PSYCHOMETRIKA—VOL. 70, NO. 2, 325–345
JUNE 2005
DOI: 10.1007/s11336-003-1083-3

SELECTING THE NUMBER OF CLASSES UNDER LATENT CLASS REGRESSION: A FACTOR ANALYTIC ANALOGUE

GUAN-HUA HUANG

NATIONAL CHIAO TUNG UNIVERSITY

Recently, the regression extension of latent class analysis (RLCA) model has received much attention in the field of medical research. The basic RLCA model summarizes shared features of measured multiple indicators as an underlying categorical variable and incorporates the covariate information in modeling both latent class membership and multiple indicators themselves. To reduce complexity and enhance interpretability, one usually fixes the number of classes in a given RLCA. Often, goodness of fit methods comparing various estimated models are used as a criterion to select the number of classes. In this paper, we propose a new method that is based on an analogous method used in factor analysis and does not require repeated fitting. Two ideas with application to many settings other than ours are synthesized in deriving the method: a connection between latent class models and factor analysis, and techniques of covariate marginalization and elimination. A Monte Carlo simulation study is presented to evaluate the behavior of the selection procedure and compare to alternative approaches. Data from a study of how measured visual impairments affect older persons' functioning are used for illustration.

Key words: categorical data, factor analysis, finite mixture model, goodness of fit test, latent profile model, marginalization, residuals in generalized linear models, Monte Carlo simulation.

1. Introduction

Latent class analysis (LCA), originally described by Green (1951) and systematically developed by Lazarsfeld and Henry (1968), Goodman (1974), has been found useful for classifying subjects based on their responses to a set of categorical items. The basic model postulates an underlying categorical latent variable with, say, J categories, and measured items are assumed independent of one another within any category of the latent variable. Observed relationships among measured variables are thus assumed to result from the underlying classification of the data produced by the categorical latent variable. Recently, several authors extended the LCA model to describe the effects of measured covariates on the underlying mechanism (Dayton and Macready, 1988; Van der Heijden, Dessens, and Bökenholt, 1996; Bandeen-Roche, Miglioretti, Zeger, and Rathouz, 1997), or on measured item distributions within latent levels (Melton, Liang, and Pulver, 1994). This paper studies the problem of determining the number of latent variable levels in latent class models that incorporate covariate effects both on the latent variable and the measured indicators themselves (Formann, 1992; Hagenaars, 1993; Vermunt, 1996; Muthén, and Muthén, 1998; Vermunt, and Magidson, 2000).

To reduce complexity and enhance interpretability, one usually fixes the number of levels or "classes" in a given latent class model. When prior knowledge does not mandate the number of classes, selecting the number of classes to fit becomes an analytic challenge. Standard practice is to base selection on either the Pearson χ^2 or the likelihood ratio goodness of fit test, and to fix J at the lowest number of classes that yields acceptable fit (Goodman, 1974; Bartholomew, and Knott,

This work was supported by National Institute on Aging (NIA) Program Project P01-AG-10184-03. The author wishes to thank Dr. Karen Bandeen-Roche for her stimulating comments and helpful discussions, and Drs. Gary Rubin and Sheila West for kindly making the Salisbury Eye Evaluation data available.

Request for reprints should be sent to Guan-Hua Huang, Institute of Statistics, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu 300, Taiwan. E-mail: ghuang@stat.nctu.edu.tw. Webpage: http://www.stat.nctu.edu.tw/subhtml/source/teachers/ghuang/

1999; Formann, 1992). One well-known problem of this procedure is that when a large number of response patterns have low expected frequencies, the χ^2 approximation for the test distribution loses validity (Titterington, Smith, and Makov, 1985). In latent class models that build in regression on covariates [henceforth, regression extension of latent class analysis (RLCA)], asymptotic χ^2 inferences certainly fail if covariates are continuous (one individual per response-covariate "cell").

Instead of testing the goodness of fit of a specified model, we might use a criterion for selecting among different numbers of classes. Unfortunately, the standard generalized likelihood ratio statistic for testing $\mathrm{H}_0 : J = J_0$ versus $\mathrm{H}_A : J = J_0 + 1$ is not asymptotically distributed as χ^2, because the null hypothesis corresponds to a boundary of the parameter space of the alternative hypothesis (Titterington et al. 1985, Section 5.4). The AIC criterion (Akaike, 1987), which trades off the value of the likelihood at the maximum likelihood solution and the number of estimated parameters, is an appropriate and commonly used alternative approach (Moustaki, 1996; Wedel, Desarbo, Bult, and Ramaswamy, 1993). However, the use of AIC criterion has been proven to favor models with a greater number of parameters. Moreover, researchers have shown that AIC is not a consistent method because it does not depend on the sample size N (Kashyap, 1982, Schwarz, 1978). Schwarz (1978) proposed an alternative method that replaces the number of estimated parameters in the AIC (say, T) by $T \log N$. This selection process, known as the BIC criterion, was motivated by a Bayesian approach, and was proven to obtain a consistent estimate of the parameter number.

One common feature of the above methods is that they all must fit the latent class model repeatedly under different numbers of classes. Due to the slow convergence of commonly used fitting methods (e.g., EM algorithm: Dempster, Laird, and Rubin, 1977; Gibbs sampler: Geman and Geman, 1984), these procedures may be infeasible in practical applications. We develop a tool for identifying the number of latent classes, which requires no model fit based on the assumed class number and synthesizes ideas from factor analysis, latent variable theory and generalized linear model residuals. In the next section, we briefly describe the RLCA model that this paper studies. Section 3 motivates the proposed method with an analogy between finite mixture and factor analytic models. To implement this connection, we also develop techniques of marginalizing and eliminating covariate effects from both the latent variable and the measured indicators. Monte Carlo simulation is provided in Section 4 to evaluate the behavior of the proposed selection procedure and the comparison to alternatives. In Section 5, data from a visual functioning study are used to illustrate the proposed methods. We conclude by discussing procedure assumptions and proposing areas that need future study.

2. Model

Let $(Y_{i1}, \ldots, Y_{iM})$ denote a set of M observable polytomous outcome indicators and S_i denote the unobservable class membership, for the ith individual in a study sample of N persons. Y_{im} can take values $\{1, \ldots, K_m\}$, where $K_m \geq 2$, $m = 1, \ldots, M$, and S_i can take values $\{1, \ldots, J\}$. The LCA model is based on the concept of conditional independence in the sense that the observed variables are assumed to be statistically independent within latent classes. Therefore, the distribution for $(Y_{i1}, \ldots, Y_{iM})$ can be expressed as the finite mixture density:

$$\Pr(Y_{i1} = y_1, \ldots, Y_{iM} = y_m) = \sum_{j=1}^{J} \left\{ \eta_j \prod_{m=1}^{M} \prod_{k=1}^{K_m} p_{mkj}^{y_{mk}} \right\}, \tag{1}$$

where $\eta_j = \Pr(S_i = j)$ are mixing probabilities, $p_{mkj} = \Pr(Y_{im} = k|S_i = j)$ and $y_{mk} = \mathrm{I}(y_m = k) = 1$ if $y_m = k$; 0 otherwise.

To incorporate covariate effects into LCA, let $(\mathbf{x}_i, \mathbf{z}_i)$ be the associated covariate vector for the ith person, where $\mathbf{x}_i = [1, x_{i1}, \ldots, x_{iP}]^{\mathrm{T}}$ are predictors associated with latent class S_i, and

$\mathbf{z}_i = [\mathbf{z}_{i1}, \ldots, \mathbf{z}_{iM}]$; $\mathbf{z}_{im} = [1, z_{im1}, \ldots, z_{imL}]^{\mathrm{T}}$ with $m = 1, \ldots, M$ are covariates used to build direct effects on measured indicators. The two sets of covariates may include any combination of continuous and discrete measures. To derive an appropriate selection process, we begin by assuming that the two sets of covariates are mutually independent. In the provided simulation study, we will address how sensitive our approach is to this assumption. The basic RLCA equation can be stated as

$$\Pr(Y_{i1} = y_1, \ldots, Y_{iM} = y_m | \mathbf{x}_i, \mathbf{z}_i) = \sum_{j=1}^{J} \left\{ \eta_j(\mathbf{x}_i) \prod_{m=1}^{M} \prod_{k=1}^{K_m} p_{mkj}^{y_{mk}}(\mathbf{z}_{im}) \right\}, \tag{2}$$

with $\eta_j(\mathbf{x}_i)$ and $p_{mkj}(\mathbf{z}_{im})$ as in the generalized linear framework (McCullagh and Nelder, 1989). Often, (2) is implemented assuming generalized logit (Agresti, 1984) link functions:

$$\log\left[\frac{\eta_j(\mathbf{x}_i)}{\eta_J(\mathbf{x}_i)}\right] = \beta_{0j} + \beta_{1j}x_{i1} + \cdots + \beta_{Pj}x_{iP} \tag{3}$$

and

$$\begin{gathered} \log\left[\frac{p_{mkj'}(\mathbf{z}_{im})}{p_{mK_mj'}(\mathbf{z}_{im})}\right] = \gamma_{mkj'} + \alpha_{1mk}z_{im1} + \cdots + \alpha_{Lmk}z_{imL} \\ i = 1, \ldots, N; \quad m = 1, \ldots, M; \quad k = 1, \ldots, (K_m - 1); \\ j = 1, \ldots, (J-1); \quad j' = 1, \ldots, J. \end{gathered} \tag{4}$$

Notice that in the conditional probability model (4), we allow unrestricted intercepts and level- and item-specific covariate coefficients, but we do not allow the coefficients to vary across classes (i.e., α_{qmk} is dependent on m, k but independent of j). This constraint is logical if the primary purpose of modeling conditional probabilities is to prevent possible misclassification by adjusting for characteristics associated with item measurements. It is also necessary to unambiguously distinguish covariate effects on measured response probabilities from covariate effects on class probabilities. Three assumptions complete (2):

(C1) $\Pr(Y_{i1} = y_1, \ldots, Y_{iM} = y_m | S_i, \mathbf{x}_i, \mathbf{z}_i) = \Pr(Y_{i1} = y_1, \ldots, Y_{iM} = y_m | S_i, \mathbf{z}_i)$;

(C2) $\Pr(S_i = j | \mathbf{x}_i, \mathbf{z}_i) = \Pr(S_i = j | \mathbf{x}_i)$;

(C3) $\Pr(Y_{i1} = y_1, \ldots, Y_{iM} = y_m | S_i, \mathbf{z}_i) = \prod_{m=1}^{M} \Pr(Y_{im} = y_m | S_i, \mathbf{z}_{im})$.

For more detail on model assumptions, identifiability and parameter estimations, readers may reference Huang and Bandeen-Roche (2004).

3. Selecting the Number of Classes to Fit

Latent class analysis may legitimately be viewed as the categorical variable analog of factor analysis (Bartholomew and Knott, 1999). Both the number of factors in factor analysis and the number of classes in latent class models can be seen as the number of dimensions needed to characterize the systematic part of the response distribution. (Notice that the number of classes required does not need to be the same as the number of latent variables required. In this paper, we focus on the models that has one categorical latent variable with J categories.) This suggests that procedures used to determine the number of factors to extract in factor analysis might provide a useful basis for choosing the number of classes to fit in LCA and RLCA.

To motivate our model selection procedure, a commonly used criterion for determining the number of components to retain in factor analysis is based on a fact derived by Guttman (1954):

the number of eigenvalues of the population correlation matrix greater than or equal to unity is a lower bound for the number of factors. This suggests using the eigenvalues of the sample correlation matrix as a criterion for selecting the number of factors. We proceed to build a process that uses Guttman's criterion to determine the number of classes to fit in the regression extension of latent class model. Building an RLCA selection process is complicated by the introduction of predictor variables. In the case where covariates are only incorporated in the mixing probabilities η_j (i.e., no $\mathbf{z}_{im}$), Bandeen-Roche et al. (1997) found that, marginalizing over covariates, the joint distribution of observed variables of RLCA had the form of LCA (1). As a result, the number of classes may be inferred as for LCA ignoring the covariates. Model (2) is more complex because of the additional covariate effects on conditional probabilities, and therefore needs further work.

In the rest of this paper, we first briefly define and justify Guttman's eigenvalue method of determining the number of components in factor analysis. We then build the connection between finite mixture models and factor analysis, thereby extending Guttman's factor analytic criterion to determine the number of classes in a LCA model. To implement this connection to the proposed RLCA, a method of marginalizing over the covariate effects of model (2) is developed to reduce the complexity of the selection process nearly to the level of standard LCA.

3.1. Linear Factor Models and Guttman's Selection Criterion

The general philosophy of factor analysis is to replace M-dimensional data by L-dimensional data where L is much smaller than M. To be more specific, let $\mathbf{r}$ be an observable $M \times 1$ random vector with mean $\boldsymbol{\mu}_\mathbf{r}$ and covariance matrix $\mathrm{D}(\mathbf{r})$. The linear factor model postulates that $\mathbf{r}$ is linearly dependent upon an unobservable $L \times 1$ random vector $\mathbf{f}$ and a $M \times 1$ random error $\boldsymbol{\varepsilon}$ (Bartholomew and Knott, 1999):

$$\mathbf{r} = \boldsymbol{\mu}_\mathbf{r} + \boldsymbol{\ell}\mathbf{f} + \boldsymbol{\varepsilon}, \tag{5}$$

where $\boldsymbol{\ell}$ is a $M \times L$ matrix of coefficients with linearly independent columns, and $\mathbf{f}$ and $\boldsymbol{\varepsilon}$ satisfy

1. $\mathbf{f}$ and $\boldsymbol{\varepsilon}$ are independent;
2. $\mathrm{E}(\mathbf{f}) = \mathbf{0}$, $\mathrm{Cov}(\mathbf{f}) = \mathbf{I}$; and
3. $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi}$, where $\boldsymbol{\Psi}$ is a diagonal matrix.

It follows that the covariance matrix of $\mathbf{r}$

$$\mathrm{D}(\mathbf{r}) = \boldsymbol{\ell}\boldsymbol{\ell}^\mathrm{T} + \boldsymbol{\Psi}. \tag{6}$$

The principal component method provides one way to estimate $\boldsymbol{\ell}$ and $\boldsymbol{\Psi}$ in (6), as well as to choose the number of components of $\mathbf{f}$. $\mathrm{D}(\mathbf{r})$ is specified in terms of its eigenvalue–eigenvector pairs $(\lambda_1, \mathbf{e}_1), \dots, (\lambda_M, \mathbf{e}_M)$ with $\lambda_1 \geq \cdots \geq \lambda_M$ and $\mathbf{e}_j^\mathrm{T}\mathbf{e}_j = 1$, $\mathbf{e}_j^\mathrm{T}\mathbf{e}_m = 0$ for $j \neq m$. By the spectral decomposition theorem, and neglecting the contribution of the last $M - L$ eigenvalues, $\mathrm{D}(\mathbf{r})$ can be approximated by

$$\begin{aligned}\mathrm{D}(\mathbf{r}) &\doteq \begin{bmatrix} \sqrt{\lambda_1}\mathbf{e}_1, & \cdots, & \sqrt{\lambda_L}\mathbf{e}_L \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1}\mathbf{e}_1^T \\ \vdots \\ \sqrt{\lambda_L}\mathbf{e}_L^T \end{bmatrix} + \begin{bmatrix} \tilde{\psi}_1 & & & \\ & \tilde{\psi}_2 & & \mathbf{0} \\ \mathbf{0} & & \ddots & \\ & & & \tilde{\psi}_M \end{bmatrix} \\ &= \tilde{\boldsymbol{\ell}}\tilde{\boldsymbol{\ell}}^\mathrm{T} + \tilde{\boldsymbol{\Psi}}, \end{aligned} \tag{7}$$

where $\tilde{\psi}_i = \mathrm{Var}(r_i) - \sum_{j=1}^{L} \tilde{\ell}_{ij}^2$, r_i is the ith component of $\mathbf{r}$, and $\tilde{\ell}_{ij}$ is the (i, j)th component of $\tilde{\boldsymbol{\ell}}$. Analytically, we have

$$\text{sum of squared entries of } \mathrm{D}(\mathbf{r}) - (\tilde{\boldsymbol{\ell}}\tilde{\boldsymbol{\ell}}^{\mathrm{T}} + \tilde{\boldsymbol{\Psi}}) \leq \lambda_{L+1}^2 + \cdots + \lambda_M^2 \tag{8}$$

(Bartholomew and Knott, 1999, pp. 55–56). Consequently, a small value for sum of squares of the neglected eigenvalues implies a small value for the sum of squared errors of approximation in (7). If the number of components L is not determined by a prior consideration, then it is reasonable to choose L equal to the number of eigenvalue-eigenvector pairs that gives a reasonable approximation in (7) (i.e., $\lambda_{L+1}^2 + \cdots + \lambda_M^2$ reasonably small). Guttman (1954) recommended setting the number of common factors L as the number of sample correlation matrix eigenvalues greater than or equal to one. He claimed that this selected number could represent the actual structure of the data. He also showed that the true number of common factors is bounded below by the number of population correlation eigenvalues ≥ 1, which gives theoretical support to this selecting method.

3.2. Connection between Finite Mixture Models and Linear Factor Analysis

Bartholomew and Knott (1999, pp. 151–155) built an important link between finite mixture models and linear factor analysis. This connection was based on moment estimations for finite mixture models as proposed by Lazarsfeld and Henry (1968). For the polytomous, categorical response Y_{im} in (1, 2), it is usually represented as a vector with elements being the indicators of each category. To implement Bartholomew and Knott's connection to our latent class models (1, 2), we need to further extend it to the case where each measured item is a vector.

Specifically, let $(\mathbf{R}_{i1}, \ldots, \mathbf{R}_{iM})$ be M observed "vectors", where $\mathbf{R}_{im}$ is a $(K_m - 1) \times 1$ vector, $m = 1, \ldots, M$. Therefore, under the finite mixture model

$$\Pr(\mathbf{R}_{i1}, \ldots, \mathbf{R}_{iM}) = \sum_{j=1}^{J} \left\{ \eta_j \prod_{m=1}^{M} f_{mj}(\mathbf{R}_{im} | S_i = j) \right\}, \tag{9}$$

where $f_{mj}(\cdot|\cdot)$ is a multivariate distribution with $\mathrm{E}(\mathbf{R}_{im}|S_i = j) = \boldsymbol{\mu}_m(j)$ and $\mathrm{Var}(\mathbf{R}_{im}|S_i = j) = \boldsymbol{\Sigma}_m(j)$,

$$\mathrm{E}[\mathbf{R}_{im}(\mathbf{R}_{im})^{\mathrm{T}}] = \sum_{j=1}^{J} \{\eta_j [\boldsymbol{\Sigma}_m(j) + \boldsymbol{\mu}_m(j)\boldsymbol{\mu}_m(j)^{\mathrm{T}}]\},$$

$$\mathrm{E}[\mathbf{R}_{im}(\mathbf{R}_{iq})^{\mathrm{T}}] = \sum_{j=1}^{J} \{\eta_j \boldsymbol{\mu}_m(j)\boldsymbol{\mu}_q(j)^{\mathrm{T}}\},$$

where $m, q = 1, \ldots, M$; $m \neq q$. We then have

$$\begin{aligned}\mathrm{Var}(\mathbf{R}_{im}) &= \mathrm{E}[\mathbf{R}_{im}(\mathbf{R}_{im})^{\mathrm{T}}] - \mathrm{E}(\mathbf{R}_{im})[\mathrm{E}(\mathbf{R}_{im})]^{\mathrm{T}} \\ &= \sum_{j=1}^{J} \{\eta_j \boldsymbol{\Sigma}_m(j)\} + \sum_{j=1}^{J} \{\eta_j [(\boldsymbol{\mu}_m(j) - \bar{\boldsymbol{\mu}}_m)(\boldsymbol{\mu}_m(j) - \bar{\boldsymbol{\mu}}_m)^{\mathrm{T}}]\} \qquad (10)\end{aligned}$$

$$\begin{aligned}\mathrm{Cov}(\mathbf{R}_{im}, \mathbf{R}_{iq}) &= \mathrm{E}[\mathbf{R}_{im}(\mathbf{R}_{iq})^{\mathrm{T}}] - \mathrm{E}(\mathbf{R}_{im})[\mathrm{E}(\mathbf{R}_{iq})]^{\mathrm{T}} \\ &= \sum_{j=1}^{J} \{\eta_j [(\boldsymbol{\mu}_m(j) - \bar{\boldsymbol{\mu}}_m)(\boldsymbol{\mu}_q(j) - \bar{\boldsymbol{\mu}}_q)^{\mathrm{T}}]\}, \quad (m \neq q), \qquad (11)\end{aligned}$$

where $\bar{\boldsymbol{\mu}}_m = \sum_{j=1}^{J} \eta_j \boldsymbol{\mu}_m(j)$. The covariance matrix for $(\mathbf{R}_{i1}, \ldots, \mathbf{R}_{iM})$ may thus be written as

$$\mathrm{D}(\mathbf{R}_{i1}, \ldots, \mathbf{R}_{iM}) = \mathbf{L}\mathbf{L}^{\mathrm{T}} + \boldsymbol{\Psi}, \tag{12}$$

where $\boldsymbol{\Psi}$ is a $\sum_{m=1}^{M}(K_m - 1) \times \sum_{m=1}^{M}(K_m - 1)$ diagonal block matrix with mth block $\boldsymbol{\Psi}_m = \sum_{j=1}^{J}\{\eta_j \boldsymbol{\Sigma}_m(j)\}$ (a $(K_m - 1) \times (K_m - 1)$ matrix), and the elements of $\sum_{m=1}^{M}(K_m - 1) \times J$ matrix $\mathbf{L} = (\mathbf{l}_{mj})$ are given by $\mathbf{l}_{mj} = \sqrt{\eta_j}\{\boldsymbol{\mu}_m(j) - \bar{\boldsymbol{\mu}}_m\}$, $m = 1, \ldots, M$; $j = 1, \ldots, J$. Notice that $\mathrm{D}(\mathbf{R}_{i1}, \ldots, \mathbf{R}_{iM})$ is of exactly the same form as the covariance matrix for the linear factor model, but with one important difference: columns of $\mathbf{L}$ are linearly dependent because $\sum_{j=1}^{J} \sqrt{\eta_j}\mathbf{l}_{mj} = 0$, $\forall m$. For this reason, the principal component approach (7) is not applicable, where columns of $\tilde{\boldsymbol{\ell}}$ are linearly independent. However, by Graybill (1983), Theorem 1.7.7, there exists a $\sum_{m=1}^{M}(K_m - 1) \times J^*$ matrix $\boldsymbol{\Omega}$ with linearly independent columns such that $\mathbf{L}\mathbf{L}^{\mathrm{T}} = \boldsymbol{\Omega}\boldsymbol{\Omega}^{\mathrm{T}}$, where $J^* = \mathrm{rank}(\mathbf{L}) < J$. This completes the correspondence. We can then write

$$\mathrm{D}(\mathbf{R}_{i1}, \ldots, \mathbf{R}_{iM}) = \boldsymbol{\Omega}\boldsymbol{\Omega}^{\mathrm{T}} + \boldsymbol{\Psi}. \tag{13}$$

By applying the principal component method and Guttman's selection criterion used in linear factor models to (13), then, it might be reasonable to choose J^* equal to the number of sample correlation matrix eigenvalues greater than or equal to one.

3.3. Marginalization of the Regression Extension of Latent Class Model

The connection (13) is based on the finite mixture model (9), where no covariates are incorporated. To apply the connection (13) to the RLCA model (2), we develop a process to "eliminate" the covariate effects, hence "marginalize" the model (2). The marginalization process we propose includes two stages. Stage 1 aims to eliminate $\mathbf{z}_i$. We then apply the marginalization property used in Bandeen-Roche et al. (1997) to average $\mathbf{x}_i$ effects out of the latent prevalences.

3.3.1. Marginalizing the Covariate Effects on Conditional Probabilities

The key to marginalizing over $\mathbf{z}_i$ is that the process must yield random variables that follow a finite mixture distribution that is both independent of $\mathbf{z}_i$ and has J mixing components. One strategy for achieving such marginalization can be motivated by the properties of added variable plots for linear regression models. In the following, we first introduce and extend this strategy to extract $\mathbf{z}_i$ from model (4); second, we develop the residuals for the generalized linear models, which are needed for implementing the above strategy; third, we formulate the orthogonal condition under model (4), which is sufficient for completing the extension; and fourth, we generalize the result from the binary-measured-indicator case to the polytomous case.

Consider the linear model

$$\mathbf{Y} = \mathbf{x}_1^{\mathrm{T}}\boldsymbol{\beta}_1 + \mathbf{x}_2^{\mathrm{T}}\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}, \tag{14}$$

where $\boldsymbol{\varepsilon}$ has mean $\mathbf{0}$ and variance matrix $\mathbf{V}$. Let $\tilde{\mathbf{Y}}$ denote the residuals of regressing $\mathbf{Y}$ on $\mathbf{x}_2$, and $\mathbf{W} = \mathbf{V}^{-1}$ be the weight matrix. Then, it is well-known that if $\mathbf{x}_1$ and $\mathbf{x}_2$ are orthogonal (i.e., $\mathbf{x}_1\mathbf{W}\mathbf{x}_2^{\mathrm{T}} = 0$), $\tilde{\mathbf{Y}}$ has mean $\mathbf{x}_1^{\mathrm{T}}\boldsymbol{\beta}_1$ and variance $\mathbf{V}$. Hence, the simple linear regression of $\tilde{\mathbf{Y}}$ on $\mathbf{x}_1$ yields exactly the same inferences about $\boldsymbol{\beta}_1$ as if we performed the analysis on the more complicated model (14) (Cook and Weisberg, 1982). Viewing the just-described stability of $\boldsymbol{\beta}_1$ as analogous to the desired stability of latent class dimension, J, we now apply the added variable property to model (4) to obtain marginalized conditional probabilities.

To present the key ideas more clearly, henceforth, the measured indicators $(Y_{i1}, \ldots, Y_{iM})$ are assumed to be binary (i.e., $K_1 = \cdots = K_M = 2$). We then generalize the result to polytomous indicators as a final step. To make the analogy to (14), notice that (4) can be viewed as fitting a logistic regression of Y_{im} on S_i adjusting for $\mathbf{z}_{im}$, separately for each m. To see this, let $S_{ij} = \mathrm{I}(S_i = j)$ for $i = 1, \dots, N$; $j = 1, \dots, J-1$. We can reparameterize (4) as

$$\mathrm{logit}[\mathrm{E}(Y_{im}|\mathbf{S}_i, \mathbf{Z}^c_{im})] = \mathbf{S}_i^{\mathrm{T}}\boldsymbol{\gamma}_m + (\mathbf{Z}^c_{im})^{\mathrm{T}}\boldsymbol{\alpha}_m \quad \text{for } i = 1, \dots, N; \;\; m = 1, \dots, M, \tag{15}$$

where
$\mathbf{S}_i = [1, S_{i1}, \dots, S_{i(J-1)}]^{\mathrm{T}}$; $\mathbf{Z}^c_{im} = [(z_{im1} - \bar{z}_{m1}), \dots, (z_{imL} - \bar{z}_{mL})]^{\mathrm{T}}$, $\bar{z}_{mp} = (1/N)\sum_{i=1}^{N} z_{imp}$ ("centered" covariate matrix); $\boldsymbol{\gamma}_m = [\gamma_{m0}, \gamma_{m1}, \dots, \gamma_{m(J-1)}]^{\mathrm{T}}$; and $\boldsymbol{\alpha}_m = [\alpha_{1m}, \dots, \alpha_{Lm}]^{\mathrm{T}}$. Therefore, for any realization of $\mathbf{S}_i$, (15) is a logistic regression with dependent variable: Y_{im} and predictors: $\mathbf{S}_i, \mathbf{Z}^c_{im}$.

Next, the problem becomes how to calculate residuals from the generalized linear model

$$\mathrm{logit}[\mathrm{E}(Y_{im}|\mathbf{Z}^c_{im})] = (\mathbf{Z}^c_{im})^{\mathrm{T}}\boldsymbol{\alpha}^*_m \quad \text{for } i = 1, \dots, N; \;\; m = 1, \dots, M. \tag{16}$$

If we estimate $\boldsymbol{\alpha}^*_m$ by the method of iteratively reweighed least-squares (IRLS), and using the fact that logit is the canonical link function for binomial data, $\hat{\boldsymbol{\alpha}}^*_m$ as of the $(t+1)$th iteration can be written

$$\hat{\boldsymbol{\alpha}}^*_m(t+1) = [\mathbf{Z}^c_m \hat{\mathbf{V}}_m(t)(\mathbf{Z}^c_m)^{\mathrm{T}}]^{-1}[\mathbf{Z}^c_m \hat{\mathbf{V}}_m(t)\mathbf{Y}^*_m(t)], \quad \text{for } m = 1, \dots, M. \tag{17}$$

Here, "hat" represents the estimated values; $\mathbf{Y}_m = [Y_{1m}, \dots, Y_{Nm}]^{\mathrm{T}}$; $\mathbf{V}_m = \mathrm{diag}(V_{1m}, \dots, V_{Nm})$, $V_{im} = \mathrm{Var}(Y_{im})$; $\mathbf{Z}^c_m = [\mathbf{Z}^c_{1m}, \dots, \mathbf{Z}^c_{Nm}]$; and

$$\mathbf{Y}^*_m(t) = (\mathbf{Z}^c_m)^{\mathrm{T}}\hat{\boldsymbol{\alpha}}_m(t) + \hat{\mathbf{V}}_m^{-1}(t)[\mathbf{Y}_m - \hat{\boldsymbol{\mu}}^*_m(t)] \tag{18}$$

with $\hat{\boldsymbol{\mu}}^*_m(t) = \mathrm{E}(\mathbf{Y}_m|\mathbf{Z}^c_m)|_{\boldsymbol{\alpha}^*_m = \hat{\boldsymbol{\alpha}}^*_m(t)}$. Comparing with the ordinary weighted linear regression, $\hat{\mathbf{V}}_m$, $\mathbf{Y}^*_m$ and $(\mathbf{Z}^c_m)^T\hat{\boldsymbol{\alpha}}^*_m$ can be thought of as the weight matrix, "pseudo-observation" and "pseudo-fitted-value". So, "pseudo-residuals" are given by

$$\mathbf{R}_m = [R_{1m}, \dots, R_{Nm}]^{\mathrm{T}} = \mathbf{Y}^*_m - (\mathbf{Z}^c_m)^{\mathrm{T}}\hat{\boldsymbol{\alpha}}^*_m = \hat{\mathbf{V}}_m^{-1}(\mathbf{Y}_m - \hat{\boldsymbol{\mu}}_m). \tag{19}$$

A logistic regression version of the partial residual plot based on the pseudo-residuals (19) was suggested by Landwehr, Pregibon, and Shoemaker (1984). They used both simulated data and real examples to show that the partial residual plot based on (19) can detect possible nonlinearity between outcomes and predictors.

To extend the orthogonality property in the linear model (14), we need to assume that

(C4) $\mathbf{S}$ and $\mathbf{Z}^c_m$ are orthogonal, that is, $\mathbf{S}\mathbf{W}_m(\mathbf{Z}^c_m)^{\mathrm{T}} = 0$,

where $\mathbf{S} = [\mathbf{S}_1, \dots, \mathbf{S}_N]$, and $\mathbf{W}_m = \mathbf{V}_m$ is the weight matrix in the model (15). This assumption cannot be verified because $\mathbf{S}$ is unobservable. However, by assumption **(C2)**, S_i and $\mathbf{z}_i$ are independent of given $\mathbf{x}_i$. Thus, if

(C5) $\mathbf{x}_i$ and $\mathbf{z}_{im}$ are independent,

then $\mathbf{S}$ and $\mathbf{Z}^c_m$ are mutually uncorrelated. Since $\mathrm{E}[\mathbf{W}_m^{1/2}(\mathbf{Z}^c_m)^{\mathrm{T}}] = 0$, this implies $\mathrm{E}[\mathbf{S}\mathbf{W}_m(\mathbf{Z}^c_m)^{\mathrm{T}}] = 0$. Thus, if **(C5)** is true, **(C4)** holds to an increasingly close approximation as $N \to \infty$. **(C5)** can be verified empirically by calculating the sample correlation matrix among covariates. As $\mathbf{x}_i$ seeks to estimate the effects of risk factors on the conceptual outcome and $\mathbf{z}_{im}$ aims to adjust for characteristics associated with measured indicators to prevent possible misclassification of latent classes (Huang and Bandeen-Roche, 2004), analysts may select two exclusive sets of covariates based on study objects and data characteristics. Appropriate statistical methods (e.g., principal component analysis) can also be used to uncover approximate linear dependencies among covariates.

Under assumption **(C5)**, we can then extract the $\mathbf{Z}_{im}^{c}$ from conditional probabilities by treating the residuals from the model (16) as new response variables and regressing them on $\mathbf{S}_i$. We propose to substitute the estimate of $\boldsymbol{\gamma}_m^*$ in the linear model

$$R_{im} = \mathbf{S}_i^{\mathrm{T}} \boldsymbol{\gamma}_m^* + \varepsilon_{im}, \quad i = 1, \dots, N; \quad m = 1, \dots, M \tag{20}$$

for the estimate of $\boldsymbol{\gamma}_m$ in the model (15). A formal justification (Appendix A) shows that $\boldsymbol{\gamma}_m^*$ and $\boldsymbol{\gamma}_m$ can be very close under reasonable regularities.

The above results can be extended to the cases where $(Y_{i1}, \dots, Y_{iM})$ are polytomous as in (1) and (2). Under polytomous item responses, the pseudo-residuals for mth item can be written as

$$\mathbf{R}_m^p = [(\mathbf{R}_{1m}^p)^{\mathrm{T}}, \dots, (\mathbf{R}_{Nm}^p)^{\mathrm{T}}]^{\mathrm{T}} = (\hat{\mathbf{V}}_m^p)^{-1}[\mathbf{Y}_m^p - \hat{\boldsymbol{\mu}}_m^p], \tag{21}$$

where
"p" denotes polytomous responses and "hat" denotes the estimated values;
$\mathbf{R}_{im}^p = [R_{im1}, \dots, R_{im(K_m-1)}]^{\mathrm{T}}$; $\mathbf{Y}_m^p = [(\mathbf{Y}_{1m}^p)^{\mathrm{T}}, \dots, (\mathbf{Y}_{Nm}^p)^{\mathrm{T}}]^{\mathrm{T}}$ with $\mathbf{Y}_{im}^p = [Y_{im1}, \dots, Y_{im(K_m-1)}]^{\mathrm{T}}$ and $Y_{imk} = \mathrm{I}(Y_{im} = k)$; $\mathbf{V}_m^p = \mathrm{diag}(\mathbf{V}_{1m}^p, \dots, \mathbf{V}_{Nm}^p)$ with $\mathbf{V}_{im}^p = \mathrm{Var}(\mathbf{Y}_{im}^p)$; $\boldsymbol{\mu}_m^p = \mathrm{E}(\mathbf{Y}_m^p | \mathbf{Z}_m^c)$; and $i = 1, \dots, N, m = 1, \dots, M, k = 1, \dots, K_m$. Thus, (20) becomes

$$\mathbf{R}_{im}^p = (\mathbf{S}_{im}^p)^{\mathrm{T}} \boldsymbol{\gamma}_m^p + \boldsymbol{\varepsilon}_{im}^p, \tag{22}$$

where $\mathbf{S}_{im}^p = [\mathbf{1}^{(K_m-1)}, \mathbf{S}_{i1}^{(K_m-1)}, \dots, \mathbf{S}_{i(J-1)}^{(K_m-1)}]^{\mathrm{T}}$ with $\mathbf{A}^{(K_m-1)} = \mathrm{diag}(\overbrace{A, \dots, A}^{K_m-1})$ and $\mathbf{A} = \mathbf{1}$ or $\mathbf{S}_{ij}$; $\boldsymbol{\gamma}_m^p = [(\boldsymbol{\gamma}_{m0}^p)^{\mathrm{T}}, (\boldsymbol{\gamma}_{m1}^p)^{\mathrm{T}}, \dots, (\boldsymbol{\gamma}_{m(J-1)}^p)^{\mathrm{T}}]^{\mathrm{T}}$ with $\boldsymbol{\gamma}_{mj}^p = [\gamma_{m1j}, \dots, \gamma_{m(K_m-1)j}]^{\mathrm{T}}$; and $i = 1, \dots, N, m = 1, \dots, M, j = 1, \dots, (J-1)$.

Notice that the variance of $\boldsymbol{\varepsilon}_{im}^p$ in (22) varies through the associated covariates $(\mathbf{x}_i, \mathbf{z}_{im})$ and latent class S_i. Therefore, the marginalization process does not marginalize the covariate effects from the variance. However, Liang and Zeger (1986) showed that, using generalized estimating equations, the parameter estimate is still consistent even if we specify an incorrect variance structure. Since the maximum likelihood approach is used for parameter estimates, we can assume that the variance of $\boldsymbol{\varepsilon}_{im}^p$ does not depend on associated covariates while still maintaining the consistency of $\boldsymbol{\gamma}_m^p$. Therefore, it is reasonable to think of the marginalization of model (2) over $\mathbf{z}_i$ as

$$\Pr(\mathbf{R}_{i1}^p, \dots, \mathbf{R}_{iM}^p | \mathbf{x}_i) = \sum_{j=1}^{J} \left\{ \eta_j(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}) \prod_{m=1}^{M} f_{mj}(\mathbf{R}_{im}^p | S_i = j) \right\}, \tag{23}$$

where $f_{mj}(\cdot|\cdot)$ is a multivariate distribution with $\mathrm{E}(\mathbf{R}_{im}^p | S_i = j) = \boldsymbol{\mu}_m(j) = \boldsymbol{\gamma}_{m0}^p + \boldsymbol{\gamma}_{mj}^p$ if $j = 1, \dots, J-1$, $\boldsymbol{\gamma}_{m0}^p$ if $j = J$; and $\mathrm{Var}(\mathbf{R}_{im}^p | S_i = j) = \boldsymbol{\Sigma}_m(j)$. The conditional independence of $(\mathbf{R}_{i1}^p, \dots, \mathbf{R}_{iM}^p)$ given S_i holds in equation (23) because $\mathbf{R}_{im}^p = \mathbf{R}_{im}^p(\mathbf{Y}_{im}^p)$ for $m = 1, \dots, M$, and therefore $(\mathbf{R}_{i1}^p, \dots, \mathbf{R}_{iM}^p)$ are conditionally independent given $\mathrm{S_i}$, $\mathbf{z}_i$ and approximately independent of $\mathbf{z}_i$ given $\mathrm{S_i}$ due to marginalization. Since the estimators of $\boldsymbol{\gamma}_m$ in (15) and $\boldsymbol{\gamma}_m^*$ in (20) are asymptotically equivalent, this marginalization process keeps the number of classes fixed.

3.3.2. Marginalizing the Covariate Effects on Latent Prevalences

Next, we need to marginalize $\mathbf{x}_i$ effects from the latent prevalences of (23) to use the connection built in the previous section. The latent variable regression model (23) possesses the nice property that the covariates associated with class prevalences, $\mathbf{x}_i$, can be ignored. This is seen by marginalizing over the covariates $\mathbf{x}_i$:

$$\begin{aligned}\Pr(\mathbf{R}^p_{i1}, \ldots, \mathbf{R}^p_{iM}) &= \int \Pr(\mathbf{R}^p_{i1}, \ldots, \mathbf{R}^p_{iM}|\mathbf{x}_i)\mathrm{d}G(\mathbf{x}_i) \\ &= \sum_{j=1}^{J}\left\{\left[\int \eta_j(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})dG(\mathbf{x}_i)\right]\left[\prod_{m=1}^{M} f_{mj}(\mathbf{R}^p_{im}|S_i = j)\right]\right\} \\ &= \sum_{j=1}^{J}\left\{\eta_j^* \prod_{m=1}^{M} f_{mj}(\mathbf{R}^p_{im}|S_i = j)\right\} \qquad (24)\end{aligned}$$

with $G(\mathbf{x}_i)$ representing the probability distribution of $\mathbf{x}_i$. Thus, the finite mixture model (24) can be thought as the marginalized version of model (2) over $(\mathbf{x}_i, \mathbf{z}_i)$.

The proposed two-stage marginalization procedure can be applied to the entire range of RLCA models. For RLCA that only incorporates covariates in latent prevalences, stage one is skipped, so that $\mathbf{R}^p_{im} = \mathbf{Y}^p_{im}$ and $f_{mj}(\cdot|\cdot)$ is a multinomial distribution in (23). The marginalized model of RLCA which only allows covariate effects in conditional probabilities is (24) with $\eta_j^* = \eta_j$.

3.4. The Eigenvalue Criterion of Selecting the Number of Classes in RLCA

By implementing the connection between finite mixture models and linear factor analysis to (24), we thereby recommend a method of choosing the number of classes to fit in RLCA (2), similar to Guttman's eigenvalue criterion in linear factor models.

Before we describe the method, let $\mathrm{R}(\mathbf{R}^p_{i1}, \ldots, \mathbf{R}^p_{iM}) = \mathbf{V}^{-1/2}\mathrm{D}(\mathbf{R}^p_{i1}, \ldots, \mathbf{R}^p_{iM})\mathbf{V}^{-1/2}$ denote the correlation matrix for $(\mathbf{R}^p_{i1}, \ldots, \mathbf{R}^p_{iM})$, where $\mathbf{V} = \mathrm{diag}(\mathrm{Var}(\mathbf{R}^p_{i1}), \ldots, \mathrm{Var}(\mathbf{R}^p_{iM}))$. Further, let

$$\mathbf{U} = \begin{bmatrix} \boldsymbol{\mu}_1(1) & \cdots & \boldsymbol{\mu}_1(J) \\ \vdots & & \vdots \\ \boldsymbol{\mu}_M(1) & \cdots & \boldsymbol{\mu}_M(J) \end{bmatrix} = \begin{bmatrix} \mathbf{U}_1 \\ \vdots \\ \mathbf{U}_M \end{bmatrix}$$

with $\boldsymbol{\mu}_m(j)$ as defined in the text following equation (23) for all m, j.

Theorem 1. Suppose that

- **(C6)** $\sum_{m=1}^{M}(K_m - 1) \geq (J - 1) \geq \max\{(K_1 - 1), \ldots, (K_M - 1)\}$;
- **(C7)** the rank of $\mathbf{U}_m$ is $(K_m - 1)$, for $m = 1, \ldots, M$;
- **(C8)** the columns of $\mathbf{U}$ are linearly independent; and
- **(C9)** the conditional variance $\boldsymbol{\Sigma}_m(j)$ as defined in the text following Equation (23) is positive definite, and η_j^* in (24) is greater than 0, $m = 1, \ldots, M$, $j = 1, \ldots, J$.

Then, the number of latent classes $J \geq (r+1)$, where r denotes the number of $\mathrm{R}(\mathbf{R}^p_{i1}, \ldots, \mathbf{R}^p_{iM})$'s eigenvalues that are greater than or equal to one.

The proof of Theorem 1 is detailed in Appendix B. In assumption **(C6)**, we assume that the dimension of the parameter space for the latent structure is not less than the one for the marginal distribution of any given pseudo-residual $\mathbf{R}^p_{im}$, but it is not greater than the one for the marginal distribution of all M pseudo-residuals. This assumption is reasonable if we think of latent variable modeling as a dimension reduction process. Assumption **(C7)** requires all $K_m - 1$ elements of $\mathbf{R}^p_{im}$ to have distinct conditional distributions given on the latent class, and assumption **(C8)** requires that distributions for pseudo-residuals are distinct across latent classes. Assumption **(C9)** excludes degenerate conditional distributions and latent classes. In practice, it is difficult to check these assumptions because the latent variable distribution is unknown. However, if $f_{mj}(\cdot|\cdot)$ in (24) follows a multinormal distribution, Yakowitz and Spragins (1968) showed that, under assumptions **(C7)**, **(C8)** and **(C9)**, the finite mixture model (24) is identifiable. Therefore, this theorem works for an identifiable regression extension of latent class model (2), which implies the identifiability of the finite mixture model (24), with the true number of latent classes met assumption **(C6)**.

In large samples, the theorem provides a theoretical justification of how the proposed selection criterion approximates the true number of classes. This result is based on the population correlation matrix, not the sample correlation matrix. In practice, we recommend using the following algorithm to estimate the number of latent classes:

Step 1. Calculate the pseudo-residual of ith participant's mth response item $\mathbf{R}^p_{im} = (\hat{\mathbf{V}}^p_{im})^{-1}(\mathbf{Y}^p_{im} - \hat{\boldsymbol{\mu}}^p_{im})$, where $\hat{\mathbf{V}}^p_{im}$, $\mathbf{Y}^p_{im}$ and $\hat{\boldsymbol{\mu}}^p_{im}$ as defined in (21), and $i = 1, \ldots, N$; $m = 1, \ldots, M$.

Step 2. Create the sample correlation matrix $\mathbf{C} := \mathbf{T}^{-1/2}\mathbf{H}\mathbf{T}^{-1/2}$, where $\mathbf{H} = (\mathbf{H}_{mq})$ with elements $\mathbf{H}_{mq} = 1/N - 1 \sum_{i=1}^{N}(\mathbf{R}^p_{im} - \bar{\mathbf{R}}^p_m)(\mathbf{R}^p_{iq} - \bar{\mathbf{R}}^p_q)^T$, $\bar{\mathbf{R}}^p_m = 1/N \sum_{i=1}^{N} \mathbf{R}^p_{im}$, $m, q = 1, \ldots, M$; and $\mathbf{T} = \operatorname{diag}(\mathbf{H}_{11}, \ldots, \mathbf{H}_{MM})$.

Step 3. Select the number of classes as one plus the number of **C**'s eigenvalues that are greater than or equal to one.

This algorithm also works for the straightforward finite mixture model, where observed outcomes Y_{im}'s can follow any distribution and no covariate effects are added. In Step 1, let $\mathbf{R}^p_{im} = Y_{im}$, and then Step 2 and Step 3 can be used and are valid because of the link between finite mixture models and factor analysis (Section 3.2).

4. Simulation Study

The simulation study contains two parts. The first part aims to examine the performance of the proposed method, and the second part focuses on the comparison to alternative approaches.

4.1. *Performance of the Proposed Method*

Here, we specifically aim to address three issues. First, we aim to compare how the estimated numbers of classes vary under different true numbers of classes and different parameter (i.e., $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$) structures. Second, the proposed method assumes that **S** and $\mathbf{Z}^c_m$ are mutually uncorrelated for each m, which excludes the common situation of correlations among $\mathbf{x}_i$ and $\mathbf{z}_i$. Thus, a second aim is to evaluate the sensitivity of the proposed method to this assumption. Third, the sample correlation matrix is used to approximate the actual population correlation matrix. Therefore, the larger the sample size, the more accurate the approximation. How the sample size affects the estimating procedure is also studied.

Two different RLCA (2) models were simulated. One was a three-class RLCA with five two-level measured indicators, two covariates associated with conditional probabilities, and two covariates associated with latent prevalences (i.e., $J = 3$, $M = 5$, $K_1 = \cdots = K_5 = 2$, $P = L = 2$). The other was a six-class RLCA with five three-level measured indicators and the same covariate setting as the three-class model ($J = 6, M = 5, K_1 = \cdots = K_5 = 3, P = L = 2$). For each model, the model parameters β_{pj} were determined through four methods. For each $p \in \{0, 1, \ldots, P\}$

- equal parameters: $\beta_{pj} = k_1,\ j = 1, \ldots, (J-1)$;
- randomly selected: $\beta_{pj} = k_2 U_j,\ U_j \sim U(0,1),\ j = 1, \ldots, (J-1)$;
- moderately decreasing: $\beta_{pj} = k_3/j,\ j = 1, \ldots, (J-1)$; and
- rapidly decreasing: $\beta_{pj} = k_4/2^j,\ j = 1, \ldots, (J-1)$,

where $k_1, \ldots, k_4$ were constants such that $\sum_{j=1}^{J-1} \beta_{pj}$ equaled the preselected total. These methods were also applied to create $\{\gamma_{jmk},\ j = 1, \ldots, J\}$ for all m, k, and $\{\alpha_{qmk},\ m = 1, \ldots, M; k = 1, \ldots, (K_m - 1)\}$ for all q. All $(\beta_{pj}, \gamma_{jmk}, \alpha_{qmk})$ pairs were generated by the same method.

To determine the effect of assuming mutually uncorrelated $\mathbf{S}$ and $\mathbf{Z}_m^c$, the covariates associated with conditional probabilities $(z_{im1}, z_{im2}, m = 1, \ldots, 5)$ and latent prevalences (x_{i1}, x_{i2}) were generated to be

- independent:
  $z_{im1} \sim$ Bernoulli(0.4), $z_{im2} \sim$ Normal(0, 1), $i = 1, \ldots, N$ for each m,
  $x_{i1} \sim$ Bernoulli(0.6), $x_{i2} \sim$ Normal(0, 1), $i = 1, \ldots, N$,
  all z_{imq} and x_{ip} are mutually independent;
- weakly correlated:
  $(x_{i1}, z_{i11}, \ldots, z_{i51}) \sim$ Multinomial(1; 0.1, 0.18, ..., 0.18),
  $(x_{i2}, z_{i12}, \ldots, z_{i52}) \sim$ Multinormal($\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\sigma}_{0.2}^2$), where $\boldsymbol{\sigma}_{0.2}^2$ has 1 in the diagonal and 0.2 in others; and
- highly correlated:
  $(x_{i1}, z_{i11}, \ldots, z_{i51}) \sim$ Multinomial(1; 0.5, 0.1, ..., 0.1),
  $(x_{i2}, z_{i12}, \ldots, z_{i52}) \sim$ Multinormal($\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\sigma}_{0.8}^2$).

We fit each model under several different sample sizes. For the three-class RLCAs, the selected sample sizes were 200, 500 and 1,500, which gave roughly 6, 16 and 50 individuals per parameter of RLCA (2), respectively. For the six-class RLCAs, we set $N = 500$, 1,500 and 3,000, which gave 5, 15 and 30 individuals per parameter, respectively. The observable measurements $\mathbf{Y}_i$ were then generated from each different model structure with 100 replications.

Distributions of the estimated numbers of classes from different model settings are displayed in Figure 1 (three-class RLCAs), and Figure 2 (six-class RLCAs). The notable features are: first, in the three-class RLCA, the proposed method tends to give correct class number estimates for randomly selected model parameters but overestimate the number of classes for the equal parameter setting. The distributions of the estimated numbers of classes are similar between moderate and rapid parameter settings where the method tends to underestimate the number as the sample size became larger. In the six-class RLCA, the proposed method is more likely to give the correct estimated values for equal and moderate parameter settings than for other parameter settings. Second, comparing the estimated numbers of classes among models with different covariate associations, the estimated numbers are more likely to be accurate for models with independent $\mathbf{x}_i$ and $\mathbf{z}_i$ than for models with correlated $\mathbf{x}_i$ and $\mathbf{z}_i$. Moreover, the higher the correlation, the more likely are the estimated numbers to be too low. We also found that moderate and rapid parameter settings and large sample sizes can inflate the underestimation created by the violation of independence assumption. Third, as the sample size per parameter increases, the proposed method's tendency to underestimate the number of classes increases. This tendency becomes more apparent when the association between $\mathbf{x}_i$ and $\mathbf{z}_i$ is stronger. Nonetheless, the proposed procedure gives a reasonable prediction of the number of classes to fit, in all cases, only rarely overestimating the number of classes at $N > 200$ or underestimating the number by more than one. Generally speaking, our proposed approach under the settings of the equal or moderate parameter structure, independent

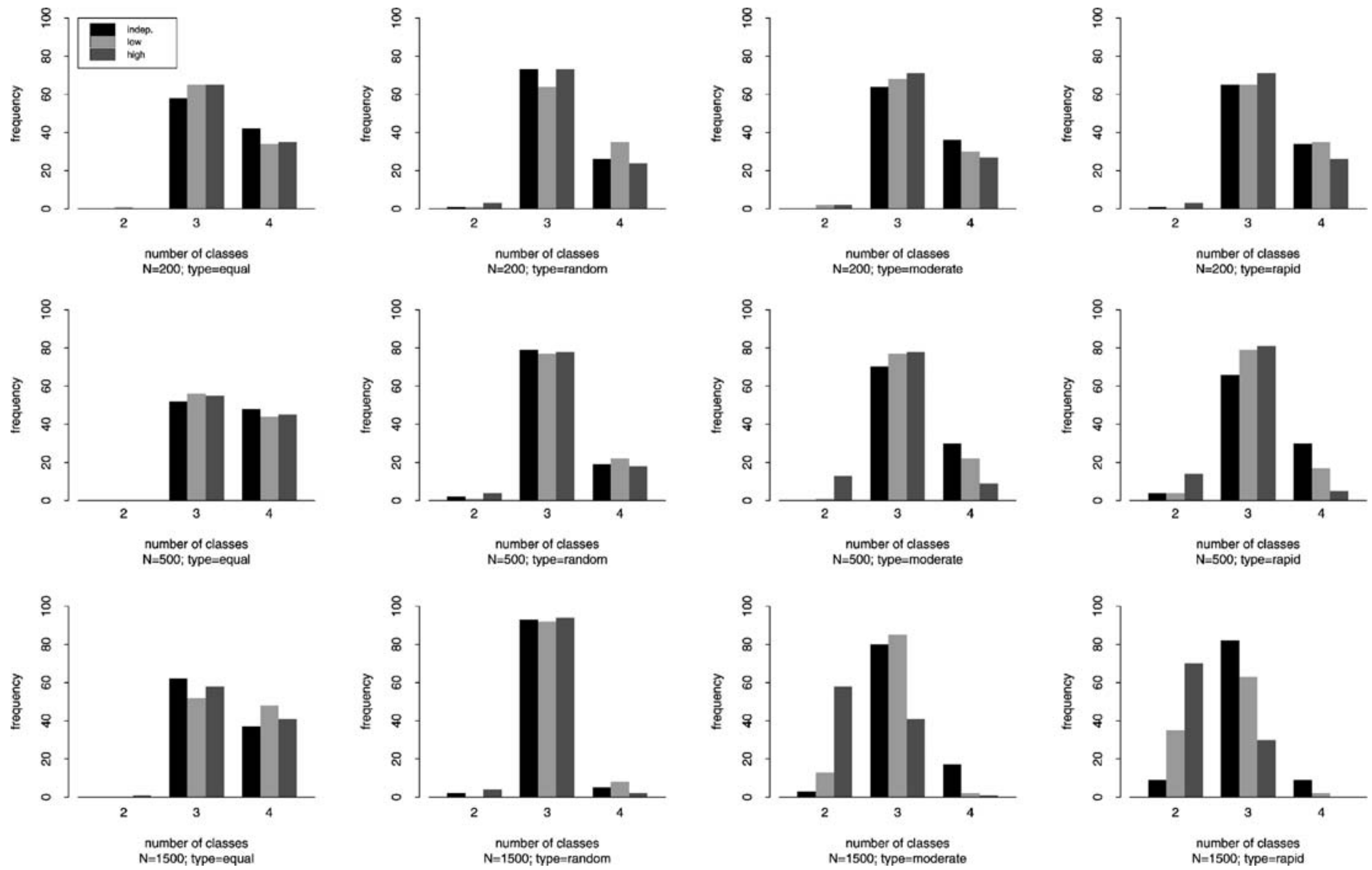

FIGURE 1.
Distributions of the estimated numbers of classes based on the simulated data from three-class RLCAs. In each plot, *black bars* are for independent covariate sets, *light bars* for weakly correlated covariates, and *dark bars* for highly correlated covariates. Rows give plots of (*left to right*) equal, randomly selected, moderately decreasing and rapidly decreasing parameters.

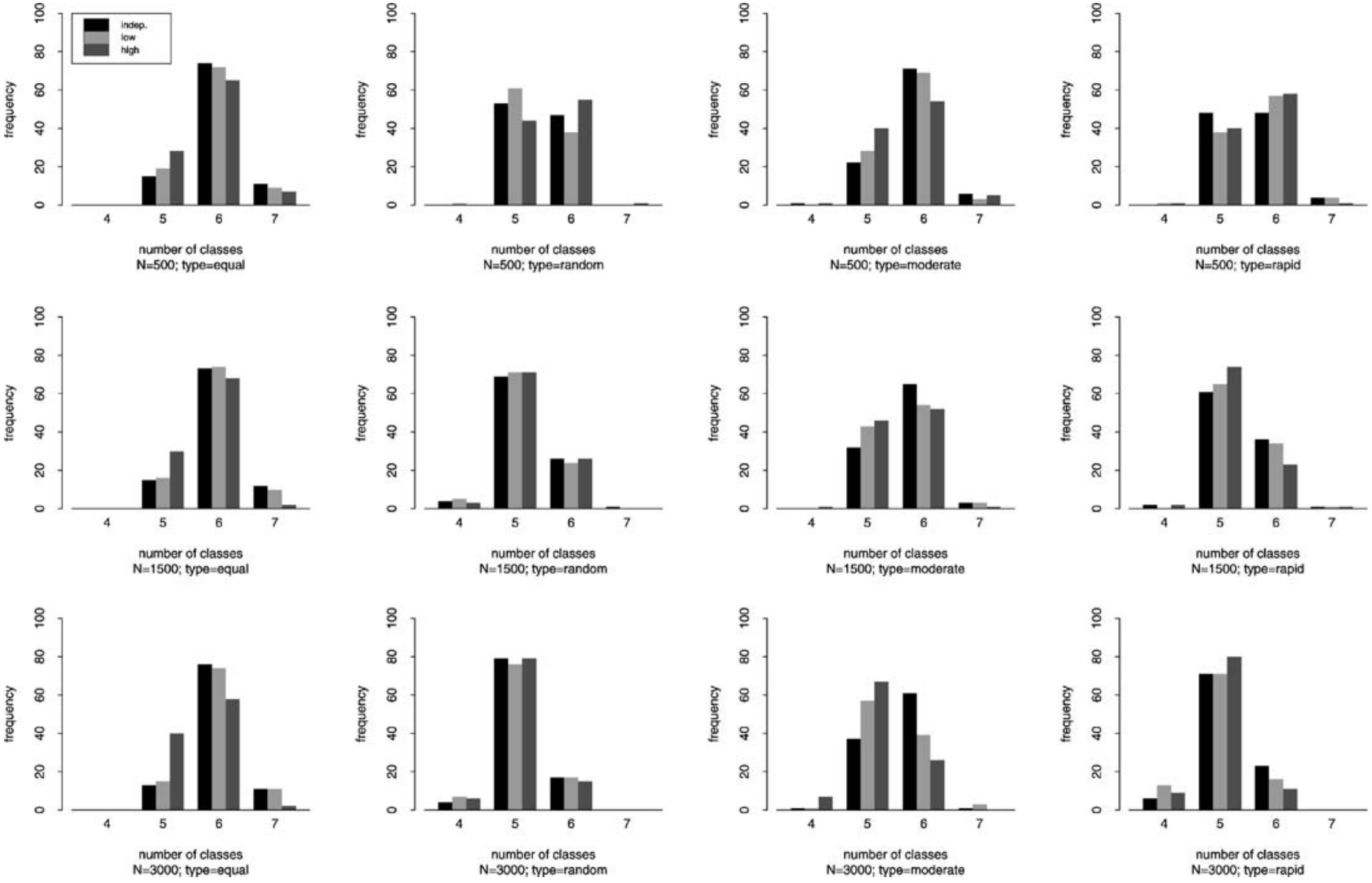

FIGURE 2.
Distributions of the estimated numbers of classes based on the simulated data from six-class RLCAs. In each plot, *black bars* are for independent covariate sets, *light bars* for weakly correlated covariates, and *dark bars* for highly correlated covariates. Rows give plots of (*left to right*) equal, randomly selected, moderately decreasing and rapidly decreasing parameters.

$\mathbf{x}_i$ and $\mathbf{z}_i$, or small sample sizes per parameter tends to give a more accurate class number estimate than under other types of settings.

4.2. Comparison to Alternatives

We compare our proposed method with the three most frequently used approaches: (a) the standard approach, where the RLCA model is fit under different numbers of classes and the selected class number is the lowest number of classes that yield acceptable fit under the likelihood ratio goodness of fit test; (b) the AIC criterion, where the estimated number of classes is fixed at the class number J that minimizes $-2\log L + 2\cdot T$ with $\log L = \sum_{i=1}^{N} \log \Pr(\mathbf{Y}_i|\mathbf{x}_i, \mathbf{z}_i)$ being the log likelihood function and T the total number of parameters in the RLCA model; and (c) the BIC criterion, where the estimated number of classes is fixed at J that minimizes $-2\log L + \log(N)\cdot T$ with N being the total number of observations.

In this simulation, we focus on comparing the performances of different approaches under "ideal" conditions. Therefore, we limited the simulation to models that fit the required assumptions of each approach. To meet the independent assumption **(C5)**, the covariates associated with conditional probabilities and the covariates associated with latent prevalences were created independently. To avoid sparse response patterns that might invalidate goodness of fit test, all incorporated covariates were binary. Each approach was applied to observed measurements $\mathbf{Y}_i$, generated by two different RLCA (2) models. One was a three-class RLCA with $M = 5$, $K_1 = \cdots = K_5 = 2$, $P = L = 1$, and the other was a six-class RLCA with $M = 5$, $K_1 = \cdots = K_5 = 3$, $P = L = 1$. For both models, all parameters were randomly generated from U(0,1) (the "random" parameter setting described in the previous subsection), and the binary covariates associated with conditional probabilities $z_{im1}, m = 1, \ldots, 5$ and latent prevalences x_{i1} were mutually independent (the "independent" covariates following the Bernoulli distribution described in the previous subsection). As we discuss in the introduction, the standard approach is sensitive to sparse response patterns. We therefore used large enough sample sizes to avoid the sparseness problem. For the three-class RLCA, the selected sample sizes were 500 and 1,500, which gave 16 and 47 individuals per response pattern of measured indicators, respectively. For the six-class RLCA, we set $N = 3{,}000$ and 6,000, which gave 12 and 25 individuals per response pattern, respectively. 100 replications were performed for each generated RLCA model.

Results of the simulation are shown in Figure 3. In the three-class model, our approach performs best and the BIC performs worst among all approaches for the sample size 500. As a sample size of 1,500, all four approaches perform almost equally well; in fact, our approach is a bit more likely to underestimate the number of classes than other approaches. When the true number of classes is six, our approach tends to underestimate the number of classes by one. Our approach's performance is the best among all approaches for the six-class model, and the outperforming is more apparent for $N = 3{,}000$ than for $N = 6{,}000$. The standard approach and the BIC criterion are likely to underestimate the number by more than two. The AIC criterion can overestimate the number of classes as the sample size becomes larger. In summary, our approach is able to provide a good number-of-classes estimation and tends to underestimate the number by one as the sample size per response pattern became larger. The standard approach and BIC criterion require very "unsparse" data to obtain correct estimates. The AIC criterion may overestimate the number of classes for large sample sizes. Our approach can outperform existing methods more when the data become sparse.

5. Example

To illustrate the proposed selecting method, we use data from the Salisbury Eye Evaluation (SEE) project. The SEE project is described in detail in West, Munoz, Rubin, Schein, Bandeen-

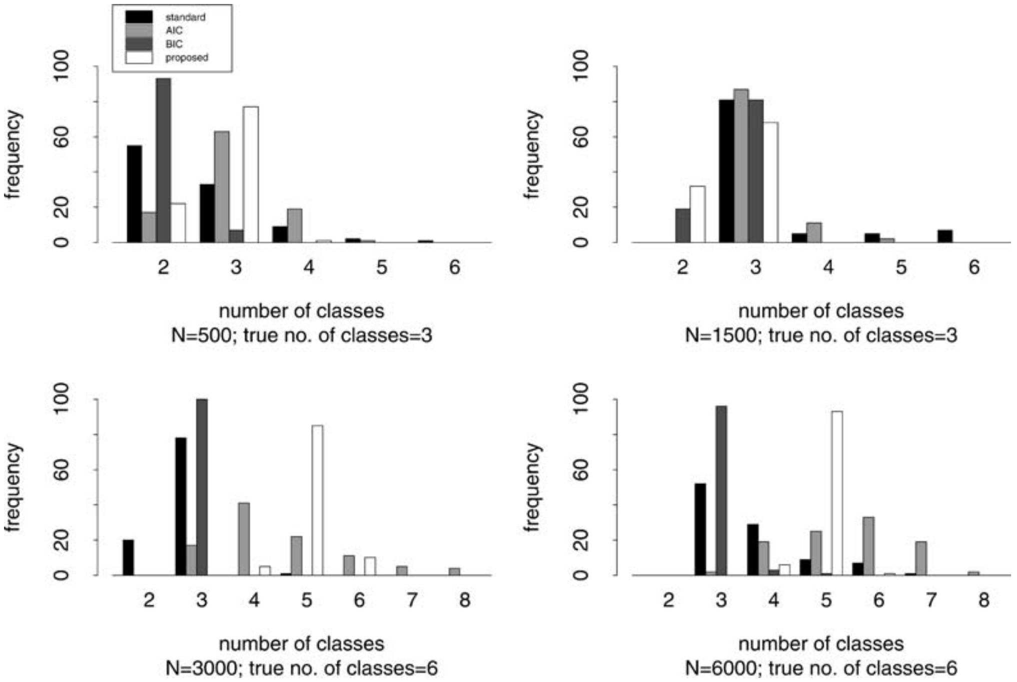

FIGURE 3.
Comparison among four different approaches in estimating the number of classes. In each plot, *black bars* are based on the standard approach, *light bars* the AIC criterion, *dark bars* the BIC criterion and *white bars* the proposed approach. Clockwise from top left, plots describe results from: the three-class model with sample size 500, three-class model with sample size 1,500, six-class model with sample size 6,000 and six-class model with sample size 3,000.

Roche, Zeger, German, and Fried (1997). Briefly, SEE is a population-based, prospective study of how vision affects functioning in older persons. An age- and race-stratified random sample of Salisbury, Maryland residents between the ages of 65 and 84 years was drawn from the Health Care Financing Administration (HCFA) Medicare Database. Twenty-five hundred and twenty persons agreed to participate in both home interview and clinic examination.

The analysis we report in this paper aims to describe the association between functioning in activities that require seeing at a distance (far vision functioning) and psychophysical measures of visual impairment, adjusting for potential confounding variables. In the SEE project, far vision functioning was determined using the self-reported difficulty on doing five activities: reading street signs in daylight, reading street signs at night, walking down steps during daylight, walking down steps in dim light, and watching TV. Here, we measured difficulty as a binary indicator (1=having difficulty; 2=no difficulty) for each activity, except for reading street signs at night which was measured as a three-level categorical indicator (1=extreme or moderate difficulty; 2=a little difficulty; 3=no difficulty). Visual impairment was determined using multiple psychophysical vision tests (Rubin, West, Munoz, Bandeen-Roche, Zeger, Schein, and Fried, 1997). Our analysis includes five tests: visual acuity, contrast sensitivity, glare sensitivity, stereoacuity, and central visual field.

Three latent class models for self-reported far vision functioning (measured indicators) were fitted. We started with an LCA model, which did not incorporate any covariates. Then, two different RLCA models (2) were fitted as a function of visual impairment variables, the number of reported comorbid diseases, and the following personal demographic characteristics: age, Mini-Mental State Examination score (Folstein, Folstein, and McHugh, 1975), years of education, gender, race, and General Health Questionnaire depression subscale score (Goldberg, 1972). The first RLCA model (RLCA1) treated the vision and disease variables as primary predictors of latent class membership ($\mathbf{x}_i$) and modelled the personal characteristics as having direct effects on measured indicators themselves within classes ($\mathbf{z}_i$). $\mathbf{x}_i$'s would help us obtain the effect of visual impairment on the underlying far vision functioning. We adjusted for $\mathbf{z}_i$'s, which were identified as extraneous influences (other than the underlying far vision functioning) that affected the individual's reporting in the questionnaire, hence hopefully could yield a more accurate latent class. It is arguable that age seems a good predictor of the underlying latent class. The second RLCA model (RLCA2) allowed age to affect both class membership and measured indicators themselves. Notice that, in RLCA2, two sets of covariates $\mathbf{x}_i$ and $\mathbf{z}_i$ are not independent. The analysis was applied to the subsample of participants who rated each far vision item and also had no missing covariates ($N = 1,641$). While estimating each model, several sets of starting points were used to ensure global maxima.

The proposed method, standard approach, AIC criterion and BIC criterion were used to determine the appropriate class numbers of different latent class models. For the LCA model, where no covariates were incorporated in predicting conditional probabilities, the proposed method used measured indicators as pseudo-residuals. Various elements of the covariance matrix of measured indicators are

$$\mathrm{Cov}(Y_{imk}, Y_{iqs}) = \begin{cases} \Pr(Y_{imk} = 1) - \Pr(Y_{imk} = 1)\Pr(Y_{iqs} = 1) & \text{if } m = q \text{ and } k = s \\ -\Pr(Y_{imk} = 1)\Pr(Y_{iqs} = 1) & \text{if } m = q \text{ and } k \neq s \\ \Pr(Y_{imk} = 1, Y_{iqs} = 1) - \Pr(Y_{imk} = 1)\Pr(Y_{iqs} = 1) & \text{if } m \neq q. \end{cases}$$

These variances were estimated by replacing the probabilities with the sample averages. Eigenvalues of the estimated correlation matrix are 2.69, 1.02, 0.81, 0.70, 0.40 and 0.38. Therefore, the proposed method estimates the number of latent classes equal to three. The goodness-of-fit tests for LCA models with 2-, 3-, 4- and 5-class result in p-values < 0.001, 0.02 and 0.07 respectively. The standard approach gives a five-class model. Results of AIC and BIC for LCA are shown in Table 1. Both AIC and BIC criteria result in four classes.

TABLE 1.
AIC and BIC criteria for selecting the number of latent classes: the SEE-project far vision functioning data

Model/Method	AIC	BIC
LCA		
2-class	6579.69	6649.93
3-class	6490.46	6598.52
4-class	6390.71	6536.60
5-class	6392.96	6576.66
RLCA1*		
2-class	6309.38	6606.55
3-class	6153.87	6521.28
4-class	6061.47	6499.11
5-class	6064.29	6572.18
RLCA2*		
2-class	6308.87	6611.44
3-class	6157.80	6536.02
4-class	6064.98	6518.84
5-class	6067.56	6597.06

Note: RLCA1 = the regression extension of latent class analysis model with age effect only on conditional probabilities; RLCA2 = the regression extension of latent class analysis model with age effect on both latent prevalences and conditional probabilities.

Because continuous covariates were incorporated in the RLCA models, sparseness of response patterns failed χ^2 approach and the standard approach was thereby invalid. For RLCA1, the sample correlation matrix of pseudo-residuals as defined in Step 2 of the proposed algorithm has eigenvalues 3.10, 1.03, 0.70, 0.52, 0.40 and 0.26, and, thus, the estimated class number based on the proposed method is three. AIC and BIC both estimate the class number of RLCA1 as four (Table 1). Because RLCA2 has the same incorporated covariates on conditional probabilities as RLCA1, the sample correlation matrices of pseudo-residuals for both models are the same, therefore, the proposed method estimates a three-class RLCA2 model. From Table 1, AIC and BIC criteria select the four-class RLCA2 model.

All selection criteria give consistent class number estimates across LCA, RLCA1 and RLCA2. The number of dimensions needed to characterize self-reported far vision functioning does not confound with adjusted factors. Huang and Bandeen-Roche (in press) analyzed the SEE far vision data and adopted a four-class RLCA1 model based on AIC and BIC criteria. The model diagnosis revealed an appropriate fit to the data. A similar diagnostic approach is applied to the three-class RLCA1 model (not shown) and shows a reasonable model fit. Comparing the estimated latent classes of three- (not shown) and four-class models (Table 1 of Huang and Bandeen-Roche, in press), we find that the three-class model maintains the able and severely disabled classes in the four-class model, but combines the reading-signs difficulty and descending-steps difficulty classes as one class. The failure of separating out the descending-steps difficulty class using the proposed method may be due to high correlation between two types of covariates $\mathbf{x}_i$ and $\mathbf{z}_i$ (e.g., the sample correlation was 0.27 between age and visual acuity, 0.34 between age and contrast sensitivity) and relatively large sample size per parameter (30, 24, 20 and 17 individuals per parameter for 2-, 3-, 4- and 5-class model, respectively). As seen in the simulation study, these factors might cause the proposed method to underestimate the number of classes.

When implementing the proposed method in analyzing the SEE data, we did the following: (1) We started with a three-class RLCA model and performed the model diagnosis. (2) Diagnostic results revealed a reasonable fit with mild model violation; thus, we augmented the class number and refit a four-class model. (3) The same diagnostic approach was applied and showed an improved and satisfied model fit. (4) We consulted with the SEE scientists to ensure the resulting four classes having meaningful interpretation.

6. Discussion

In this article, a computationally simple method was proposed to choose the number of classes to fit. A connection between finite mixture models and factor analysis was built, such that commonly used rules for determining the number of components to retain in factor analysis could be applied to select the number of classes to fit in finite mixture models. We then used the marginalization technique to reduce the complexity of latent class models with covariate effects on both latent prevalences and conditional probabilities to the level of latent class analysis, so that the built connection could be applied to the marginalized models. A computer module of implementing the RLCA model (2) is created using statistical package S-PLUS (Statistical Science Inc., 1995) and programming language C. It provides initial values for the estimation, parameter and variance estimates, model identifiability checking, the number of latent classes selection, and graphical displays for model diagnosis. This computer module is available from the author (e-mail: ghuang@stat.nctu.edu.tw).

The proposed procedure does not require repeatedly fitting RLCA model as in traditional goodness-of-fit methods, and can provide an estimate of the number of classes when a prior knowledge does not mandate the class number. A simulation study showed that this method provides a reasonably accurate estimate of the class number and performs better than other existing approaches. The proposed method tends to give a lower class-number estimate when there is high correlation between covariates for latent prevalences and covariates for conditional probabilities and large sample sizes. In practical use, we suggest that readers select an initial estimate of the class number by combining the proposed procedure and scientific considerations (e.g., prior knowledge about the class number and enhancing the interpretability of the resulting latent class). The inferences of RLCA may be done by fixing the number of classes at the selected number, and the analyst may then diagnose the model fit. If a poor fit is found, the implications for inference and interpretation must be elucidated. For descriptive analyses, the analyst might augment the class number and re-estimate the model.

From the simulation results, the proposed methods were more likely to underestimate the number of classes at larger sample sizes. This differs from the phenomenon seen in determining the number of components in principal component analysis, where the trend is toward more accurate estimates as the sample size increases (Humphreys, 1964; Francisco and Finch, 1979). The possible reasons for this difference may be: (a) From Theorem 1, the proposed method is actually based on a lower bound. When the sample size gets larger, population and sample correlation matrices become closer; therefore, we are more likely to obtain the lower bound. (b) The proposed method must marginalize the covariate effects from conditional probabilities, and large sample sizes inflate errors from the marginalization process. The simulation study shows that in most cases, the proposed method underestimated the number of classes by less than two.

We adopted Guttman's (1954) criterion to choose a class number that approximates (24) reasonably well. This criterion is the most commonly used rule in factor analysis, and several studies have found that it often leads to accurate results when the eigenvalues of the correlation matrix are high and the number of measured variables is moderate (Humphreys, 1964; Francisco and Finch, 1979). However, many authors have pointed out that Guttman's criterion can grossly overestimate the number of factors and be inconsistent in cases with low eigenvalues and a large number of variables (Cattell and Vogelmann, 1977; Linn, 1968). One alternative method for choosing the number of factors is the scree test (Cattell, 1966). This test successively plots the eigenvalues of the correlation matrix from large to small, and picks the estimated number of factors as the lowermost point that contributes to "substantive" down of the plot (Cattell, 1966). Unlike Guttman's criterion looking at the absolute value, the scree test provides the distribution of eigenvalues, and researchers can then incorporate scientific considerations to choose the number of factors. In our selecting procedure, the scree plot might be useful in cases where several eigenvalues are closed around one, and the plot can provide useful information to modify the estimated number of classes.

In developing our method of selecting the number of classes to fit, we derived a marginalization technique and a connection between finite mixture models and factor analysis. These two properties are applicable to many applications besides ours. First, in ordinary linear regression, graphical diagnostic displays have proven very useful for detecting lack of fit of a model to data. The discreteness of categorical outcomes makes it difficult to interpret such displays. Several authors (Landwehr et al., 1984; Wang, 1985, 1987; O'Hara Hines and Carter, 1993) had developed residuals under the generalized linear model framework, and successfully implemented these residuals in creating diagnostic plots. In marginalizing covariate effects of RLCA, we provided a formula for residuals of categorical responses, which were a "vector"-version extension of previous residuals, and could be applied broadly. Second, the link between finite mixture and factor analysis has another important implication. The method of fitting used in factor analysis could be used to estimate $\mathbf{\Omega}$ and $\mathbf{\Psi}$ in (13), and then the estimated conditional expectations $\hat{\boldsymbol{\mu}}_m(j)$ and variances $\hat{\mathbf{\Sigma}}_m(j)$ could be obtained through (10) and (11). These estimated conditional expectations and variances could help to determine the structure of latent class model. By appropriate transformation, they might provide alternative estimates for model parameters in LCA.

Appendix A: Justification of the Marginalization Procedure for $\mathbf{z}_i$

To formally justify the marginalization procedure for $\mathbf{z}_i$, let's treat $\mathbf{S}_{im}$ as one FIXED matrix (i.e., the class memberships of all individuals are pre-determined). We first estimate $\boldsymbol{\gamma}_m$ and $\boldsymbol{\alpha}_m$ from (15). By IRLS, the $(t+1)$th iteration can be written as

$$\begin{bmatrix} \hat{\boldsymbol{\gamma}}_m(t+1) \\ \hat{\boldsymbol{\alpha}}_m(t+1) \end{bmatrix} = \left(\begin{bmatrix} \mathbf{S} \\ \mathbf{Z}^c_m \end{bmatrix} \hat{\mathbf{V}}^a_m(t) \begin{bmatrix} \mathbf{S}^{\mathrm{T}} & (\mathbf{Z}^c_m)^{\mathrm{T}} \end{bmatrix} \right)^{-1} \left(\begin{bmatrix} \mathbf{S} \\ \mathbf{Z}^c_m \end{bmatrix} \hat{\mathbf{V}}^a_m(t) \quad \mathbf{Y}^a_m(t) \right),$$

where $\hat{\mathbf{V}}^a_m(t)$ refers to the value of $\mathbf{V}_m$ evaluated at $\hat{\boldsymbol{\gamma}}_m(t)$ and $\hat{\boldsymbol{\alpha}}_m(t)$, and $\mathbf{Y}^a_m(t) = \mathbf{S}^{\mathrm{T}}\hat{\boldsymbol{\gamma}}_m(t) + (\mathbf{Z}^c_m)^{\mathrm{T}}\hat{\boldsymbol{\alpha}}_m(t) + \hat{\mathbf{V}}^a_m(t)^{-1}[\mathbf{Y}_m - \hat{\boldsymbol{\mu}}^a_m(t)]$ with $\hat{\boldsymbol{\mu}}^a_m(t) = \mathrm{E}(\mathbf{Y}_m|\mathbf{S}_m, \mathbf{Z}^c_m)|_{\boldsymbol{\gamma}_m=\hat{\boldsymbol{\gamma}}_m(t);\boldsymbol{\alpha}_m=\hat{\boldsymbol{\alpha}}_m(t)}$. Since $\mathbf{S}\mathbf{W}_m(\mathbf{Z}^c_m)^{\mathrm{T}} = 0$ ((**C4**)), then

$$\begin{bmatrix} \hat{\boldsymbol{\gamma}}_m(t+1) \\ \hat{\boldsymbol{\alpha}}_m(t+1) \end{bmatrix} = \begin{bmatrix} (\mathbf{S}\hat{\mathbf{V}}^a_m(t)\mathbf{S}^{\mathrm{T}})^{-1}(\mathbf{S}\hat{\mathbf{V}}^a_m(t)\mathbf{Y}^a_m(t)) \\ (\mathbf{Z}^c_m\hat{\mathbf{V}}^a_m(t)(\mathbf{Z}^c_m)^{\mathrm{T}})^{-1}(\mathbf{Z}^c_m\hat{\mathbf{V}}^a_m(t)\mathbf{Y}^a_m(t)) \end{bmatrix}.$$

Also, the $(t+1)$th iteration estimate of $\boldsymbol{\gamma}^*_m$ from linear model (20) is

$$\hat{\boldsymbol{\gamma}}^*_m(t+1) = (\mathbf{S}\hat{\mathbf{V}}^*_m(t)\mathbf{S}^{\mathrm{T}})^{-1}(\mathbf{S}\hat{\mathbf{V}}^*_m(t)\mathbf{R}_m) = (\mathbf{S}\hat{\mathbf{V}}^*_m(t)\mathbf{S}^{\mathrm{T}})^{-1}(\mathbf{S}\hat{\mathbf{V}}^*_m(t)\mathbf{Y}^*_m),$$

where $\hat{\mathbf{V}}^*_m(t)$ is defined as its counterpart above but evaluated at $\hat{\boldsymbol{\gamma}}^*_m(t)$. Suppose that both variance estimates are very close to the true variance (i.e., $\hat{\mathbf{V}}^a_m(t) \approx \hat{\mathbf{V}}^*_m(t) \approx \mathbf{V}_m$). Then, $\hat{\boldsymbol{\gamma}}^*_m$ from (20) is close to $\hat{\boldsymbol{\gamma}}_m$ from (15) if the difference between $\mathbf{Y}^a_m$ and $\mathbf{Y}^*_m$ is negligible. The difference between $\mathbf{Y}^a_m$ and $\mathbf{Y}^*_m$ can expressed as

$$\mathbf{Y}^a_m - \mathbf{Y}^*_m = [(\mathbf{S}^{\mathrm{T}}\hat{\boldsymbol{\gamma}}_m + (\mathbf{Z}^c_m)^{\mathrm{T}}\hat{\boldsymbol{\alpha}}_m) - (\mathbf{Z}^c_m)^T\hat{\boldsymbol{\alpha}}^*_m] - \mathbf{V}^{-1}_m[\hat{\boldsymbol{\mu}}^a_m - \hat{\boldsymbol{\mu}}^*_m].$$

The first part of right-hand-side is the difference between pseudo-fitted-values from (15) and (16), and the second part is the difference of pseudo-residuals between two models. This negligible assumption assumes that $\mathbf{S}_m$ contributions on fitted values and residuals are almost the same. Under ordinary linear models (i.e., identical link and identity correlation matrix), two differences are the same. Therefore, it is reasonable to make such an approximation.

Appendix B: Proof of Theorem 1

For simplicity, we assume $K_1 = \cdots = K_M = K$ (i.e., the levels of items are all the same) in the following proof. Extension to allow the levels being different is straightforward.

From (10, 11, 12),

$$\mathrm{R}(\mathbf{R}^p_{i1}, \ldots, \mathbf{R}^p_{iM}) = \mathbf{V}^{-1/2}\mathbf{L}\mathbf{L}^T\mathbf{V}^{-1/2} + \mathbf{V}^{-1/2}\boldsymbol{\Psi}\mathbf{V}^{-1/2} = \mathbf{G}\mathbf{G}^{\mathrm{T}} + \mathbf{E}, \tag{A1}$$

where

$$\mathbf{V} = \mathrm{diag}(\mathrm{Var}(\mathbf{R}^p_{i1}), \ldots, \mathrm{Var}(\mathbf{R}^p_{iM})) =: \mathrm{diag}(\mathbf{D}_{11}, \ldots, \mathbf{D}_{MM}),$$

$$\mathbf{G} = \mathbf{V}^{-1/2}\mathbf{L} = \begin{bmatrix} \mathbf{D}_{11}^{-\frac{1}{2}}\mathbf{l}_{11} & \mathbf{D}_{11}^{-\frac{1}{2}}\mathbf{l}_{12} & \cdots & \mathbf{D}_{11}^{-\frac{1}{2}}\mathbf{l}_{1J} \\ \mathbf{D}_{22}^{-\frac{1}{2}}\mathbf{l}_{21} & \mathbf{D}_{22}^{-\frac{1}{2}}\mathbf{l}_{22} & \cdots & \mathbf{D}_{22}^{-\frac{1}{2}}\mathbf{l}_{2J} \\ \vdots & \vdots & & \vdots \\ \mathbf{D}_{MM}^{-\frac{1}{2}}\mathbf{l}_{M1} & \mathbf{D}_{MM}^{-\frac{1}{2}}\mathbf{l}_{M2} & \cdots & \mathbf{D}_{MM}^{-\frac{1}{2}}\mathbf{l}_{MJ} \end{bmatrix} =: \begin{bmatrix} \mathbf{G}_1 \\ \mathbf{G}_2 \\ \vdots \\ \mathbf{G}_M \end{bmatrix},$$

$$\mathbf{E} = \mathbf{V}^{-1/2}\boldsymbol{\Psi}\mathbf{V}^{-1/2} = \mathrm{diag}\left(\sum_{j=1}^{J}\{\eta_j^*\mathbf{D}_{11}^{-\frac{1}{2}}\boldsymbol{\Sigma}_1(j)\mathbf{D}_{11}^{-\frac{1}{2}}\}, \cdots, \sum_{j=1}^{J}\{\eta_j^*\mathbf{D}_{MM}^{-\frac{1}{2}}\boldsymbol{\Sigma}_M(j)\mathbf{D}_{MM}^{-\frac{1}{2}}\}\right).$$

Let

$$\mathbf{B} = \mathrm{R}(\mathbf{R}_{i1}^{p}, \cdots, \mathbf{R}_{iM}^{p}) - \mathrm{I}_{(K-1)M}, \tag{A2}$$

where $\mathrm{I}_{(K-1)M}$ is a $(K-1)M \times (K-1)M$ identity matrix. From (A1) and (A2),

$$\mathbf{G}\mathbf{G}^{\mathrm{T}} = \mathbf{B} + (\mathrm{I}_{(K-1)M} - \mathbf{E}). \tag{A3}$$

Since $\mathbf{E}$ is a block diagonal matrix, the eigenvalues of $(\mathrm{I}_{(K-1)M} - \mathbf{E})$ are equal to the eigenvalues of each of the blocks $(\mathrm{I}_{K-1} - \mathbf{E}_1), \ldots, (\mathrm{I}_{K-1} - \mathbf{E}_M)$, where $\mathbf{E}_m = \sum_{j=1}^{J}\{\eta_j^*\mathbf{D}_{mm}^{-\frac{1}{2}}\boldsymbol{\Sigma}_m(j)\mathbf{D}_{mm}^{-\frac{1}{2}}\}$, $m = 1, \ldots, M$. Notice that, from (A1),

$$\mathbf{E}_m = \mathrm{I}_{K-1} - \mathbf{G}_m\mathbf{G}_m^{\mathrm{T}}, \; m = 1, \ldots, M. \tag{A4}$$

Since $J - 1 \geq K - 1$ (assumption **(C6)**) and conditions **(C7)** and **(C9)**, the rank of $\mathbf{G}_m$ is equal to $K - 1$. $(\mathrm{I}_{K-1} - \mathbf{E}_1), \ldots, (\mathrm{I}_{K-1} - \mathbf{E}_M)$ are then all positive definite (Graybill, 1983, Theorem 1.7.6 and 1.7.7). So, $(\mathrm{I}_{(K-1)M} - \mathbf{E})$ is positive definite and symmetric. Therefore, there exists a non-singular matrix $\mathbf{F}$ such that $(\mathrm{I}_{(K-1)M} - \mathbf{E}) = \mathbf{F}\mathbf{F}^{\mathrm{T}}$ (Strang 1976, p. 253).

Let $\mathbf{G_F} = \mathbf{F}^{-1}\mathbf{G}\mathbf{G}^{\mathrm{T}}(\mathbf{F}^{\mathrm{T}})^{-1}$ and $\mathbf{B_F} = \mathbf{F}^{-1}\mathbf{B}(\mathbf{F}^{\mathrm{T}})^{-1}$. Then premultiplying both sides of (A3) by $\mathbf{F}^{-1}$ and postmultiplying by $(\mathbf{F}^{\mathrm{T}})^{-1}$ yield

$$\mathbf{G_F} = \mathbf{B_F} + \mathrm{I}_{(K-1)M}. \tag{A5}$$

From (A5), the eigenvalues of $\mathbf{G_F}$ are the same as those of $\mathbf{B_F}$ each increased by one. So, if g_F is the number of positive eigenvlaues of $\mathbf{G_F}$ and b_F is the number of non-negative eigenvalues of $\mathbf{B_F}$, it must be that

$$g_F \geq b_F. \tag{A6}$$

From Sylvester's law of intertia (Strang 1976, p. 259), the number of eigenvalues of $\mathbf{G_F}$ of a given sign is the same as for $\mathbf{G}\mathbf{G}^{\mathrm{T}}$. The same invariance holds between the signs of eigenvalues of $\mathbf{B}$ and $\mathbf{B_F}$. By Theorem 1.7.6 of Graybill (1983), the rank of $\mathbf{G}\mathbf{G}^{\mathrm{T}}$ is equal to the rank of $\mathbf{G}$. Because of conditions **(C8)** and **(C9)**, the rank of $\mathbf{G}$ is $J - 1$. $\mathbf{G}\mathbf{G}^{\mathrm{T}}$ is positive semi-definite, and therefore its rank equals the number of positive eigenvalues of $\mathbf{G}\mathbf{G}^{\mathrm{T}}$, i.e., $g_F = J - 1$. Also since (A2), the number of $\mathrm{R}(\mathbf{R}_{i1}^{p}, \ldots, \mathbf{R}_{iM}^{p})$'s eigenvalues that are not less than one (i.e., r) is the same as the number of non-negative eigenvalues of $\mathbf{B}$. Therefore, by (A6), we can get $J - 1 \geq r$, which gives the lower bound of the true number of classes.

References

Agresti, A. (1984). *Analysis of Categorical Data*. New York: J. Wiley and Sons.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*, 317–332.

Bandeen-Roche, K., Miglioretti, D.L., Zeger, S.L., & Rathouz, P.J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, *92*, 1375–1386.

Bartholomew, D.J., & Knott, M. (1999). *Latent Variable Models and Factor Analysis* (2nd ed.). Kendall Library of Statistics. London: Arnold.

Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245–276.

Cattell, R.B., & Vogelmann, S. (1977). A comprehensive trial of the scree and KG criteria for determining the number of factors. *Multivariate Behavioral Research*, *12*, 289–325.

Cook, R.D., & Weisberg, S. (1982). *Residuals and Influence in Regression*. London: Chapman Hall.

Dayton, C.M., & Macready, G.B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, *83*, 173–178.

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

Folstein, M.F., Folstein, S.E., & McHugh, P.R. (1975). Mini-mental state: a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*, 189.

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

Francisco, C.A., & Finch, M.D. (1979). A comparison of methods used for determining the number of factors to retain in factor analysis. *American Statistical Association Proceedings of the Statistical Computing Section*, 105–110.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Goldberg, D. (1972). *GHQ The Selection of Psychiatric Illness by Questionnaire*. London: Oxford University Press.

Goodman, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*, 215–231.

Graybill, F.A. (1983). *Matrices with Applications in Statistics*. Belmont: Wadsworth.

Green, B.F. (1951). A general solution of the latent class model of latent structure analysis and latent profile analysis. *Psychometrika*, *16*, 151–166.

Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika*, *19*, 149–161.

Hagenaars, J.A. (1993). *Loglinear Models with Latent Variables* Sage. University Paper series on Quantitative Applications in the Social Sciences, series no. 07–094. Newbury Park, CA: Sage Publications.

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

Humphreys, L.G. (1964). Number of cases and number of factors: an example where *N* is very large. *Educational and Psychological Measurement*, *24*, 457–466.

Kashyap, R.L. (1982). Optimal choice of AR and MA parts in autoregressive moving average models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *4*, 99–104.

Landwehr, J.M., Pregibon, D., & Shoemaker, C. (1984). Graphical methods for assessing logistic regression models. *Journal of the American Statistical Association*, *79*, 61–71.

Lazarsfeld, P.F., & Henry, N.W. (1968). *Latent Structure Analysis*. New York: Houghton-Mifflin.

Liang, K.Y., & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*, 13–22.

Linn, R. (1968). A monte carlo approach to the number of factors problem. *Psychometrika*, *33*, 37–71.

McCullagh, P., & Nelder, J.A. (1989). *Generalized Linear Models*, (2nd ed.). London: Chapman and Hall.

Melton, B., Liang, K.Y., & Pulver, A.E. (1994). Extended latent class approach to the study of familial/sporadic forms of a disease: its application to the study of the heterogeneity of schizophrenia. *Genetic Epidemiology*, *11*, 311–327.

Moustaki, I. (1996). A latent trait and a latent class model for mixed observed variables. *British Journal of Mathematical and Statistical Psychology*, *49*, 313–334.

Muthén, L.K., & Muthén, B.O. (1998). *Mplus User's Guide*. Los Angeles, CA: Muthén & Muthén.

O'Hara Hines, R.J., & Carter, E.M. (1993). Improved added variable and partial residual plots for the detection of influential observation in generalized linear models. *Applied Statistics*, *42*, 3–20.

Rubin, G.S., West, S.K., Munoz, B., Bandeen-Roche, K., Zeger, S.L., Schein, O., & Fried, L.P. (1997). A comprehensive assessment of visual impairment in an older american population: SEE study. *Investigative Ophthalmology and Visual Science*, *38*, 557–568.

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

Statistical Sciences Inc., (1995). S-PLUS User's Manual, Version 3.3 for Windows, Seattle: Statistical Sciences Inc.

Strang, G. (1976). *Linear Algebra and Its Applications*. New York: Academic Press.

Titterington, D.M., Smith, A.F.M., & Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Chichester, U. K.: Wiley.

Van der Heijden, P.G.M., Dessens, J., & Böckenholt, U. (1996). Estimating the concomitant-variable latent-class model with the EM algorithm, *Journal of Educational and Behavioral Statistics*, *21*, 215–229.

Vermunt, J.K. (1996). *Log-linear Event History Analysis: A General Approach with Missing Data, Unobserved Heterogeneity, and Latent Variables*. Tilburg: Tilburg University Press.

Vermunt, J.K., & Magidson, J. (2000). *Latent GOLD 2.0 User's Guide*. Belmont, MA: Statistical Innovations Inc.

Wang, P.C. (1985). Adding a variable in generalized linear models. *Technometrics*, *27*, 273–276.

Wang, P.C. (1987). Residual plots for detecting nonlinearity in generalized linear models. *Technometrics*, *29*, 435–438.

Wedel, M., Desarbo, W.S., Bult, J.R., & Ramaswamy, V. (1993). A latent class poisson regression model for heterogeneous count data. *Journal of Applied Econometrics*, *8*, 397–411.

West, S.K., Munoz, B., Rubin, G.S., Schein, O.D., Bandeen-Roche, K., Zeger, S.L., German, P.S., & Fried, L.P. (1997). Function and visual impairment in a population-based study of older adults: SEE project. *Investigative Ophthalmology and Visual Science*, *38*, 72–82.

Yakowitz, S.J., & Spragins, J.D. (1968). On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, *39*, 209–214.

Manuscript received 29 APR 2003
Final version received 7 JAN 2004