mu}_{kl})^2}{\hat{\mu}_{kl}} = \sum_{k\in\{0,1\}} \sum_{l\in\{0,1\}} \frac{r_{kl}^2}{\hat{\mu}_{kl}} \tag{4}$$

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 $y_{j'}$ $(j \neq j')$. Since the marginals are perfectly reproduced by the latent class model, all residuals are equal in absolute value, so that the BVR reduces to

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

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 or not.

The BVR has the same form as a Pearson residual and is often treated in applied research as though its asymptotic distribution (as the sample size N approaches infinity) converged to a chi-square distribution. However, this is not the case because the expected frequencies are not independent; the conditioning of the $\hat{\mu}_{kl}$ on parameters not used in the estimation of the bivariate cross-table is disregarded. Most applied studies ignore this issue and refer the BVR to a chi-square distribution. An alternative method is to instead refer the BVR to its empirical distribution as obtained from a parametric bootstrap (Efron, 1982; Langeheine et al., 1996). We propose using the parametric bootstrap as a better approach to obtaining p-values for the BVR statistic.

A different approach to evaluating the hypothesis of local independence is the classical score test (Rao, 1948). This approach was discussed in the context of item response theory by Glas (1999) and van der Linden and Glas (2010). Define the score $s_{jj'}$ for a local dependence between the observed variables y_j and $y_{j'}$ $(j \neq j')$ as $s_{jj'} := \partial\ell(\boldsymbol{\theta})/\partial\psi_{jj'}$, where $\psi_{jj'}$ is the element of $\boldsymbol{\psi}$ corresponding to the two-way interaction between y_j and $y_{j'}$. Let the hessian $\mathbf{H} := \partial^2\ell(\boldsymbol{\theta})/\partial(\psi_{jj'}, \boldsymbol{\theta}')'\partial(\psi_{jj'}, \boldsymbol{\theta}')$.

Then the "modification index" (MI) allowing a test for the presence of a local dependence between y_j and $y_{j'}$ conditional on the latent class is the statistic

$$\text{MI} := \frac{s_{jj'}^2}{\text{Var}(s_{jj'})} = \frac{s_{jj'}^2}{h_{\psi\psi} - \mathbf{H}_{\psi\boldsymbol{\theta}}\mathbf{H}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\mathbf{H}_{\psi\boldsymbol{\theta}}'}, \tag{6}$$

where $h_{\psi\psi}$, $\mathbf{H}_{\psi\boldsymbol{\theta}}$, and $\mathbf{H}_{\boldsymbol{\theta}\boldsymbol{\theta}}$ denote submatrices of the hessian (e.g. Sörbom, 1989), and it is assumed that there is at least one degree of freedom. A score test can be constructed by replacing $\boldsymbol{\theta}$ by consistent estimates such as $\hat{\boldsymbol{\theta}}$. Under the null hypothesis that $\psi_{jj'} = 0$, the MI then asymptotically (as $N \to \infty$) approaches a chi-square distribution with one degree of freedom. Furthermore, as long as the alternative model is not strongly misspecified, when $\psi_{jj'} \neq 0$ so that local independence is violated, the MI approaches a noncentral chi-square distribution with noncentrality parameter (ncp) equal to the population improvement in χ^2 (equation 3) obtained by freeing $\psi_{jj'}$ (Satorra, 1989). An advantage of the score test based on the MI relative to bootstrapping the BVR is that its computation does not require resampling methods and is therefore preferable when computational convenience is an issue.

The MI statistic defined in equation 6 may appear rather different from the BVR defined in equation 4. However, there is a strong connection between these two statistics since $s_{jj'} = 4r_{11}$. To see this, let $\mathbf{x}_{(y_j y_{j'})}$ be the design vector corresponding to $\psi_{jj'}$, i.e. an R-vector that equals +1 when $y_{rj} = y_{rj'}$ and -1 when $y_{rj} \neq y_{rj'}$. Then, by differentiating equation 1, we obtain

$$\begin{aligned} s_{jj'} = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \psi_{jj'}} &= \mathbf{n}' \sum_{t \in 1..T} \Pr(\xi = t|\mathbf{Y})[\mathbf{x}_{(y_j y_{j'})} - \mathbf{x}'_{(y_j y_{j'})}\Pr(\mathbf{Y}|\xi)] \\ &= \mathbf{n}'[\mathbf{x}_{(y_j y_{j'})} - \mathbf{x}'_{(y_j y_{j'})}\Pr(\mathbf{Y})] \\ &= \sum_{k=l}(n_{kl} - \hat{\mu}_{kl}) - \sum_{k \neq l}(n_{kl} - \hat{\mu}_{kl}) = 4r_{11} \end{aligned} \tag{7}$$

where the second step is due to the fact that $\psi_{jj'}$ is class-independent, and the last step follows because the off-diagonal residuals have a sign opposite to the diagonal residuals. If dummy coding is chosen instead of effect coding, $s_{jj'} = r_{11}$ and the

variance $\text{Var}(s_{jj'})$ is scaled accordingly. The difference between the BVR and the MI for introducing a class-independent local dependence parameter is therefore that the BVR does not take the dependency between the expected cell frequencies into account, while the MI uses the correct asymptotic variance. This makes the MI an attractive alternative to the more intuitively defined BVR statistic.

4 Monte Carlo simulation

4.1 Study design

The performance of the three procedures was evaluated by Monte Carlo simulation under the conditions resulting from fully crossing the following factors:

- Loglinear effect of latent on observed variables ("loadings") $\lambda \in \{.50, .80\}$;
- Local dependence between last two items $\psi \in \{-0.4, -0.2, -0.05, 0, +0.05, +0.2, +0.4\}$;
- Sample size $N \in \{200, 500, 1000, 5000\}$.

This $2 \times 7 \times 4$ design yields 56 conditions, in 8 of which the null hypothesis holds ($\psi = 0$), and 48 of which violate local independence to various degrees ($\psi \neq 0$).

Under each condition, 200 samples of size N were drawn from a two-class population with five binary observed variables. The latent and observed variable intercepts were set to zero, meaning 50% of the observations fell in either class. Conditional on the latent class, the last two observed variables were locally dependent (except when $\psi = 0$). In each of the 200 samples, the MI and BVR were calculated. Subsequently, a parametric bootstrap of the BVR with 500 replicates was performed conditional on the sample parameter estimates. We then obtained p-values for the BVR by 1) Referring the sample BVR to a chi-square distribution ("naive" p-value), 2) Referring the sample BVR to its bootstrapped empirical distribution, and 3) Referring the MI to a chi-square distribution. All analyses used R version 2.15.2 (R Core Team, 2012), while an experimental version of the software Latent Gold (Vermunt and Magidson, 2005) was used to obtain the bootstrapped p-values and check the results.

4.2 Results

We first consider the performance of the BVR and MI under the eight conditions in which the null hypothesis holds.

Condition			α for nominal 5%			Empirical distribution			
			BVR			MI		BVR	
λ	ψ	n	Naive	Boot	MI	Mean	Var	Mean	Var
0.5	0	200	0.000	0.050	0.051	0.97	1.7	0.33	0.2
0.5	0	500	0.000	0.020	0.050	1.06	2.3	0.36	0.2
0.5	0	1000	0.000	0.060	0.065	0.96	1.9	0.33	0.2
0.5	0	5000	0.000	0.085	0.055	0.97	2.0	0.34	0.2
0.8	0	200	0.000	0.065	0.040	1.04	1.7	0.25	0.1
0.8	0	500	0.000	0.070	0.060	1.05	2.0	0.25	0.1
0.8	0	1000	0.000	0.060	0.090	1.22	2.6	0.30	0.2
0.8	0	5000	0.000	0.035	0.060	1.16	3.1	0.28	0.2

Table 1 Rejection rates with a nominal α-level of 5%, and empirical distribution of BVR and MI under the null hypothesis.

Under the null hypothesis and choosing a nominal α-level (probability of type-I error) of 5%, approximately 5% of the 200 simulated p-values should be smaller than 0.05. Table 1 shows that this is approximately the case for the bootstrap p-values for the BVR and the asymptotic p-value of the MI. The naive p-value which refers the BVR to a chi-square distribution, however, does not provide the nominal α-level; in fact, the null hypothesis was not rejected in any of the 200 simulated samples.

The last four columns of Table 1 compare the empirical distribution of the BVR and MI to that of a chi-square distribution with one degree of freedom, which has mean equal to unity and variance 2. Table 1 clearly shows that the empirical distribution of the MI conforms to this expectation but the BVR uniformly has both mean and variance that are too low. For each of the eight conditions, a Kolmogorov-Smirnoff test of the hypothesis that the MI's are sampled from a central chi-square distribution with one degree of freedom produces p-values larger than 0.15, while the same test for the BVR yields p-values smaller than 10^{-10}. It is therefore clear that under the null hypothesis, the MI appears to follow this

asymptotic distribution closely under all conditions, while the BVR does not follow this distribution under any condition.

The null hypothesis did not hold in the 48 conditions with local dependencies ψ not equal to zero. Figure 1 plots the power to reject the null hypothesis under these conditions as a function of the (log) sample size. In each plot, the three lines correspond to the three methods studied here: the solid red line indicates the power of the naive BVR p-values, the dotted green line the power of the bootstrap p-values, and the dashed blue line the power of the asymptotic p-values for the MI. The twelve different plots correspond to the twelve conditions resulting from crossing the local dependence size (rows) with the loadings (columns).

A striking feature of Figure 1 is that all three procedures have a rather low power to detect small (± 0.05) to medium (± 0.2) local dependencies, especially when the loadings are large. The MI and bootstrap provide adequate power to detect dependencies of ± 0.2 or stronger when the sample size is 1000 or above (here we define adequate power to mean a power larger than 0.8). The figure shows that the power of the naive BVR p-values to detect the varying sizes of local dependence is uniformly lower than that of the other two procedures. Furthermore, this power is only satisfactory when the local dependence is -0.4. When the local dependence is -0.2, +0.2, or +0.4, a sample of at least 5000 is needed to attain adequate power. The smaller local dependencies of ± 0.05 are almost never detected by the naive BVR p-value.

Both the bootstrap p-values for the BVR and the asymptotic p-values for the MI perform much better than the naive BVR p-values. As shown in Figure 1, these two alternative procedures generally yield similar power, except in the case of low loadings and large local dependencies. In these conditions the MI test appears to be more powerful than a parametric bootstrap of the BVR. This difference is most pronounced in the condition with a large positive local dependence of +0.4 and low loadings (lower-left graph in Figure 1): with a sample size of 200, the power of the bootstrap is about 0.25 while the MI test yields a power of about 0.50.

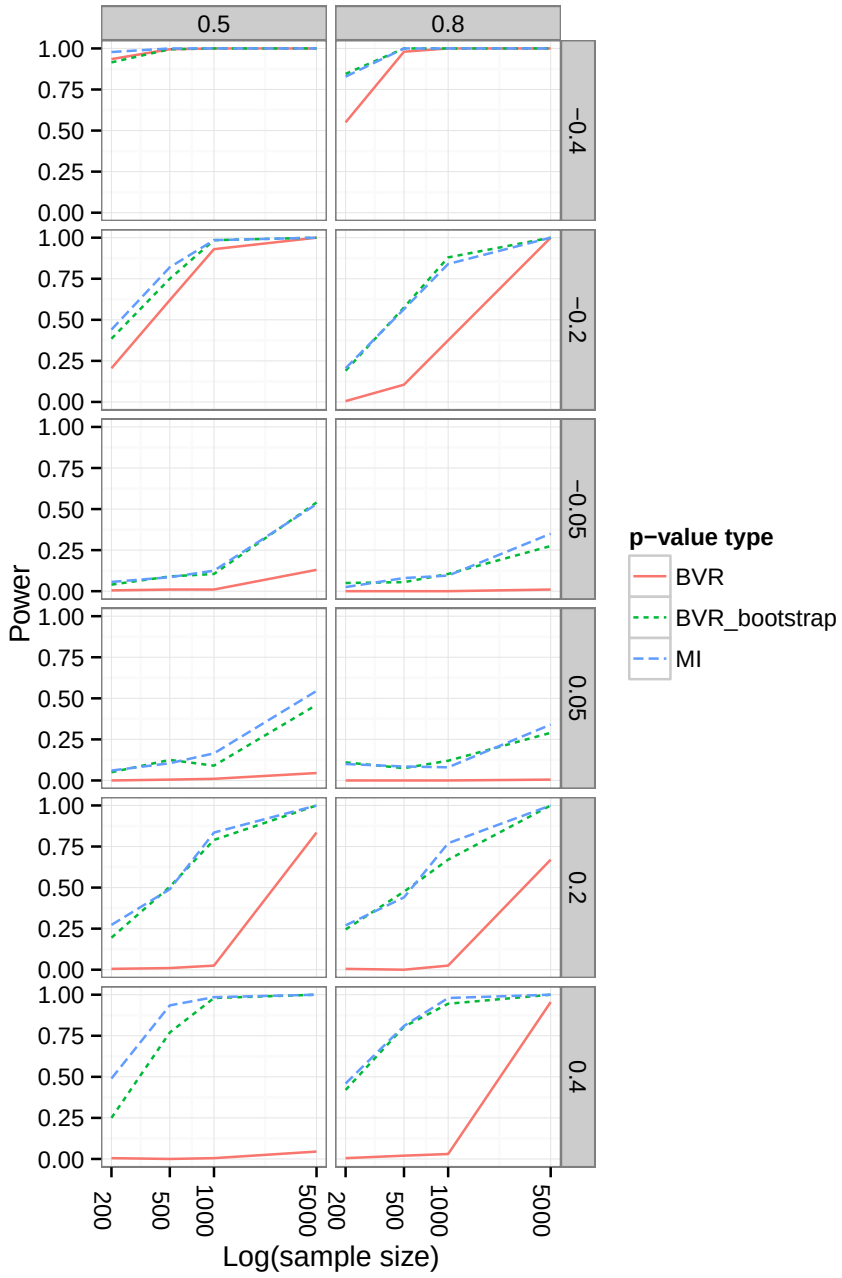

Fig. 1 Power to detect the local dependence under 48 conditions using the bootstrap p-value for the BVR ("bootstrap"), referring the BVR to a chi-square distribution ("BVR"), and referring the MI to a chi-square distribution ("MI"). Rows indicate the size of the local dependence, while columns correspond to the size of the loadings.

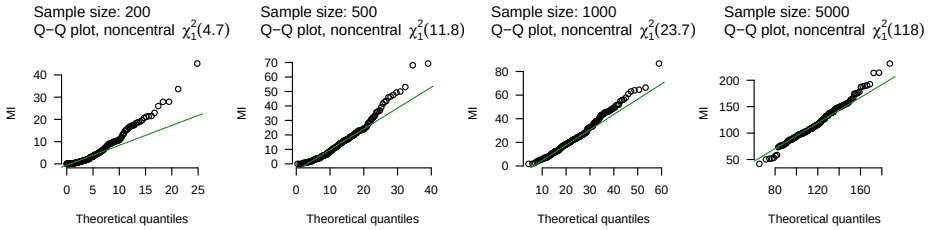

Fig. 2 Quantile-quantile plots of the MI under the simulation condition $\lambda = 0.8$, $\psi = 0.4$.

When local independence is violated, the MI asymptotically follows a noncentral chi-square distribution with one degree of freedom and noncentrality parameter (ncp) equal to the population shift in the χ^2 goodness-of-fit measure relative to the true model. To investigate whether this asymptotic result holds in finite samples, we applied Kolmogorov-Smirnoff (KS) tests of the hypothesis that the sample test statistics in each condition indeed followed a noncentral chi-square distribution with ncp corresponding to the population χ^2 of that condition. Performing this test for the BVR leads to a rejection ($p < 10^{-14}$) in all cases. For the MI, in contrast, the resulting p-values were larger than 0.01 for all but seven out of the 48 conditions, the average p-value for the KS test being 0.27. The most problematic condition in this regard is the condition in which the loading is 0.8 and the local dependence equals +0.4. Figure 2 demonstrates the fit of the sample MI values to their theoretical noncentral chi-squared distributions under this condition. For small sample sizes, the theoretical distribution does not appear to hold in this particular case. Figure 2 does demonstrate how increasing the sample size leads to a convergence to the theoretical distribution, as the fit improves with the sample size. Thus, in the exceptional case of small sample sizes, large positive local dependence, and large loadings, caution is warranted. In all other cases the MI appears to follow its theoretical distribution quite closely.

5 Concluding remarks

Binary data latent class analysis is a commonly applied model-based clustering method, in which a key assumption is that of local independence. We evaluated three methods to investigate the source of model misfit to this hypothesis by examining residuals in the bivariate cross-tables between observed variables. These methods were: 1) referring the bivariate residual (BVR) to a chi-square distribution, 2) referring the BVR to its parametric bootstrap distribution, and 3) referring the modification index (MI) to a chi-square distribution, also known as the score (or "Lagrange multiplier") test. The latter two methods are novel to the field of latent class analysis.

A Monte Carlo simulation study under various conditions demonstrated that judging the size of the BVR as though it were a chi-square variate (method 1) will yield α-levels lower than the nominal rate, and leads to inadequate power. The bootstrap and MI (methods 2 and 3) performed very similarly, showing adequate power and reproducing the nominal alpha levels. In the few cases where differences occurred, the MI appeared to be more powerful than the bootstrap BVR test. Furthermore, except in one condition, the MI approached its theoretical distribution. This suggests that when computational convenience is an issue, the MI provides an attractive alternative to the bootstrap for assessing the source of misfit to the hypothesis of local independence in latent class models.

Although it was already known theoretically that the BVR should not be regarded as a chi-square variate, method 1 is often encountered in applied research, possibly due to its intuitive appeal and convenience. The simulation study reported here clearly demonstrates that this practice may not always work as intended, in the sense that low BVR values cannot not be seen as indicative of good fit of the bivariate cross-tables to the hypothesis of local independence.

Acknowledgements This research was supported by grant ——– of the Netherlands Organization for Scientific Research (NWO).

References

Agresti, A. (2002). *Categorical data analysis, 2nd ed.* Wiley-Interscience, New York.

Albert, P. and Dodd, L. (2004). A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics*, 60(2):427–435.

Baughman, A., Bisgard, K., Cortese, M., Thompson, W., Sanden, G., and Strebel, P. (2008). Utility of composite reference standards and latent class analysis in evaluating the clinical accuracy of diagnostic tests for pertussis. *Clinical and Vaccine Immunology*, 15(1):106–114.

Chen, F., Mackey, A., Vermunt, J., and Roos, D. (2007). Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One*, 2(4):e383.

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.

Faraone, S. and Tsuang, M. (1994). Measuring diagnostic accuracy in. *Am J Psychiatry*, 1(51):651.

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.

Gaffikin, L., McGrath, J., Arbyn, M., and Blumenthal, P. (2007). Visual inspection with acetic acid as a cervical cancer test: accuracy validated using latent class analysis. *BMC medical research methodology*, 7(1):36.

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

Glas, C. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika*, 64(3):273–294.

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. P. (1988). 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. L. (2002). *Applied Latent Class Analysis*. Cambridge University Press.

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

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196.

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.

Hybels, C., Blazer, D., Pieper, C., Landerman, L., and Steffens, D. (2009). Profiles of depressive symptoms in older adults diagnosed with major depression: a latent cluster analysis. *The American journal of geriatric psychiatry: official journal of the American Association for Geriatric Psychiatry*, 17(5):387.

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.

Maydeu-Olivares, A. and Joe, H. (2005). Limited-and full-information estimation and goodness-of-fit testing in 2^n contingency tables. *Journal of the American Statistical Association*, 100(471):1009–1020.

McLachlan, G. and Peel, D. (2000). *Finite mixture models*, volume 299. Wiley-Interscience.

Nyholt, D., Gillespie, N., Heath, A., Merikangas, K., Duffy, D., and Martin, N. (2004). Latent class and genetic analysis does not support migraine with aura and migraine without aura as separate entities. *Genetic epidemiology*, 26(3):231–244.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

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

Satorra, A. (1989). Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrika*, 54(1):131–151.

Sörbom, D. (1989). Model modification. *Psychometrika*, 54(3):371–384.

Tay, L., Newman, D., and Vermunt, J. (2011). Using mixed-measurement item response theory with covariates (mm-irt-c) to ascertain observed and unobserved measurement equivalence. *Organizational Research Methods*, 14(1):147–176.

Torrance-Rynard, V. and Walter, S. (1998). Effects of dependent errors in the assessment of diagnostic test performance. *Statistics in Medicine*, 16(19):2157–2175.

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

van der Linden, W. and Glas, C. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika*, 75(1).

Vermunt, J. K. and Magidson, J. (2005). Technical guide for latent GOLD 4.0: Basic and advanced. *Belmont Massachusetts: Statistical Innovations Inc.*

Walter, S. and Irwig, L. (1988). Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *Journal of clinical epidemiology*, 41(9):923–937.