PSYCHOMETRIKA—VOL. 76, NO. 4, 584–611
OCTOBER 2011
DOI: 10.1007/S11336-011-9227-3

OPTIMIZATION-BASED MODEL FITTING FOR LATENT CLASS AND LATENT PROFILE ANALYSES

GUAN-HUA HUANG, SU-MEI WANG, AND CHUNG-CHU HSU

INSTITUTE OF STATISTICS
NATIONAL CHIAO TUNG UNIVERSITY, TAIWAN

Statisticians typically estimate the parameters of latent class and latent profile models using the Expectation-Maximization algorithm. This paper proposes an alternative two-stage approach to model fitting. The first stage uses the modified k-means and hierarchical clustering algorithms to identify the latent classes that best satisfy the conditional independence assumption underlying the latent variable model. The second stage then uses mixture modeling treating the class membership as known. The proposed approach is theoretically justifiable, directly checks the conditional independence assumption, and converges much faster than the full likelihood approach when analyzing high-dimensional data. This paper also develops a new classification rule based on latent variable models. The proposed classification procedure reduces the dimensionality of measured data and explicitly recognizes the heterogeneous nature of the complex disease, which makes it perfect for analyzing high-throughput genomic data. Simulation studies and real data analysis demonstrate the advantages of the proposed method.

Key words: classification, finite mixture, hierarchical clustering, high-dimensional data, k-means, microarray, two-stage approach.

1. Introduction

In many psychometric studies, the conceptually or clinically most meaningful outcome is unobservable. Hence, a set of multiple indicators is measured in place of this outcome. Latent variable models explore the relationships between unobservable outcomes and their measured indicators. These models assume that all measured indicators reflect the same unobservable outcome (the assumption of unidimensionality), and that this outcome fully explains the associations between observed indicators. This unobservable outcome can be described by a continuous variable representing individuals' positions on a scale of ability (the latent trait) (Rasch, 1960; Lazarsfeld & Henry, 1968; Moustaki, 1996), or a categorical variable identifying subpopulations or "classes" each of which has homogeneous outcome status (Goodman, 1974; Titterington, Smith, & Makov, 1985; Bandeen-Roche, Miglioretti, Zeger, & Rathouz, 1997; Huang & Bandeen-Roche, 2004). This paper focuses on the cases involving an underlying categorical variable (the latent class variable). In this case, measured indicators are independent of one another within any category of the latent variable (i.e., exhibit conditional independence). When measured indicators are categorical variables, such models are called latent class (LC) models. Models with continuous indicators are called latent profile (LP) models.

The parameters of LC/LP models are typically estimated by maximum likelihood (ML) for a fixed number of classes. Viewing the class membership as unobservable, the LC/LP model becomes a typical incomplete-data problem and the Expectation-Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) can estimate the ML parameters. Notice that the EM approach does not assign objects to classes as part of the procedure, but rather uses estimated model parameters to infer the latent classes. Assumptions used to derive the likelihood function, such

Requests for reprints should be sent to Guan-Hua Huang, Institute of Statistics, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu 30010, Taiwan. E-mail: ghuang@stat.nctu.edu.tw

as unidimensionality and conditional independence, are empirically checked through analyses stratifying on inferred class memberships (Bandeen-Roche et al., 1997). This is problematic because these inferred latent classes can be wrong if the model assumptions are violated. Also, for large number of indicators and moderate or small samples, using the EM algorithm to estimate parameters in LC/LP models is typically time-consuming and the algorithm may have difficulty converging.

LC/LP analyses may legitimately be viewed as the analog of cluster analysis. Cluster analysis uses a number of different algorithms and methods for grouping similar objects into their respective categories. Grouping methods optimize a criterion that measures the compatibility of clustering parameters with the data (Celeux & Govaert, 1992). This paper uses modified k-means and hierarchical clustering methods to group objects into classes with a purposively selected optimization criterion that reflects the conditional independence assumption. Treating class membership assigned through the clustering algorithm as the observed value of latent class makes it easy to estimate parameters in the LC/LP model. The proposed method can easily handle high-dimensional data (i.e., many indicators) and directly obtain estimated latent classes that best describe the association among indicators.

This paper also develops a classification rule for predicting the statuses of interest (e.g., diseased/not diseased, cured/not cured) of new objects based on LC/LP models. The proposed classification rule is especially useful for high-dimensional data (e.g., microarray data). Classification using high-dimensional data is difficult due to many noisy and overlapping features in such data, which can disturb the classification. Thus, it is important to carry out dimension reduction before classifying objects on the basis of high-dimensional data. The latent variable model explicitly recognizes and, hence, mitigates errors in measurement, and accurately summarizes measured indicators. Therefore, the latent variable model is the perfect tool for dimension reduction. The proposed parameter estimating procedure can easily perform LC/LP analyses on high-dimensional data, creating classification rules based on the inferred latent classes.

2. Model

Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iM})^T$ denote a set of M observable indicators and let S_i denote the unobservable class membership for the ith individual in a study sample of N persons. Y_{im} can be either continuous, ordinal, or categorical, and S_i can take values $\{1, \ldots, J\}$, where J is the number of latent classes. LC/LP models are based on the concept of conditional independence in the sense that the observed indicators are statistically independent within any latent class. Therefore, the density function for the observation $(y_1, \ldots, y_M)$ of $\mathbf{Y}_i$ can be expressed as the finite mixture:

$$f(y_1, \ldots, y_M) = \sum_{j=1}^{J} \left\{ \eta_j \prod_{m=1}^{M} f_{mj}(y_m | S_i = j) \right\}, \tag{1}$$

where

$$\Pr(S_i = j) = \eta_j, \quad \text{and} \quad Y_{im} | S_i = j \sim f_{mj}(\cdot | S_i = j). \tag{2}$$

Several authors have extended LC/LP models to describe the effects of covariates on the underlying outcome or on measured indicators within latent levels. It is possible to summarize the effect of covariates on the underlying outcome by allowing covariates $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{iP})^T$ to be functionally related to latent class S_i (Dayton & Macready, 1998; Bandeen-Roche et al., 1997;

Huang & Bandeen-Roche, 2004). This paper uses a generalized linear framework (McCullagh & Nelder, 1989) to incorporate covariate effects into S_i:

$$\log\left[\frac{\eta_j(\mathbf{x}_i)}{\eta_J(\mathbf{x}_i)}\right] = \beta_{0j} + \beta_{1j}x_{i1} + \cdots + \beta_{Pj}x_{iP}, \quad \text{for } j = 1, \ldots, J-1. \tag{3}$$

To adjust for characteristics associated with indicators and, hence, avoid possible misclassification of underlying variable categories, we can incorporate covariates in the within-class distributions of measured indicators (Melton, Liang, & Pulver, 1994; Huang & Bandeen-Roche, 2004; Muthén & Muthén, 2007). Let $\mathbf{z}_i = (\mathbf{z}_{i1}, \ldots, \mathbf{z}_{iM})$ with $\mathbf{z}_{im} = (z_{im1}, \ldots, z_{imL})^T$, $m = 1, \ldots, M$ be the covariates used to build direct effects on measured indicators within latent classes for the ith individual (i.e., $f_{mj}(\cdot|S_i = j) = f_{mj}(\cdot|S_i = j, \mathbf{z}_{im})$). If Y_{im} can take values $\{1, \ldots, K_m\}$, where $K_m \geq 2$, $m = 1, \ldots, M$, then assume that $(Y_{im}|S_i = j, \mathbf{z}_{im}) \sim$ Multinomial$(1; p_{m1j}(\mathbf{z}_{im}), p_{m2j}(\mathbf{z}_{im}), \ldots, p_{mK_mj}(\mathbf{z}_{im}))$, and

$$\log\left[\frac{p_{mkj}(\mathbf{z}_{im})}{p_{mK_mj}(\mathbf{z}_{im})}\right] = \gamma_{mkj} + \alpha_{1mk}z_{im1} + \cdots + \alpha_{Lmk}z_{imL}, \quad \text{for } k = 1, \ldots, (K_m - 1). \tag{4}$$

When indicators are continuous variables, we assume that $(Y_{im}|S_i = j, \mathbf{z}_{im}) \sim$ Normal$(\mu_{mj}(\mathbf{z}_{im}), \sigma_m^2)$, and

$$\mu_{mj}(\mathbf{z}_{im}) = \theta_{mj} + \tau_{1m}z_{im1} + \cdots + \tau_{Lm}z_{imL}. \tag{5}$$

Some features of the above extended LC/LP models that incorporate covariate effects can be found in Huang and Bandeen-Roche (2004). Notice that, after adjusting covariate effects, the conditional independence assumption is also conditioning on $\mathbf{z}_i$:

$$f_j(y_1, \ldots, y_M|S_i = j, \mathbf{z}_i) = \prod_{m=1}^{M} f_{mj}(y_m|S_i = j, \mathbf{z}_{im}). \tag{6}$$

3. Two-Stage Optimization-Based Parameter Estimation

This paper proposes an alternative strategy for estimating parameters. The proposed method consists of two stages. The first stage obtains the underlying latent class membership through some optimization procedure. The second stage treats the obtained class membership as a known variable and then estimates the other parameters.

LC/LP analysis is a useful tool for classifying objects based on their responses to a set of indicators. The basic model postulates an underlying categorical latent variable $S_i \in \{1, \ldots, J\}$ and assumes that the measured indicators within any category of the latent variable are independent of one another. The proposed method obtains S_i by grouping objects into J subgroups such that objects in one subgroup have a set of statistically independent indicators. Grouping methods optimize a criterion that measures the independence among indicators within each subgroup.

Let C_i denote the class membership assigned through the above optimization procedure. Then estimate β_{pj} in (3) by treating C_i as a response variable and fitting a polytomous logistic regression (McCullagh & Nelder, 1989) with covariates $\mathbf{x}_i$. Estimate $(\gamma_{mkj}, \alpha_{lmk})$ in (4) and $(\theta_{mj}, \tau_{lm}, \sigma_m^2)$ in (5) by plugging $C_{ij} = \mathrm{I}(C_i = j) = 1$ if $C_i = j$, 0 otherwise, $j = 1, \ldots, (J-1)$

into reparameterized models:

$$\begin{aligned}\log&\left[\frac{\Pr(Y_{im}=k|C_{i1},\ldots,C_{i(J-1)},\mathbf{z}_{im})}{\Pr(Y_{im}=K|C_{i1},\ldots,C_{i(J-1)},\mathbf{z}_{im})}\right]\\&=\gamma_{mk0}+\gamma_{mk1}C_{i1}+\cdots+\gamma_{mk(J-1)}C_{i(J-1)}\\&\quad+\alpha_{1mk}z_{im1}+\cdots+\alpha_{Lmk}z_{imL},\quad\text{for } k=1,\ldots,(K_m-1)\end{aligned}\tag{7}$$

and

$$\begin{aligned}\mathrm{E}(Y_{im}|C_{i1},\ldots,C_{i(J-1)},\mathbf{z}_{im})&=\theta_{m0}+\theta_{m1}C_{i1}+\cdots+\theta_{m(J-1)}C_{i(J-1)}\\&\quad+\tau_{1m}z_{im1}+\cdots+\tau_{Lm}z_{imL},\end{aligned}\tag{8}$$

and fitting polytomous logistic regressions and linear regressions, respectively. The corresponding standard error estimations can also be obtained. Notice that these standard error estimations do not reflect the "true" variations of parameter estimates. When performing the proposed two-stage procedure, β_{pj}'s and $(\gamma_{mkj},\alpha_{lmk})$'s/$(\theta_{mj},\tau_{lm},\sigma_m^2)$'s are estimated separately; therefore, the variation of estimates of β_{pj}'s does not account for the variation of $(\gamma_{mkj},\alpha_{lmk})/(\theta_{mj},\tau_{lm},\sigma_m^2)$ estimates, and vice versa. The following simulation study evaluates the closeness to the true values.

The following analysis theoretically justifies the proposed two-stage optimization-based model fitting. The Appendix provides the proof.

Theorem 1. *Treat the observable data as a sequence* $\{(\mathbf{Y}_1,\mathbf{x}_1,\mathbf{z}_1),(\mathbf{Y}_2,\mathbf{x}_2,\mathbf{z}_2),\ldots\}$ *of mutually independent vector triples, and let subscript-n denote quantities estimated from the first n triples in the sequence. Suppose that*

(a) *there exists object partition* C_{in} *that satisfies the conditional independence assumption* (6): $f_j(\mathbf{y}|C_{in}=j,\mathbf{z}_i)=\prod_{m=1}^{M}f_{mj}(y_m|C_{in}=j,\mathbf{z}_{im})$ *for* $j=1,\ldots,J$; *and*

(b) $\widehat{\beta}_{pjn}$, $(\widehat{\gamma}_{mkjn},\widehat{\alpha}_{lmkn})$ *and* $(\widehat{\theta}_{mjn},\widehat{\tau}_{lmn},\widehat{\sigma}^2_{mn})$, *for all* j,m,p,l,k, *are the estimates of parameters in* (3), (7) *and* (8), *with* C_{in} *being the class membership indicator and these parameter estimates converge in probability to* β^*_{pj}, $(\gamma^*_{mkj},\alpha^*_{lmk})$ *and* $(\theta^*_{mj},\tau^*_{lm},\sigma^{*2}_m)$.

Then, the underlying distribution of $\mathbf{Y}_i$

$$f_{\mathbf{Y}_i}(\mathbf{y}|\mathbf{x}_i,\mathbf{z}_i)\xrightarrow[n\to\infty]{}\sum_{j=1}^{J}\left\{\eta^*_j(\mathbf{x}_i)\prod_{m=1}^{M}f^*_{mj}(y_m|S_i=j,\mathbf{z}_{im})\right\}$$

for each $\mathbf{y}=(y_1,\ldots,y_M)\in Supp_{\mathbf{Y}_i}$ *and* $i=1,2,\ldots$, *where* $Supp_{\mathbf{Y}_i}$ *denotes the support of* $\mathbf{Y}_i$, *and* $\eta^*_j(\mathbf{x}_i)$ *and* $f^*_{mj}(y_m|S_i=j,\mathbf{z}_{im})$ *are class probabilities and within-class distributions in* (3), (4) *and* (5) *evaluated at* β^*_{pj}, $(\gamma^*_{mkj},\alpha^*_{lmk})$ *and* $(\theta^*_{mj},\tau^*_{lm},\sigma^{*2}_m)$.

Assumption (a) is met if the optimization procedure obtains an object partition with a small enough loss value. Part (b) assumes the consistency of parameter estimators when the underlying latent variable is known and equal to the class membership estimates from (a). The polytomous logistic and linear regressions are used to estimate these parameters, which leads to maximum likelihood estimates. Thus, the convergence is ensured under some regularity conditions. Theorem 1 conveys many key features of the proposed approach. First, if the optimization procedure can obtain satisfactory results, the estimates obtained from the proposed approach can be viewed as a derivation of a LC/LP model (1). Second, its statement as an asymptotic result highlights the necessity of adequate precision for estimation, which is driven by convergence of

$(\widehat{\beta}_{pjn}, \widehat{\gamma}_{mkjn}, \widehat{\alpha}_{lmkn}, \widehat{\theta}_{mjn}, \widehat{\tau}_{lmn}, \widehat{\sigma}^2_{mn})$ to $(\beta^*_{pj}, \gamma^*_{mkj}, \alpha^*_{lmk}, \theta^*_{mj}, \tau^*_{lm}, \sigma^{*2}_m)$. Third, in the proposed optimization-based approach, object partition is treated as an unknown parameter, and the maximization is over all possible partitions as well as over values of model parameters. Under this approach, the model parameters and object partitions increase in number with the number of observations. Note that $(\beta^*_{pj}, \gamma^*_{mkj}, \alpha^*_{lmk}, \theta^*_{mj}, \tau^*_{lm}, \sigma^{*2}_m)$ may not be the true population parameters (Marriott, 1975; Bryant & Williamson, 1978; Clogg, 1995). Nevertheless, this approach appears to be fruitful since it allows us to find out the underlying unobservable characteristics. As a result, the conditional independence assumption that cannot be directly checked in the likelihood approach can be evaluated as a part of our two-stage algorithm. This two-stage approach splits the traditional estimation algorithm into two subsets and operates two subsets individually. Because each subset has fewer parameters and is easier to converge, this algorithm should be faster than the full likelihood approach.

4. Optimization Algorithm

To obtain the underlying latent class S_i, the proposed method modifies k-means and hierarchical clustering algorithms to optimize the conditional independence criterion. This section first details the proposed algorithms when covariates $\mathbf{z}_i$ are not incorporated in the conditional distribution $f_{mj}(\cdot|S_i = j)$, and then extends the algorithms to allow covariate effects. Finally, the proposed method selects the number of latent classes based on the results of these optimization algorithms.

4.1. Latent Class Membership Assignment When Not Incorporating Covariate Effects

To illustrate the proposed approach, this section first describes the sample covariance matrix among indicators used. For continuous indicators, the sample covariance matrix is an $M \times M$ matrix with component (m, t) being the sample covariance between Y_{im} and Y_{it}. For polytomous categorical indicators, each component of $(Y_{i1}, \ldots, Y_{iM})$ is represented as a vector with elements being the indicators of each category:

$$\begin{aligned}\widetilde{\mathbf{Y}}_i &= \big(\widetilde{\mathbf{Y}}_{i1}, \widetilde{\mathbf{Y}}_{i2}, \ldots, \widetilde{\mathbf{Y}}_{iM}\big)\\ &= (Y_{i11}, \ldots, Y_{i1(K_1-1)}, Y_{i21}, \ldots, Y_{i2(K_2-1)}, \ldots, Y_{iM1}, \ldots, Y_{iM(K_M-1)})\end{aligned}$$

with $Y_{imk} = \mathrm{I}(Y_{im} = k)$; $m = 1, \ldots, M$; $k = 1, \ldots, (K_m - 1)$. Then,

$$\mathrm{Cov}\big(\widetilde{\mathbf{Y}}_i\big) = \big\{\mathrm{Cov}(Y_{imk}, Y_{its})\big\} = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} & \cdots & \mathbf{B}_{1M} \\ \mathbf{B}_{21} & \mathbf{B}_{22} & \cdots & \mathbf{B}_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{B}_{M1} & \mathbf{B}_{M2} & \cdots & \mathbf{B}_{MM} \end{bmatrix}, \tag{9}$$

where $\mathbf{B}_{mt} = \mathrm{Cov}(\widetilde{\mathbf{Y}}_{im}, \widetilde{\mathbf{Y}}_{it})$ is a $(K_m - 1) \times (K_t - 1)$ block matrix. Various components of the above covariance matrix are

$$\mathrm{Cov}(Y_{imk}, Y_{its}) = \begin{cases} \Pr(Y_{imk} = 1) - \Pr(Y_{imk} = 1)\Pr(Y_{its} = 1) & \text{if } m = t \text{ and } k = s, \\ -\Pr(Y_{imk} = 1)\Pr(Y_{its} = 1) & \text{if } m = t \text{ and } k \neq s, \\ \Pr(Y_{imk} = 1, Y_{its} = 1) - \Pr(Y_{imk} = 1)\Pr(Y_{its} = 1) & \text{if } m \neq t. \end{cases} \tag{10}$$

Obtain the sample covariance matrix by replacing the probabilities with the sample averages. Second, define the "loss of independence," as the distance measure used when performing k-means and hierarchical clustering. Let ACov_j be the average of absolute values of entries in

off-diagonal elements (continuous indicators)/blocks (polytomous indicators) of the sample covariance matrix using objects in class j. ACov_j represents the magnitude of between-indicator covariances for the jth class. Thus, "loss of independence" is defined as

$$\mathrm{LoI} = \sum_{j=1}^{J} w_j \, \mathrm{ACov}_j \quad \text{with } w_j = \frac{\text{the number of objects in class } j}{N}. \tag{11}$$

Notice that LoI is the weighted average of ACov_j over all classes with weights proportional to the number of objects in each class. The weighted average can account for the sizes of classes to avoid inflating the effect of small classes. However, this also increases the risk of conditional dependency in small classes. In the planning stage of our approach, we tried the equal weighting LoI, but the results were not as good as the weighted one.

This paper modifies the k-means algorithm, which groups objects using the criterion of distance to cluster mean. The modified algorithm is termed as the "k-groups" clustering algorithm, which can more appropriately reflect the main idea of the algorithm. The k-groups algorithm uses the following steps to obtain the estimated class membership:

K1. Randomly partition the objects into J initial classes.
K2. Proceed through the list of objects, assigning an object to the class with the minimum "loss of independence."
K3. Repeat Step K2 until no more reassignments take place.

Step K2, for a given object, defines $\mathrm{LoI}^{(u)} = \sum_{j=1}^{J} w_j^{(u)} \, \mathrm{ACov}_j^{(u)}$ as the loss of independence when the object is assigned to class u. Assigning the object through all J classes obtains $\mathrm{LoI}^{(1)}, \ldots, \mathrm{LoI}^{(J)}$. A smaller value of $\mathrm{LoI}^{(u)}$ increases the independence of the observed indicators within latent classes when assigning the object to class u. Thus, assign a given object to the class with the minimum loss of independence. Figure 1 shows an example of this k-groups procedure.

Hierarchical clustering techniques proceed through either a series of successive mergers (agglomerative) or a series of successive divisions (divisive). Agglomerative hierarchical methods are precise at the bottom of the clustering tree and adequate when looking for small or many classes. Divisive hierarchical methods are precise at the top of the tree and adequate when looking for large or few classes. The agglomerative hierarchical clustering algorithm includes the following steps:

AH1. Start with N classes, each containing a single object.
AH2. Consider the union of every possible pairs of classes. Merge the two classes whose combination results in the minimum loss of independence.
AH3. Repeat Step AH2 until all objects are in a single class. Record the identity of classes merged and the losses of independence at which the mergers take place.

In Step AH2, for current $J(\leq N)$ classes, if classes u and v are merged, label the newly formed class (uv). Let $\mathrm{LoI}^{(uv)}$ be the loss of independence when merging classes u and v (i.e., $\mathrm{LoI}^{(uv)} = w_{(uv)} \, \mathrm{ACov}_{(uv)} + \sum_{j \neq u,v} w_j \, \mathrm{ACov}_j$). We can get $\binom{J}{2}$ $\mathrm{LoI}^{(uv)}$'s. A smaller value of $\mathrm{LoI}^{(uv)}$ increases the independence of the observed indicators within latent classes when merging classes u and v. Thus, merge the two classes whose combination results in the minimum loss of independence. Figure 2 illustrates the proposed agglomerative hierarchical procedure. The results of the agglomerative hierarchical clustering method can be graphically displayed as a dendrogram whose vertical axis gives the values of the loss of independence at which the mergers occur.

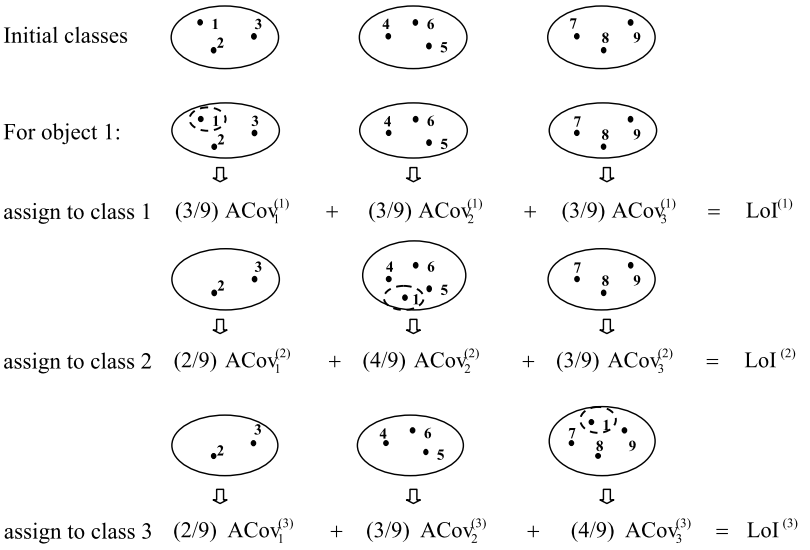

FIGURE 1.
The proposed k-groups procedure. K1: Partition the 9 objects into 3 initial classes. K2: What class will object 1 be assigned to? Assigning object 1 to classes 1, 2, and 3 yields LoI$^{(1)}$, LoI$^{(2)}$, and LoI$^{(3)}$, respectively. Assign object 1 to the class corresponding to the minimum LoI$^{(j)}$. Proceed through objects 2–9, repeating the above procedure. K3: Repeat Step K2 until no more reassignments take place.

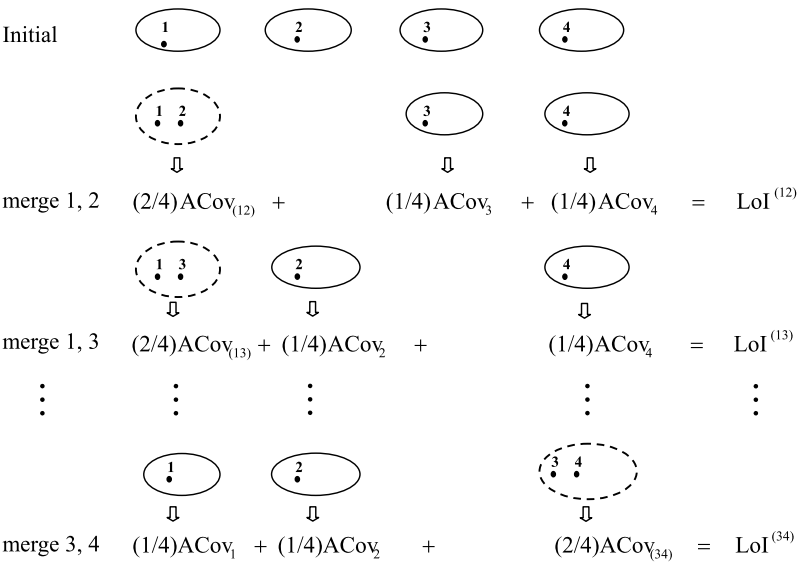

FIGURE 2.
The proposed agglomerative hierarchical procedure. AH1: Star with four initial classes, each containing a single object. AH2: Which pair of classes will be merged? Considering the union of all six ($=\binom{4}{2}$) possible pairs of classes, we get LoI$^{(12)}$, LoI$^{(13)}$, LoI$^{(14)}$, LoI$^{(23)}$, LoI$^{(24)}$, and LoI$^{(34)}$. Merge the pair of classes whose combination yields the minimum LoI$^{(uv)}$. AH3: Repeat Step AH2 until all objects are in a single class.

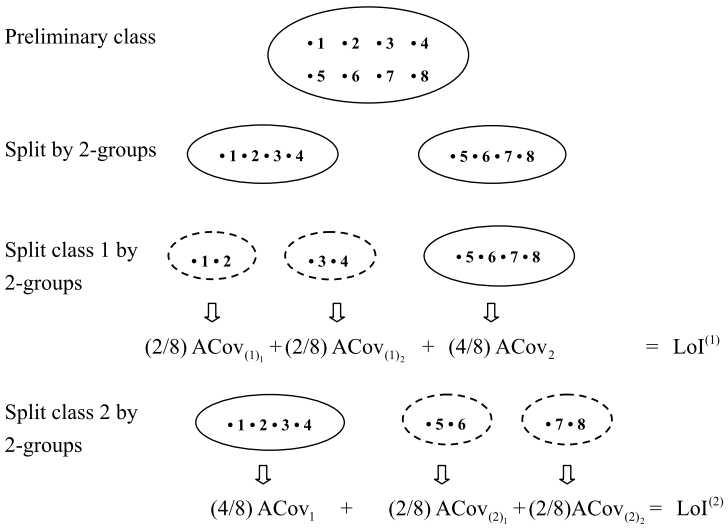

FIGURE 3.
The proposed divisive hierarchical procedure. DH1: Start with a single class that consists of all objects. DH2: Use 2-groups approach to divide the initial class into classes 1 and 2. DH3: Which class will be split first? Splitting class 1 produces $\text{LoI}^{(1)}$. Splitting class 2 produces $\text{LoI}^{(2)}$. Split the class whose division yields the minimum $\text{LoI}^{(u)}$. DH4: Repeat Step DH3 until each object is in its own singleton class.

The divisive hierarchical algorithm is implemented as follows:

DH1. Start with a single class containing all objects.
DH2. Divide the preliminary class into two smaller classes, using the k-groups approach above with $k = 2$.
DH3. Split one of the existing classes, as in Step DH2. Split the class whose division yields the minimum loss of independence.
DH4. Repeat Step DH3 until each object is in its own singleton class. Record the identity of classes split and the losses of independence at which the splits take place.

In Step DH3, for current $J\,(\geq 1)$ classes, if the class u is split by the 2-groups approach, label the two newly formed classes $(u)_1$ and $(u)_2$. Let $\text{LoI}^{(u)}$ be the loss of independence when the class u is split (i.e., $\text{LoI}^{(u)} = w_{(u)_1}\,\text{ACov}_{(u)_1} + w_{(u)_2}\text{ACov}_{(u)_2} + \sum_{j\neq u} w_j\,\text{ACov}_j$). Of those J $\text{LoI}^{(u)}$'s, split the class whose division yields the minimum loss of independence. Figure 3 illustrates the proposed divisive hierarchical procedure. Similar to the agglomerative procedure, the results of the divisive hierarchical clustering method can be graphically displayed as a dendrogram whose vertical axis represents the values of the loss of independence at which the splits occur. Notice that the dendrogram can contain inversions or reversals in which a larger loss of independence takes place at a larger number of classes. To make the plot more interpretable, when inversion happens, set the height of J classes in the dendrogram as the height of $J - 1$ classes.

4.2. Latent Class Membership Assignment When Incorporating Covariate Effects

The k-groups and hierarchical clustering algorithms above assume that observed indicators are statistically independent within any latent class. If covariates $\mathbf{z}_{im}$ are incorporated into the conditional distributions of models (4) and (5), the conditional independence assumption is also conditioning on incorporated covariates (i.e., assumption (6)). To apply these algorithms to models (4) and (5), it is necessary to "eliminate" the covariate effects, and hence "marginalize" models (4) and (5).

This study adopts the marginalization process developed in Section 3.3.1 of Huang (2005). The strategy for achieving such marginalization can be motivated by the properties of added variable plots for linear regression models (Cook & Weisberg, 1982). To present the process, first reparameterize models (4) and (5) as models (7) and (8), respectively. This process assumes that the incorporated covariates $\mathbf{z}_{im}$ and the class membership C_{ij}, $j = 1, \ldots, J-1$ are orthogonal, and calculates the residual of regressing Y_{im} on $\mathbf{z}_{im}$ separately for each $m \in \{1, \ldots, M\}$. It is then possible to extract $\mathbf{z}_{im}$ from conditional distributions by treating these residuals as new response variables and regressing them on C_{ij}. Therefore, the conditional independence assumption (6) is satisfied if objects belonging to the same latent class have a set of M statistically independent residuals. Thus, apply the clustering algorithms developed in the previous section to these residuals to obtain a class membership that satisfies assumption (6).

When Y_{im}'s are continuous, compute the typical residuals of linear regressions R_{im} (i.e., the differences between observed responses and their modeled predictors). When Y_{im}'s are categorical, the problem becomes how to calculate residuals from the generalized linear model

$$\log\left[\frac{\Pr(Y_{im}=k|\mathbf{z}_{im})}{\Pr(Y_{im}=K|\mathbf{z}_{im})}\right] = \alpha_{1mk}z_{im1} + \cdots + \alpha_{Lmk}z_{imL}, \quad \text{for } k = 1, \ldots, (K_m - 1). \tag{12}$$

We propose to use the "pseudo-residual"

$$\widetilde{\mathbf{R}}_{im} = \big(\widehat{\mathrm{Cov}}\big(\widetilde{\mathbf{Y}}_{im}\big)\big)^{-1}\big(\widetilde{\mathbf{Y}}_{im} - \widehat{\mathbf{p}}_{im}\big), \tag{13}$$

where $\widetilde{\mathbf{Y}}_{im}$ is defined as in Section 4.1, $\mathbf{p}_{im} = \mathrm{E}(\widetilde{\mathbf{Y}}_{im}|\mathbf{z}_{im})$, and "hat" denotes the estimated values based on (12). The pseudo-residual (13) is defined by analogizing the iteratively reweighted least-squares of generalized linear models with the least-square estimates of linear regressions (Landwehr, Pregibon, & Shoemaker, 1984; Huang, 2005). Then, classify objects based on new response variables R_{im} (continuous indicators) or $\widetilde{\mathbf{R}}_{im}$ (categorical indicators), as in the previous subsection.

The orthogonality assumption between $\mathbf{z}_{im}$ and the class membership is a strong assumption in most applications. However, the orthogonality assumption holds to an increasingly close approximation as $N \to \infty$ if $\mathbf{x}_i$ and $\mathbf{z}_{im}$ are independent (Huang, 2005). This assumption can be verified empirically by calculating the sample correlation matrix among covariates.

4.3. Selecting the Number of Latent Classes

When using the k-groups algorithm to infer the underlying class membership, first fit the LC/LP models (1), (3), (4) and/or (5) repeatedly under different numbers of classes, and record their corresponding loss of independence at which the k-groups algorithm stops. The number of latent classes to be selected is the number that yields the minimum loss of independence. For the agglomerative/divisive hierarchical clustering algorithm, examine the dendrogram for large changes in vertical values, which indicate the best number of latent classes to fit.

5. Classification Using LC/LP Models

Many studies attempt to predict new observations' unknown disease statuses based on indicator measurements. Here, the LC/LP model is used to predict the disease status from the indicators, on the assumption that the latent class mediates the relationship fully. So from a psychometric point of view, the LC or LP model is measurement invariant with respect to the disease status (Meredith, 1993).

Consider a set of N objects with known disease statuses D_i and measured indicators $\mathbf{Y}_i$ plus incorporated covariates $\mathbf{x}_i, \mathbf{z}_i$ if existing, for $i = 1, \ldots, N$, where D_i takes values $\{1, \ldots, A\}$. Use these objects to fit LC/LP models (1), (3), (4) and/or (5) following the methods described in Section 3. Then, obtain estimations C_i, $\widehat{\beta}_{pj}$, $(\widehat{\gamma}_{mkj}, \widehat{\alpha}_{lmk})$ and $(\widehat{\theta}_{mj}, \widehat{\tau}_{lm}, \widehat{\sigma}^2_m)$, for all i, j, m, p, l, k. The posterior possibility of classifying a new object with measurements on indicators $\mathbf{Y}^* = (Y_1^*, \ldots, Y_M^*)$ and covariates $\mathbf{x}^*, \mathbf{z}^* = (\mathbf{z}_1^*, \ldots, \mathbf{z}_M^*)$ as the disease status $D^* = a$ is

$$\Pr\big(D^* = a \big| \mathbf{Y}^*, \mathbf{x}^*, \mathbf{z}^*\big) = \sum_{j=1}^{J} \big\{\Pr\big(D^* = a \big| S^* = j\big) \Pr\big(S^* = j \big| \mathbf{Y}^*, \mathbf{x}^*, \mathbf{z}^*\big)\big\}, \tag{14}$$

where S^* is the presumed latent class membership of the new object. Equation (14) is true when assuming

$$\Pr\big(D^* = a \big| S^* = j, \mathbf{Y}^*, \mathbf{x}^*, \mathbf{z}^*\big) = \Pr\big(D^* = a \big| S^* = j\big); \tag{15}$$

in other words, latent classes can fully capture the association between the disease status and observed indicators. The ability of the psychometric model to relate indicators to a latent variable does not generally imply that the relationship between any external variable and the indicators should run via the latent variable (which would ensure (15)). See Lux and Kendler (2010) for further discussion. Fortunately, it is possible to verify (15) empirically by performing the regression relating D_i to C_i, $\mathbf{Y}_i$, $\mathbf{x}_i$ and $\mathbf{z}_i$.

The components in the right hand side of (14) can be estimated by

$$\widehat{\Pr}\big(D^* = a \big| S^* = j\big) = \frac{\sum_{i=1}^{N} \mathrm{I}(C_i = j)\mathrm{I}(D_i = a)}{\sum_{i=1}^{N} \mathrm{I}(C_i = j)} \tag{16}$$

and

$$\widehat{\Pr}\big(S^* = j \big| \mathbf{Y}^* = \big(y_1^*, \ldots, y_M^*\big), \mathbf{x}^*, \mathbf{z}^*\big) = \frac{\widehat{\eta}_j(\mathbf{x}^*) \prod_{m=1}^{M} \widehat{f}_{mj}(y_m^* | S_i = j, \mathbf{z}_m^*)}{\sum_{t=1}^{J} \{\widehat{\eta}_t(\mathbf{x}^*) \prod_{m=1}^{M} \widehat{f}_{mt}(y_m^* | S_i = t, \mathbf{z}_m^*)\}}. \tag{17}$$

Here, $\widehat{\eta}_j(\mathbf{x}^*)$ is the estimated latent prevalence of the jth class for the new observation $\mathbf{x}^*$, evaluated at estimator $\widehat{\beta}_{pj}$. The term $\widehat{f}_{mj}(y_m^* | S_i = j, \mathbf{z}_m^*)$ is the estimated conditional distribution of the mth indicator given the jth class for the new observation $(y_m^*, \mathbf{z}_m^*)$, evaluated at estimators $(\widehat{\gamma}_{mkj}, \widehat{\alpha}_{lmk})$ or $(\widehat{\theta}_{mj}, \widehat{\tau}_{lm}, \widehat{\sigma}^2_m)$. Allocate $\mathbf{Y}^*$ to $D^* = a^*$ at which the maximum estimated posterior probability is reached, i.e.,

$$a^* = \arg \max_{a \in \{1, \ldots, A\}} \widehat{\Pr}\big(D^* = a \big| \mathbf{Y}^*, \mathbf{x}^*, \mathbf{z}^*\big). \tag{18}$$

The proposed classification rule predicts a new object's disease status using his/her inferred latent class variable S^*. The S^* term essentially summarizes the new object's measured indicators through the training set $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_N\}$. This summarization process can reduce dimensionality and errors in measurements. This is especially important for high-dimensional data due to many overlapping and noisy features, which can disturb the classification. Also, if the disease of interest does not follow typical Mendelian inheritance patterns (i.e., is a "complex" disease), objects with the same disease status can originate from different indicator response patterns. As a result, traditional linear/quadratic discriminant models that directly summarize the indicator response patterns under each disease status can fail. The proposed LC/LP approach recognizes the heterogeneous nature of a complex disease; it first predicts the new object's likelihood of being in homogeneous indicator response patterns and then weights each homogeneous pattern

with its prediction probability (i.e., $\Pr(D^* = a|S^* = j)$) to classify the new object. Genetic and nongenetic (environmental) contributions to the disease can be adjusted through incorporated covariates $\mathbf{x}^*$ and/or $\mathbf{z}^*$ (Lubke, Carey, Lessem, & Hewitt, 2008).

6. Simulation Study

The simulations in this study examine characteristics of the proposed approaches in three aspects: (i) the performance of the proposed approach, including the accuracy of model parameter estimations, agreement between simulated class membership and assigned (by proposed clustering algorithms) class membership, and satisfaction of independence among indicators within each latent class; (ii) the sensitivity of the proposed method to model assumptions, including conditional independence among indicators given the latent class, and orthogonality between $\mathbf{z}_{im}$ and the class membership; and (iii) a comparison of the current results with the traditional EM approach used in Huang and Bandeen-Roche (2004) (regression extension of latent class analysis (RLCA) model). This section only presents results for categorical indicator measurements (i.e., Y_{im} takes values $\{1, \ldots, K_m\}$, for $m = 1, \ldots, M$). A similar pattern of characteristics appeared in continuous indicators.

6.1. Study Design

This study simulates two different latent class models. The first model was a three-class model 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 model with five three-level measured indicators, two covariates associated with conditional probabilities, and two covariates associated with latent prevalences ($J = 6$, $M = 5$, $K_1 = \cdots = K_5 = 3$, $P = L = 2$). The covariates used in the simulation were obtained from the schizophrenia syndrome scale study described in the following section. In each model, the covariates associated with conditional probabilities were the variables "Sex" and "Age" (in years), and the covariates associated with latent prevalences were variables "Occupation" (with versus without occupation) and "Dprime," which is the sensitivity index from the Continuous Performance Test (Rosvold, Mirsk, Sarason, Bransome, & Bech, 1956). The true model parameter values of β_{pj}, γ_{mkj} and α_{lmk} were adopted from the parameter estimates of fitting the RLCA model to a subset of the data used in the following section. For the three-class model, the sample size $N = 200$ with approximately six individuals per parameter. For the six-class model, $N = 500$ with approximately five individuals per parameter. The indicator measurements $\mathbf{Y}_i$ were then generated from the model with 100 replications.

This study also simulates latent class models with "conditional dependency" among measured indicators given the latent class membership. This represents a model often encountered in practice, and makes it possible to evaluate the impact of the conditional independency assumption on various model fitting approaches. The conditional probabilities were obtained through

$$\log\left[\frac{p_{mkj}(\mathbf{z}_{im})}{p_{mK_mj}(\mathbf{z}_{im})}\right] = \gamma_{mkj} + b_{ij} + \alpha_{1mk}z_{im1} + \cdots + \alpha_{Lmk}z_{imL}, \tag{19}$$

where $b_{ij} \sim \text{Normal}(0, \nu_j^2)$, and β_{pj}, γ_{mkj} and α_{lmk} equal the values used in the conditional independence model. Model (19) allows for conditional dependence among indicators by incorporating a single Gaussian random effect into (4) (Qu, Tan, & Kunter, 1996; Albert, McShane, & Shih, 2001).

The covariates associated with conditional probabilities ("Sex" and "Age") and the covariates associated with latent prevalences ("Occupation" and "Dprime") are not independent;

older age is associated with a lower "Dprime" value (p-value = 0.02) and having an occupation (p-value = 0.04). This indicates the violation of the orthogonality assumption between $\mathbf{z}_{im}$ and the class membership described in Section 4.2, but it is often encountered in real applications. The orthogonality assumption is not required to implement the EM approach. To evaluate the sensitivity of the proposed marginalization method to this assumption, compare the results of a latent class model with two independent sets of covariates: $z_{im1} \sim \text{Bernoulli}(p_z)$, $z_{im2} \sim \text{Normal}(\mu_z, \sigma_z^2)$; $x_{i1} \sim \text{Bernoulli}(p_x)$, $x_{i2} \sim \text{Normal}(\mu_x, \sigma_x^2)$; and all z_{iml} and x_{ip} are mutually independent. We chose $p_z = 0.54$ (the proportion of males in "Sex"), $\mu_z = 33.33$, $\sigma_z^2 = 64.06$ (the sample mean and variance of "Age"), $p_x = 0.28$ (the proportion of having occupation in "Occupation"), and $\mu_x = 3.05$, $\sigma_x^2 = 2.36$ (the sample mean and variance of "Dprime").

Initial values must be chosen carefully to ensure reasonable convergence of fitting algorithms and avoid local maxima in parameter estimation. To initialize EM fitting of (3) and (4), first fit a latent class model without the incorporated covariates, and then randomly assign each object to a class with posterior probabilities of class membership calculated from the initial no-covariate model. Then, regress the estimated membership on the covariates to obtain initial values of coefficients for (3) and (4). Given that the simulated data are somewhat sparse in this study, iterated EM steps quit frequently due to extremely large values of parameter estimates, and the algorithm did not converge. Thus, the absolute values of parameter estimates were trimmed to 20 in each EM iteration, which redirected the algorithm at boundary values and improved the convergence rate of the EM procedure. Initial classes of the proposed k-groups optimization algorithm were obtained through clustering algorithms based on k-means or hierarchical clustering, but using the criterion of distance to cluster mean instead of the proposed criterion of conditional independence.

6.2. Simulation Results

6.2.1. Performance of the Two-Stage and EM Approaches The "true" conditional probabilities and the latent prevalences calculated from the true parameter values were examined (results are not shown). Apparently, the simulated data in this study contain sparse response patterns, and the six-class model exhibits a different sparseness pattern than the three-class model. Thus, this simulation can evaluate the proposed approaches in various sparse-data situations.

Table 1 presents estimations of the regression coefficients for conditional probabilities under the three-class model. The EM approach performed poorly in estimating γ_{mkj}'s. The estimations from the k-groups, agglomerative hierarchical, and divisive hierarchical methods were close to the true values. Table 1 also displays the standard errors of parameter estimates in polytomous logistic regressions (7) and the sample standard errors of the parameters estimates from 100 simulation replications (i.e., "true" standard errors). Both standard errors were much larger in the EM approach than the clustering approaches. The EM approach also yielded very large standard errors of γ_{mkj} estimates. This is due to the convergence to boundary solutions in γ_{mkj} estimates for some replicates, which may have been caused by sparse response patterns. As expected, the proposed clustering methods underestimated the standard error estimates from polytomous logistic regressions. This underestimation was small, especially for the covariate coefficients ("Sex" and "Age").

Table 2 presents estimates for latent prevalence regression (3) under the three-class model. Similar to the results for conditional probabilities, the proposed clustering methods achieved more accurate parameter estimation than the EM approach. The EM approach did not create the extremely large standard error estimates that it did in conditional probabilities, indicating more stable parameter estimation in the latent prevalence regression.

Table 3 reveals agreement between simulated and obtained/inferred class membership under the three-class model. The simulated ("true") class membership for individual i was obtained

TABLE 1.
Estimations of the regression coefficients in the conditional probability regression (7) under the three-class model.

	True	EM algo.	K-groups	Hier. aggl.	Hier. divi.
			Indicator 1		
Intercept	−4.40	−0.66 (428.66/8.13/8.89)[a]	−3.77 (1.46/1.99/2.08)	−4.23 (2.18/2.54/2.53)	−3.19 (1.15/1.52/1.94)
Class 1	11.15	3.59 (619.57/12.88/14.84)	9.48 (16.69/4.19/4.48)	8.44 (10.96/4.18/4.96)	6.46 (5.66/3.13/5.62)
Class 2	3.09	0.71 (606.24/13.41/13.51)	4.09 (2.21/2.09/2.31)	5.84 (6.69/3.68/4.57)	4.22 (5.06/4.02/4.16)
Sex	0.39	0.68 (0.52/0.68/0.73)	0.55 (0.74/1.04/1.05)	0.56 (0.78/1.04/1.04)	0.54 (0.49/0.50/0.52)
Age	0.08	0.10 (0.04/0.07/0.07)	0.05 (0.03/0.04/0.04)	0.05 (0.04/0.04/0.05)	0.04 (0.03/0.03/0.05)
			Indicator 2		
Intercept	−1.43	−1.13 (63.52/5.42/5.38)	−1.02 (1.26/2.12/2.15)	−2.22 (3.44/3.20/3.28)	−1.44 (1.95/3.03/3.01)
Class 1	4.90	2.57 (98.82/7.00/7.33)	4.95(1.00/1.84/1.83)	6.16 (3.21/3.12/3.35)	5.16 (1.68/2.72/2.72)
Class 2	3.05	−0.32 (303.17/9.70/10.20)	3.02 (0.99/1.88/1.87)	3.14(4.59/4.59/4.57)	2.38 (3.49/3.79/3.82)
Sex	−0.64	−0.57 (0.39/0.46/0.46)	−0.45 (0.40/0.48/0.52)	−0.46 (0.43/0.56/0.59)	−0.45 (0.40/0.47/0.51)
Age	−0.05	−0.05 (0.02/0.03/0.03)	−0.07 (0.03/0.03/0.03)	−0.07 (0.03/0.03/0.03)	−0.07 (0.03/0.03/0.03)
			Indicator 3		
Intercept	−4.10	−4.00 (37.04/4.78/4.74)	−4.24 (1.67/2.55/2.54)	−4.98 (2.35/2.95/3.06)	−4.99 (2.72/3.35/3.45)
Class 1	4.33	2.28 (42.68/5.71/6.03)	5.02 (1.31/3.08/3.14)	5.80 (2.72/3.46/3.74)	5.33 (2.41/3.08/3.22)
Class 2	1.91	0.70 (70.83/7.13/7.17)	2.24 (2.11/2.82/2.83)	−0.05 (10.20/5.43/5.75)	1.89 (4.20/4.80/4.77)
Sex	−0.06	0.04 (0.39/0.48/0.49)	0.13 (0.41/0.48/0.52)	0.15 (0.44/0.60/0.63)	0.12 (0.38/0.45/0.48)
Age	0.03	0.04 (0.02/0.04/0.04)	0.01 (0.03/0.03/0.03)	0.02 (0.03/0.03/0.03)	0.01 (0.02/0.03/0.03)
			Indicator 4		
Intercept	−20.67	−9.39 (47.18/10.29/15.21)	−6.70 (3.30/3.39/14.38)	−7.90 (3.86/4.33/13.48)	−5.18 (2.45/2.84/15.75)
Class 1	13.04	3.16 (5.38/10.29/14.20)	6.75 (3.64/3.25/7.08)	7.19 (3.36/3.79/6.96)	5.82 (3.79/2.99/7.81)
Class 2	12.97	0.98 (44.45/13.15/17.71)	5.41 (3.51/3.06/8.15)	7.70 (9.36/4.61/6.99)	3.96 (3.79/3.65/9.71)
Sex	1.09	0.99 (14.62/4.91/4.87)	0.77 (0.67/0.94/0.99)	0.82 (0.74/1.03/1.06)	0.70 (0.49/0.59/0.71)
Age	0.36	0.11 (3.11/1.63/1.64)	0.10 (0.04/0.04/0.26)	0.11 (0.04/0.07/0.26)	0.07 (0.03/0.03/0.29)

TABLE 1.
(*Continued.*)

	True	EM algo.	K-groups	Hier. aggl.	Hier. divi.
			Indicator 5		
Intercept	−1.10	2.26 (7.11/6.48/7.26)	−0.68 (0.84/0.94/1.03)	−1.03 (1.13/1.41/1.40)	−0.62 (0.83/0.87/0.99)
Class 1	12.97	2.29 (14.08/11.89/15.91)	6.28 (8.28/3.70/7.64)	5.64 (4.83/3.50/8.12)	3.60 (1.91/2.19/9.62)
Class 2	0.69	0.08 (11.03/10.34/10.28)	0.95 (0.43/0.68/0.73)	0.27 (1.28/2.45/2.47)	0.93 (0.49/1.79/1.80)
Sex	−0.86	−0.80 (0.39/0.52/0.52)	−0.72 (0.39/0.47/0.49)	−0.68 (0.40/0.45/0.48)	−0.68 (0.38/0.42/0.46)
Age	0.02	0.02 (0.02/0.03/0.03)	0.003 (0.02/0.03/0.03)	0.01 (0.02/0.03/0.03)	0.002 (0.02/0.02/0.03)

[a]Numbers in parentheses: average of standard errors from the conditional probability regression/empirical standard error based on simulation replications/root mean square error.

TABLE 2.
Estimations of the regression coefficients in the latent prevalence regression (3) under the three-class model.

	True	EM algo.	K-groups	Hier. aggl.	Hier. divi.
			Class 1 vs. Class 3		
Intercept	−1.87	−0.68 (0.48/2.43/2.69)[b]	−1.21 (0.44/0.39/0.77)	−1.10 (0.44/0.56/0.95)	−0.76 (0.43/0.80/1.37)
Occupation	1.13	0.26 (0.39/0.92/1.26)	0.97 (0.39/0.39/0.42)	0.95 (0.40/0.37/0.41)	0.86 (0.41/0.41/0.49)
Dprime	0.53	0.17 (0.13/0.63/0.73)	0.35 (0.12/0.10/0.21)	0.35 (0.12/0.11/0.21)	0.28 (0.12/0.11/0.27)
			Class 2 vs. Class 3		
Intercept	−0.23	−0.03 (0.43/1.44/1.45)	−0.79 (0.42/0.49/0.74)	−0.70 (0.43/0.61/0.77)	−0.74 (0.45/0.75/0.90)
Occupation	0.33	−0.06 (0.42/1.03/1.09)	0.47 (0.45/0.40/0.42)	0.44 (0.49/0.52/0.53)	0.39 (0.49/0.49/0.49)
Dprime	0.01	−0.01 (0.12/0.43/0.42)	0.10 (0.12/0.11/0.14)	0.06 (0.13/0.13/0.14)	0.08 (0.13/0.18/0.19)

[b]Numbers in parentheses: average of standard errors from the conditional probability regression/empirical standard error based on simulation replications/root mean square error.

TABLE 3.
Matched number of individuals between simulated and obtained/inferred class membership under the three-class model with total 200 individuals, averaging over 100 replications.

Sim.\Est.	EM algo.			K-groups			Hier. aggl.			Hier. divi.		
	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3
Class 1	34.73	22.13	22.13	66.36	11.03	0.87	68.79	9.13	0.34	65.09	12.01	1.16
Class 2	23.76	17.44	17.00	15.15	33.17	9.43	18.07	32.72	6.96	26.40	18.45	12.90
Class 3	9.50	24.61	28.71	0.65	4.38	58.96	0.53	4.85	58.61	1.04	12.69	50.26

TABLE 4.
Average correlation coefficient of pseudo-residuals within each latent class under the three-class model.

	EM algo.	K-groups	Hier. aggl.	Hier. divi.
Class 1	0.14	0.09	0.12	0.08
Class 2	0.14	0.06	0.17	0.10
Class 3	0.13	0.09	0.12	0.11

by randomly assigning the individual to a class with probabilities $\eta_1(\mathbf{x}_i), \ldots, \eta_3(\mathbf{x}_i)$ calculated from the true model parameter values. The class membership inferred from the EM approach was created through the posterior latent prevalence, as described by Bandeen-Roche et al. (1997) and Huang and Bandeen-Roche (2004). The overall proportions of agreement for EM, k-groups, agglomerative hierarchical, and divisive hierarchical approaches were 40.44%, 79.25%, 80.06%, and 66.9%, respectively.

Table 4 shows the average of pair-wise sample correlations of pseudo-residuals $\{\widetilde{\mathbf{R}}_{im}, m = 1, \ldots, 5\}$ of (13) within each estimated class under the three-class model. The conditional independence assumption is satisfactory for all approaches.

Sparseness of the simulated data significantly affects the EM approach. When fitting the model with the EM algorithm, only 62 out of 100 replicated three-class models actually converged (i.e., the difference in log likelihood between two consecutive EM-iterations was smaller than 0.001 within 200 iterations). All 100 replications converged when using the proposed two-stage approaches.

Tables 5, 6, 7, and 8 present the corresponding results under the six-class model, which are similar to the results of the three-class model. However, due to the much more complex model structure of the six-class model, the proposed methods have worse accuracy, agreement, and independence. The degree of underestimation in standard errors of γ_{mkj} estimates can be large under the six-class model when using the proposed two-stage approach. Fifty-two out of 100 replicated six-class models converged when using the EM approach, whereas all 100 replications converged when using the proposed approaches.

6.2.2. Sensitivity to Model Assumptions A three-class model was generated with conditionally independent indicators within the 1st and 3rd classes but conditionally dependent indicators in the second class ($\nu_j^2 = 49$). Its average root mean square errors (RMSEs) for estimating parameters in conditional probabilities were 8.26, 3.16, 3.54, and 3.46 using the EM, k-groups, agglomerative hierarchical, and divisive hierarchical approaches, respectively. These results were worse than the estimates under the conditional independence model with average RMSE values of 6.70, 2.65, 3.18, and 3.26, respectively. In estimating parameters in latent prevalences, the average RMSE values were 2.53, 0.80, 1.26, and 0.65 under the conditional dependence model, and 1.27, 0.45, 0.50, and 0.62 under the conditional independence model, respectively. A noticeable difference was in the agreement proportion between simulated and inferred class membership for

TABLE 5.
Estimations of the regression coefficients in the conditional probability regression (7) under the six-class model.

	Level	True	EM algo.	K-groups	Hier. aggl.	Hier. divi.
				Indicator 1[a]		
Intercept	1 vs. 3	−6.03	5.43 (216.12/11.70/16.30)[b]	−0.11 (1.00/2.66/6.49)	−1.89 (1.11/6.15/7.38)	−2.59 (1.06/3.95/5.23)
	2 vs. 3	0.42	5.39 (40.84/10.57/11.59)	1.19 (0.90/2.58/2.68)	2.91 (1.01/7.86/8.21)	−0.52 (0.85/2.32/2.49)
Class 1	1 vs. 3	11.92	−4.05 (341.05/16.04/22.53)	10.17 (0.49/5.72/5.95)	12.81 (0.59/7.80/7.81)	9.69 (0.73/6.64/6.97)
	2 vs. 3	2.92	−1.89 (96.12/12.01/12.83)	6.95 (0.50/5.68/6.94)	5.28 (0.59/7.98/8.28)	5.50 (0.65/5.72/6.25)
Class 2	1 vs. 3	5.93	−5.39 (807.34/18.22/21.30)	0.25 (0.70/3.28/6.55)	1.29 (0.69/9.25/10.31)	3.34 (0.73/7.23/7.65)
	2 vs. 3	−0.13	−3.40 (291.25/14.95/15.16)	−0.19 (0.65/3.31/3.29)	−2.09 (0.60/11.16/11.28)	1.86 (0.61/4.56/4.95)
Class 3	1 vs. 3	7.98	−5.52 (821.14/18.02/22.38)	5.51 (0.63/6.62/7.03)	3.97 (0.70/8.68/9.52)	5.52 (0.73/5.86/6.33)
	2 vs. 3	1.54	−4.16 (286.89/16.94/17.72)	4.56 (0.61/6.66/7.29)	−1.00 (0.63/9.31/9.61)	4.10 (0.58/4.62/5.26)
Class 4	1 vs. 3	6.88	−2.83 (675.12/17.24/19.65)	2.97 (0.66/5.96/7.11)	0.67 (0.58/9.51/11.32)	4.16 (0.78/6.45/6.97)
	2 vs. 3	2.56	−1.44 (182.13/14.12/14.55)	2.51 (0.62/5.82/5.79)	−0.30 (0.74/12.66/12.91)	2.78 (0.68/4.91/4.89)
Class 5	1 vs. 3	0.35	−0.68 (625.10/17.16/17.02)	−5.30 (0.74/3.88/6.83)	−5.56 (0.74/7.60/9.59)	−3.42 (0.78/5.14/6.35)
	2 vs. 3	−4.83	−0.84 (249.99/15.48/15.84)	−3.97 (0.55/2.62/2.74)	−7.55 (0.68/8.65/9.02)	−2.72 (0.48/3.59/4.15)
Sex	1 vs. 3	−0.84	−0.94 (0.43/2.32/2.30)	−0.04 (0.41/0.60/1.00)	0.11 (0.46/0.56/1.10)	0.08 (0.38/0.48/1.04)
	2 vs. 3	−0.21	−0.44 (0.39/2.28/2.27)	0.18 (0.37/0.49/0.63)	0.27 (0.43/0.54/0.72)	0.27 (0.34/0.39/0.62)
Age	1 vs. 3	0.04	0.03 (0.03/0.04/0.05)	0.04 (0.03/0.03/0.03)	0.03 (0.03/0.03/0.03)	0.04 (0.02/0.03/0.03)
	2 vs. 3	0.02	0.01 (0.03/0.04/0.04)	0.03 (0.02/0.02/0.03)	0.02 (0.03/0.03/0.03)	0.02 (0.02/0.02/0.02)

TABLE 5.
(*Continued.*)

	Level	True	EM algo.	K-groups	Hier. aggl.	Hier. divi.
				Indicator 5		
Intercept	1 vs. 3	0.20	1.38 (3.03/5.74/5.81)	0.45 (0.73/0.79/0.83)	1.13 (0.90/6.08/6.12)	−0.29 (0.77/3.19/3.21)
	2 vs. 3	0.89	0.81 (1.07/6.05/5.99)	0.97 (0.69/0.77/0.77)	2.36 (0.87/6.06/6.21)	0.28 (0.68/1.66/1.76)
Class 1	1 vs. 3	3.30	−0.07 (7.30/8.05/8.65)	2.94 (0.61/0.88/0.95)	2.93 (0.74/6.42/6.40)	4.88 (0.61/4.18/4.45)
	2 vs. 3	0.48	0.71 (1.85/7.33/7.26)	0.57 (0.65/0.87/0.87)	−0.63 (0.79/6.59/6.65)	2.25 (0.59/3.65/4.04)
Class 2	1 vs. 3	0.11	−0.89 (6.07/8.83/8.81)	0.13 (0.54/0.93/0.93)	−2.74 (0.61/10.72/11.04)	0.95 (0.62/4.53/4.58)
	2 vs. 3	−0.38	0.26 (1.54/8.79/8.73)	−0.10 (0.52/0.90/0.94)	−3.15 (0.60/10.80/11.09)	0.91 (0.57/2.74/3.02)
Class 3	1 vs. 3	1.80	0.10 (9.32/8.47/8.55)	1.39 (0.57/2.42/2.44)	0.56 (0.63/7.70/7.76)	1.74 (0.57/3.79/3.77)
	2 vs. 3	2.06	1.05 (1.60/7.67/7.67)	1.77 (0.54/2.47/2.48)	0.99 (0.63/8.76/8.78)	1.89 (0.47/2.52/2.51)
Class 4	1 vs. 3	0.37	1.09 (3.25/6.97/6.94)	0.31 (0.54/0.84/0.83)	−4.12 (0.47/10.44/11.32)	1.14 (0.60/3.44/3.51)
	2 vs. 3	−0.07	1.52 (1.43/7.15/7.26)	0.02 (0.53/0.90/0.90)	−4.02 (0.47/12.12/12.69)	0.56 (0.56/1.74/1.85)
Class 5	1 vs. 3	−2.10	1.12 (3.61/7.14/7.76)	−1.82 (0.50/0.71/0.76)	−3.88 (0.69/6.99/7.18)	−1.40 (0.61/3.41/3.46)
	2 vs. 3	−1.23	1.32 (3.31/8.16/8.47)	−1.25 (0.43/0.68/0.67)	−3.60 (0.63/6.88/7.25)	−0.55 (0.46/1.54/1.68)
Sex	1 vs. 3	−0.77	−0.87 (0.29/0.37/0.38)	−0.55 (0.28/0.33/0.40)	−0.55 (0.32/0.41/0.47)	−0.52 (0.28/0.31/0.40)
	2 vs. 3	−0.59	−0.65 (0.27/0.30/0.30)	−0.49 (0.27/0.28/0.29)	−0.53 (0.31/0.36/0.37)	−0.47 (0.27/0.28/0.30)
Age	1 vs. 3	0.0003	−0.01 (0.02/0.02/0.02)	−0.001 (0.02/0.02/0.02)	−0.001 (0.02/0.02/0.02)	0.001 (0.02/0.02/0.02)
	2 vs. 3	−0.01	−0.01 (0.02/0.02/0.02)	−0.01 (0.02/0.02/0.02)	−0.01 (0.02/0.02/0.02)	−0.01 (0.02/0.02/0.02)

[a]Only results for Indicators 1 and 5 are shown. Model fit for Indicator 1 is the worst among five indicators, whereas model fit for Indicator 5 is the best.

[b]Numbers in parentheses: average of standard errors from the conditional probability regression/empirical standard error based on simulation replications/root mean square error.

TABLE 6.
Estimations of the regression coefficients in the latent prevalence regression (3) under the six-class model.

	True	EM algo.	K-groups	Hier. aggl.	Hier. divi.
			Class 1 vs. Class 6		
Intercept	−1.09	0.43 (0.30/1.26/1.97)[a]	−0.23 (0.33/0.43/0.96)	0.04 (0.40/0.89/1.43)	−0.48 (0.35/0.71/0.93)
Occupation	0.27	0.02 (1.73/1.98/1.98)	0.65 (0.39/0.39/0.54)	0.73 (1.21/1.49/1.55)	0.53 (0.40/0.36/0.44)
Dprime	0.69	−0.05 (0.10/0.47/0.87)	0.38 (0.10/0.12/0.33)	0.50 (0.13/0.15/0.24)	0.42 (0.11/0.17/0.32)
			Class 2 vs. Class 6		
Intercept	0.43	−0.02 (0.35/1.72/1.76)	−0.002 (0.33/0.45/0.62)	0.23 (0.43/1.22/1.23)	−0.45 (0.38/1.001/1.33)
Occupation	−1.74	−0.13 (2.97/2.60/3.04)	−0.10 (0.48/0.48/1.71)	−0.34 (3.28/1.91/2.36)	−0.01 (0.49/0.53/1.81)
Dprime	0.01	−0.06 (0.11/0.45/0.45)	0.03 (0.11/0.13/0.13)	0.01 (0.16/0.21/0.21)	0.16 (0.12/0.24/0.28)
			Class 3 vs. Class 6		
Intercept	−0.67	0.12 (0.33/1.53/1.70)	0.05 (0.32/0.42/0.83)	0.81 (0.39/1.31/1.97)	0.33 (0.31/0.92/1.35)
Occupation	−0.85	−0.40 (1.82/3.40/3.39)	−0.08 (0.43/0.47/0.90)	−0.06 (1.25/1.52/1.70)	−0.16 (0.42/0.45/0.83)
Dprime	0.56	−0.06 (0.10/0.49/0.78)	0.19 (0.11/0.13/0.40)	0.25 (0.13/0.18/0.36)	0.18 (0.10/0.18/0.42)
			Class 4 vs. Class 6		
Intercept	−0.46	0.12 (0.32/1.33/1.43)	−0.16 (0.34/0.45/0.54)	−0.03 (0.46/1.27/1.34)	−0.35 (0.36/0.89/0.89)
Occupation	−0.26	−0.22 (1.90/2.13/2.11)	0.001 (0.47/0.51/0.57)	−0.13 (3.28/1.93/1.93)	0.10 (0.47/0.57/0.67)
Dprime	0.25	−0.03 (0.10/0.51/0.58)	0.10 (0.11/0.14/0.20)	0.10 (0.16/0.21/0.26)	0.19 (0.12/0.20/0.21)
			Class 5 vs. Class 6		
Intercept	0.17	−0.15 (0.34/1.22/1.25)	0.98 (0.28/0.38/0.90)	1.03 (0.36/0.90/1.24)	0.46 (0.30/0.93/0.97)
Occupation	−0.62	0.11 (1.79/1.89/2.01)	0.004 (0.41/0.41/0.75)	0.02 (1.24/1.53/1.65)	−0.08 (0.44/0.44/0.70)
Dprime	0.14	0.08 (0.11/0.43/0.43)	−0.09 (0.10/0.10/0.25)	0.02 (0.13/0.16/0.20)	−0.03 (0.11/0.18/0.25)

[a]Numbers in parentheses: average of standard errors from the conditional probability regression/empirical standard error based on simulation replications/root mean square error.

TABLE 7.
Matched number of individuals between simulated and obtained/inferred class membership under the six-class model with total 500 individuals, averaging over 100 replications.

Sim.\Est.	EM algo.						K-groups					
	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
Class 1	20.50	16.67	18.67	19.40	28.98	28.81	110.26	4.44	8.91	7.64	0.19	2.96
Class 2	15.08	10.52	11.31	11.92	7.92	11.48	4.77	16.23	10.36	10.59	13.21	13.36
Class 3	25.42	12.33	9.75	17.14	19.23	13.96	14.07	12.19	43.55	16.16	1.15	10.60
Class 4	12.23	9.71	8.94	10.40	9.46	9.40	9.39	9.68	15.11	15.13	1.75	8.82
Class 5	10.31	16.64	18.75	10.73	13.71	9.73	0.07	2.05	0.35	0.99	72.54	2.38
Class 6	16.73	7.35	10.10	10.14	6.56	10.02	0.24	9.37	5.67	6.85	24.68	14.29
Sim.\Est.	Hier. aggl.						Hier. divi.					
	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
Class 1	111.18	1.95	15.66	4.82	0.03	0.76	88.91	12.84	11.16	17.75	0.19	3.55
Class 2	4.72	17.23	18.78	13.07	6.35	8.37	5.19	11.18	21.71	9.96	9.93	10.55
Class 3	10.44	5.66	63.29	14.16	0.30	3.87	13.78	12.84	49.39	17.38	1.03	3.30
Class 4	7.20	7.44	27.31	15.04	0.43	2.46	10.35	7.87	22.38	14.53	1.63	3.12
Class 5	0.01	3.81	1.59	2.72	64.63	5.62	0.64	5.84	1.77	2.54	50.14	17.45
Class 6	0.08	10.51	11.47	8.80	12.39	17.85	0.72	6.74	11.83	5.17	18.90	17.74

TABLE 8.
Average correlation coefficient of pseudo-residuals within each latent class under the six-class model.

	EM algo.	K-groups	Hier. aggl.	Hier. divi.
Class 1	0.09	0.06	0.08	0.08
Class 2	0.12	0.05	0.17	0.08
Class 3	0.10	0.05	0.11	0.06
Class 4	0.10	0.05	0.19	0.09
Class 5	0.10	0.05	0.10	0.07
Class 6	0.10	0.05	0.19	0.06

the second class in which the conditional independence assumption was not satisfied. The second class agreement proportions were 9.08%, 3.93%, 3.96%, and 8.40% under the conditional dependence model, and 8.72%, 16.59%, 16.36%, and 9.23% under the conditional independence model, respectively. Also, the conditional dependence model resulted in a larger within-class sample correlation of pseudo-residuals (0.28, 0.21, 0.29, and 0.21, respectively) than the conditional independence model (0.14, 0.08, 0.14, and 0.10, respectively). This shows that the conditional independence assumption can be checked by examining the correlation coefficient of pseudo-residuals within each latent class.

When applying the proposed approaches to a three-class model with two independent sets of covariates $\mathbf{x}_i$ and $\mathbf{z}_{im}$, the obtained results were not significantly different from those produced by the model with correlated $\mathbf{x}_i$ and $\mathbf{z}_{im}$. When applying the EM, k-groups, agglomerative hierarchical, and divisive hierarchical approaches, the average RMSE values for parameters in conditional probabilities were 7.32, 2.92, 5.26, and 3.21, respectively; the average RMSE values for parameters in latent prevalences were 1.43, 0.42, 0.51, and 0.59, respectively; the overall class agreement proportions were 30.76%, 80.37%, 79.52%, and 68.27%, respectively; and the within-class sample correlations of pseudo-residuals were 0.10, 0.07, 0.13, and 0.09, respectively.

6.3. Summary

In summary, the proposed clustering methods outperform the traditional EM approach in sparse response patterns for measured indicators. The proposed clustering methods yield accurate parameter and standard error estimates. As for assigning individuals to the class that they belong to, the proposed clustering methods' overall correct rates are high, with the divisive hierarchy ranking lowest among three methods. Based on observations from model (1), the conditional independence assumption appears to be satisfied in estimates from the proposed approaches. When the true underlying model contains conditionally dependent indicators within latent classes, this assumption violation can be detected by examining sample correlations of pseudo-residuals within each latent class. The proposed two-stage approach does not seem sensitive to the assumption of orthogonality between $\mathbf{z}_{im}$ and the class membership.

7. Example

7.1. Breast Cancer Data

This paper uses data from a study using gene expression profiling to predict breast cancer outcomes (van't Veer, Dai, van de Vijver, He, Hart, Mao, Peterse, van der Kooy, Marton, Witteveen, Schreiber, Kerkhoven, Roberts, Linsley, Bernards, & Friend, 2002). A total of 78 sporadic lymph-node-negative breast cancer patients under 55 years of age were examined for a prognostic signature in their gene expression profiles. Forty-four patients remained disease-free

for an interval of at least five years after their initial diagnosis (good prognosis group), while 34 patients had developed distant metastases within five years (poor prognosis group).

Gene expression is the level of the process where the DNA of a gene is transcribed to RNA. Differences in gene expression of certain genes from good versus poor prognostic patients suggest that these genes may induce different function in the two patient groups. Therefore, the expression profile over a set of pre-selected genes can be used to distinguish good from poor prognostic patients. In this study, microarray technology was used to simultaneously measure the gene expression of 25,000 pre-selected genes for each patient (Hughes, Mao, Jones, Burchard, Marton, Shannon, Lefkowitz, Ziman, Schelter, Meyer, Kobayashi, Davis, Dai, He, Stephaniants, Cavet, Walker, West, Coffey, Shoemaker, Stoughton, Blanchard, Friend, & Linsley, 2001). Microarrays consist of thousands of individual DNA fragments spotted in a high-density glass slide to which fluorescently labeled RNA isolated from patients is hybridized. After thorough washing, the slide is imaged using a laser scanner and fluorescence measurements are made for each spot. The fluorescence measurement indicates the abundance of the corresponding DNA fragment in the RNA sample.

This study uses a selection of 25,000 gene expression to predict group membership (good versus poor prognosis) by positing a latent variable that mediates between the gene expression and the group membership, and then uses the results to classify new cases into the two prognosis groups. We adopt the latent profile approach to reduce the dimensionality of microarray data to a latent variable, which could reduce noise while still capturing the main biological properties of the original data. A preliminary two-step selection process was performed to retain genes in the analysis. The first step selected 4741 genes with the intensity ratio >2 or <0.5 (i.e., more than two-fold difference) and the significance of regulation p-value <0.01 in more than three patients. This was used in the original paper to focus the attention on the most informative genes. The second step selected genes based on the ratio of their between-group to within-group sums of squares, as suggested by Dudoit, Fridlyand, and Speed (2002). For a gene m, this ratio is

$$\mathrm{BW}(m) = \frac{\sum_i \sum_a \mathrm{I}(d_i = a)(\bar{y}_{am} - \bar{y}_{.m})^2}{\sum_i \sum_a \mathrm{I}(d_i = a)(y_{im} - \bar{y}_{am})^2}, \tag{20}$$

where y_{im} denotes the intensity ratio of gene m in patient i, d_i is the indicator of good (=1) or poor (=0) prognosis group of patient i, and $\bar{y}_{am}$ and $\bar{y}_{.m}$ are the average intensity ratios of gene m across samples belonging to prognosis group a only and across all patients, respectively. Equation (20) was used to compute the BW ratio for each gene and the top 70 genes with the largest BW ratios were selected for the finite mixture analysis.

The latent profile model of (1), (3), and (5) was then fitted using the 70 selected gene expression ratios as observed indicators. In the fitted model, the age at diagnosis (year) was correlated with conditional probabilities, and latent prevalence was modeled as depending on age at diagnosis. Figures 4 and 5 present the heatmaps for the 70-gene expression profile, showing patients ordered by the dendrograms created from the proposed agglomerative hierarchical (AH1–AH3) and divisive hierarchical (DH1–DH4) clustering methods, respectively. The agglomerative hierarchical method divided patients into two classes of 32 and 46, while the divisive hierarchical method divided patients into three classes of 13, 19, and 46. The k-groups clustering approach (K1–K3) grouped patients in three classes of 16, 24, and 38. The heatmaps showed that the 70 included genes can be divided into two different sets. Patients in the first class from the agglomerative hierarchical method were those with high expression levels on the first set of genes but lower expression levels on the second set of genes, while patients in the second class behaved reversely. The divisive hierarchical method further divided the first class from the agglomerative hierarchical method into two with one having higher expression levels on the first set of genes than the other. The divisive hierarchical method and k-groups method produced similar expression profiles for classes. All three clustering algorithms obtained satisfactory class allocation

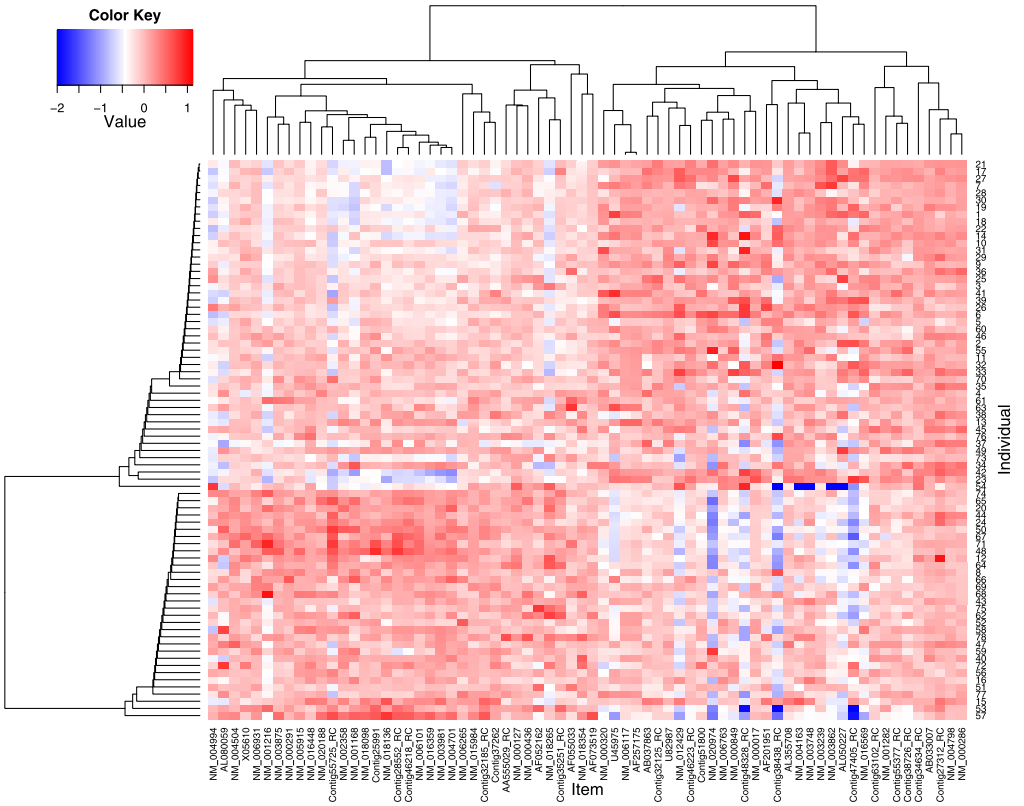

FIGURE 4.
Heatmap for breast cancer data with patients clustered using the proposed agglomerative hierarchical clustering method (AH1–AH3). The column dendrogram for genes is based on the traditional agglomerative hierarchical clustering method with the distance measure being one minus correlation between two genes.

with average within class correlations of 0.21, 0.21, and 0.22 for the k-groups, agglomerative hierarchy, and divisive hierarchy, respectively.

The leave-one-out cross-validation scheme was performed to estimate the misclassification rate of the proposed classification rule in classifying patients between good and poor prognosis groups. The k-groups, agglomerative hierarchical, and divisive hierarchical approaches produced misclassification rates of 24.36%, 26.92%, and 29.49%, respectively.

As in the original paper, an additional independent set of primary tumors from 19 young, lymph-node-negative breast cancer patients was used to validate the above 70-gene prognosis classifier. This group included seven patients who remained free of disease for at least five years, and 12 patients who developed distant metastases within five years. Consequently, the k-groups, agglomerative hierarchical, and divisive hierarchical approaches had three, three, and two out of 19 incorrect classifications, respectively.

7.2. Schizophrenia Syndrome Scale Data

This section uses data from a series of studies, investigating the clinical manifestations of schizophrenia and searching for the neuropsychological, environmental, and genetic factors underlying schizophrenia. The details of study design and eligibility criteria have been described previously (Liu, Hwu, & Chen, 1997; Chen, Liu, Chang, Lien, Chang, & Hwu, 1998; Chang, Chen, Liu, Cheng, Ou Yang, Chang, Lane, Lin, Yang, & Hwu, 2002). The analyzed data include

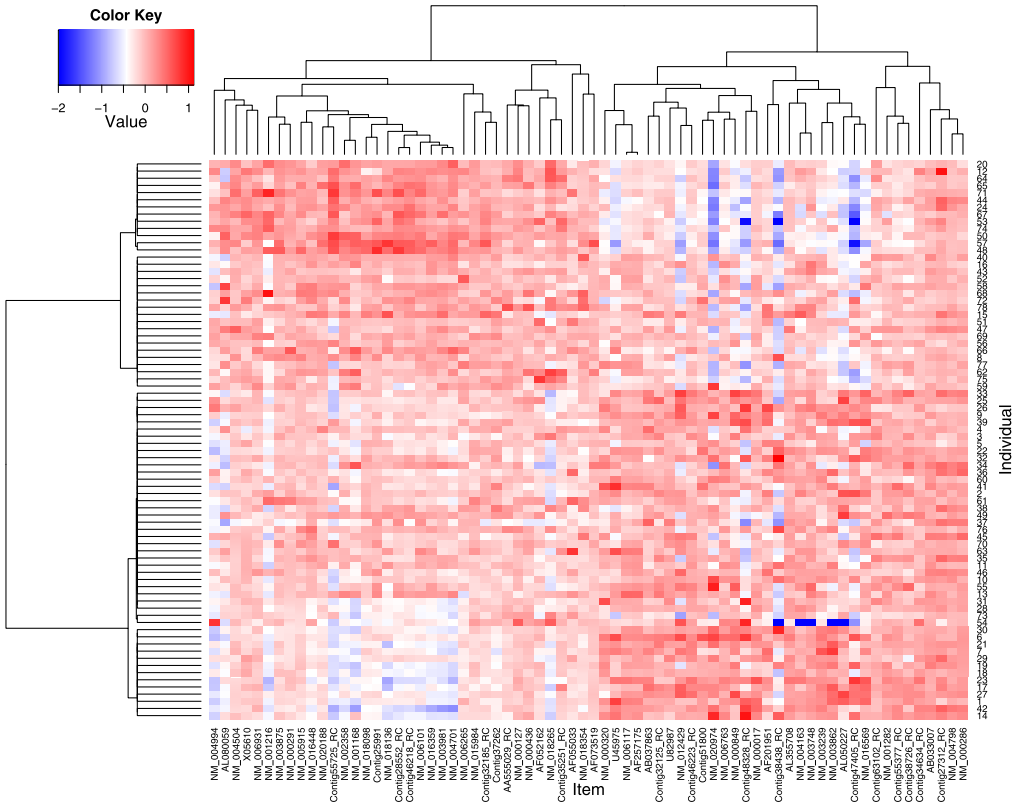

FIGURE 5.
Heatmap for breast cancer data with patients clustered using the proposed divisive hierarchical clustering method (DH1–DH4). The column dendrogram for genes is based on the traditional agglomerative hierarchical clustering method with the distance measure being one minus correlation between two genes.

169 patients with acute schizophrenia recruited within one week of index admission and 160 subsided state patients living in the community under family care.

The schizophrenia symptoms used in this study were assessed by the Positive and Negative Syndrome Scale (PANSS) (Cheng, Ho, Chang, Lane, & Hwu, 1996). The PANSS has 30 items and consists of three subscales: positive (seven symptoms: P1–P7), negative (seven symptoms: N1–N7) and general psychopathology (sixteen symptoms: G1–G16). Each item was originally rated on a 7-point scale (1 = absent, 7 = extreme), but this scale was reduced by merging the points with response percentages of less than 10%. This study considered external covariates including demographic variables and environmental/neuropsychological factors. Demographic variables included gender, age at recruitment, years of education, and occupation (versus no occupation). Environmental/neuropsychological factors included the onset-age of psychotic symptoms and the sensitivity index of the Continuous Performance Test (CPT), which is widely used to measure sustained attention deficits in psychotic disorders (Chen et al., 1998).

This study explores the subtypes (groups) of schizophrenia patients based on PANSS measurements. In this application, the latent class model of (1), (3), and (4) was applied to 30 PANSS items. Given that demographic variables might act as the extraneous influences to affect an individual's measurements of PANSS, the analysis included these variables as z_{iml}'s in (4) to obtain latent classes that more accurately reflected underlying schizophrenia subtypes. Environmental/neuropsychological factors were denoted as x_{ip}'s in (3), which enables us to model schizophrenia subtypes as depending on these factors. Figures 6 and 7 show the heatmaps for

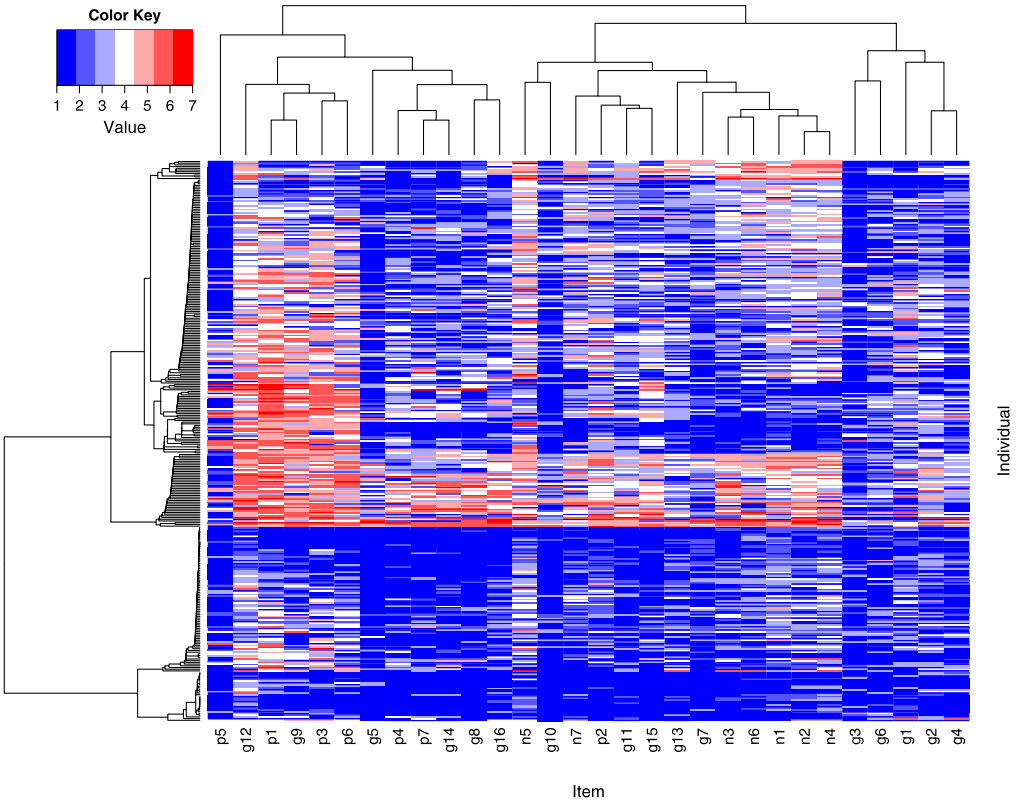

FIGURE 6.
Heatmap for schizophrenia data with patients clustered using the proposed agglomerative hierarchical clustering method (AH1–AH3). The column dendrogram for genes is based on the traditional agglomerative hierarchical clustering method with the distance measure being one minus correlation between two genes.

PANSS symptom patterns with patient subtypes identified by the proposed agglomerative hierarchical and divisive hierarchical clustering methods, respectively. Both clustering methods divided patients into four subtypes. The k-groups clustering approach was also implemented, producing four subtypes. Three clustering methods identified similar symptom patterns for classes. Class 1 generally represented a group with severe/extreme positive symptoms and moderate negative symptoms. Class 2 exhibited severe positive and negative symptoms. Class 3 had moderate positive symptoms but mild negative symptoms. Finally, class 4 was a remitted group with only rare symptoms. Satisfactory class allocations were obtained with average within class correlations of 0.16, 0.18, and 0.18 for k-groups, agglomerative hierarchy, and divisive hierarchy, respectively.

Several authors have pointed out that the symptom structure in schizophrenia may depend on the phase of disease chronicity (Mohr, Cheng, Claxton, Conley, Feldman, Hargreaves, Lehman, Lenert, Mahmoud, Marder, & Neumann, 2004). This study thus aims to use the PANSS ratings to predict patients' phases of disease chronicity (acute versus subsided). The leave-one-out cross-validation was performed to evaluate the proposed classification method. As a result, the misclassification rates were 23.10%, 24.01%, and 28.27% for the k-groups, agglomerative hierarchical, and divisive hierarchical approaches, respectively.

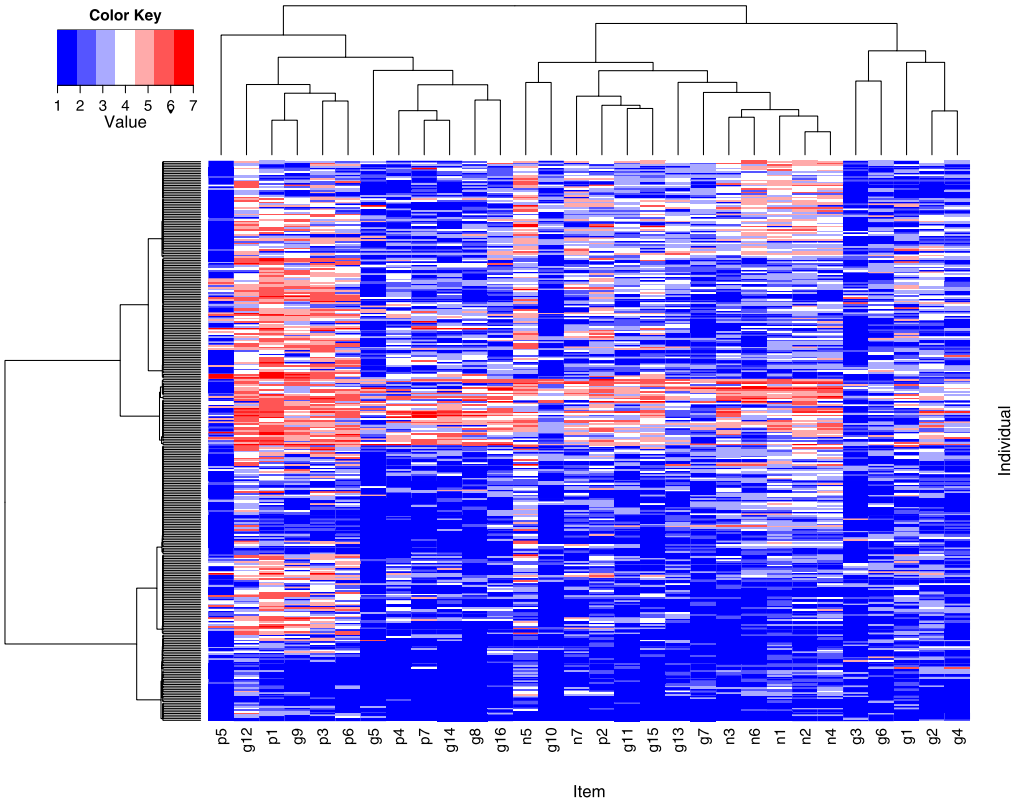

FIGURE 7.
Heatmap for schizophrenia data with patients clustered using the proposed divisive hierarchical clustering method (DH1–DH4). The column dendrogram for genes is based on the traditional agglomerative hierarchical clustering method with the distance measure being one minus correlation between two genes.

8. Discussion

This paper presents the k-groups and hierarchical clustering methods to search for the optimal class allocation that makes measured indicators as independent as possible for objects belonging to the same class. These proposed methods adopt a clustering algorithm based on k-means and hierarchical clustering, but using the psychometric criterion of local (conditional on latent class) independence rather than the usual criterion of distance to cluster mean. Treating the identified class allocation as a known predictor makes it possible to estimate the parameters underlying LC/LP models. This approach is theoretically justifiable, allows direct checking of the conditional independence assumption, and converges much faster than the full likelihood approach when analyzing high-dimensional data. This paper further develops a classification rule based on the finite mixture model. Simulation results show that the proposed clustering methods outperform the traditional EM approach when measured indicators exhibit sparse response patterns. The real data analysis in this study confirms the ability of the proposed methods to handle high-dimensional data, and confirms the accuracy of the proposed classification rule in predicting the disease statuses of new observations.

The current study can be improved and extended in several aspects. First, this study used pair-wise sample covariance as the measure of independence among indicators, which is straightforward and easy to calculate. However, it is not appropriate to use the sample covariance to represent the association between two random variables when the sample size is small. At the

early stage of the agglomerative hierarchical approach, each class contains only very few objects (e.g., at the initial stage, there is only one object for each class). Thus, any wrong reallocation of objects at an early stage will result in wrong reallocation of objects in the following stages. Alternative measures of independence among variables can be used for the improvement. Second, for large dimensional covariance matrices, the sample covariance matrix can be non-invertible, numerically ill-conditioned, and a very inaccurate estimate of the true covariance matrix (Ledoit & Wolf, 2004). As a result, a modified covariance matrix estimate (e.g., the one proposed by Ledoit & Wolf, 2004) should be used when handling high dimensional data. Third, a relatively small unknown subset of observed indicators contributes to the clustering of high-dimensional data. Therefore, algorithms that simultaneously perform variable selection and object clustering (e.g., Brusco & Cradit, 2001; Friedman & Meulman, 2004) can be used to replace regular clustering algorithms.

Acknowledgements

This research was partially supported by grants from the National Science Council, Taiwan (NSC98-2118-M-009-002-MY2 and NSC98-3112-B-001-027). The authors wish to thank Dr. Hai-Gwo Hwu for kindly providing the schizophrenia syndrome scale data. We are grateful to the National Center for High-performance Computing for computer time and facilities. We also thank the Editor and two referees for their valuable comments. Conflict of Interest: None declared.

Appendix: Proof of Theorem 1

For simplicity, the following proofs focus on the latent class model (i.e., measured indicators are categorical variables) with $K_1 = \cdots = K_M = K$ (i.e., the levels of items are all the same). Extension to allow the levels to be different and to the latent profile model is straightforward.

First consider the following two equalities:

$$\begin{aligned}
\Pr(\mathbf{Y}_i = \mathbf{y} | C_{in} = j, \mathbf{z}_i) &= \prod_{m=1}^{M} \Pr(Y_{im} = y_m | C_{in} = j, \mathbf{z}_{im}) \\
&= \prod_{m=1}^{M} \mathrm{E}\big[\Pr\big(Y_{im} = y_m \big| C_{in} = j, \mathbf{z}_{im}, \widehat{\boldsymbol{\gamma}}_{mn}, \widehat{\boldsymbol{\alpha}}_{mn}\big)\big] \\
&= \prod_{m=1}^{M} \mathrm{E}\left[\prod_{k=1}^{K} \big(\widehat{p}_{imkjn}\big)^{y_{mk}}\right] \\
&\xrightarrow[n\to\infty]{} \prod_{m=1}^{M} \prod_{k=1}^{K} \big(p^*_{imkj}\big)^{y_{mk}}, \qquad \text{(A.1)}
\end{aligned}$$

where $\widehat{\boldsymbol{\gamma}}_{mn} = (\widehat{\gamma}_{mkjn},\ \text{for all } k, j)$; $\widehat{\boldsymbol{\alpha}}_{mn} = (\widehat{\alpha}_{lmkn},\ \text{for all } l, k)$; $\boldsymbol{\gamma}^*_m = (\gamma^*_{mkj},\ \text{for all } k, j)$; $\boldsymbol{\alpha}^*_m = (\alpha^*_{lmk},\ \text{for all } l, k)$; $p_{imkj} = p_{mkj}(\mathbf{z}_{im})$; $\widehat{p}_{imkjn}$ is p_{imkj} evaluated at $(\widehat{\boldsymbol{\gamma}}_{mn}, \widehat{\boldsymbol{\alpha}}_{mn})$; p^*_{imkj} is p_{imkj} evaluated at $(\boldsymbol{\gamma}^*_m, \boldsymbol{\alpha}^*_m)$; and $y_{mk} = 1$ if $y_m = k$; 0 otherwise. The first moment convergence of the last line of (A.1) follows from the uniform integrability of $\widehat{p}_{imkjn}$ and the convergence in probability of $(\widehat{\boldsymbol{\gamma}}_{mn}, \widehat{\boldsymbol{\alpha}}_{mn})$ to $(\boldsymbol{\gamma}^*_m, \boldsymbol{\alpha}^*_m)$. Similarly,

$$\lim_{n\to\infty} \Pr(C_{in} = j | \mathbf{x}_i) = \lim_{n\to\infty} \mathrm{E}\big[\Pr\big(C_{in} = j | \mathbf{x}_i, \widehat{\boldsymbol{\beta}}_n\big)\big] = \lim_{n\to\infty} \mathrm{E}\big[\widehat{\eta}_{ijn}\big] = \eta^*_{ij}, \qquad \text{(A.2)}$$

where $\widehat{\boldsymbol{\beta}}_n = (\widehat{\beta}_{pjn},\ \text{for all } p, j)$; $\boldsymbol{\beta}^* = (\beta^*_{pj},\ \text{for all } p, j)$; $\eta_{ij} = \eta_j(\mathbf{x}_i)$; $\widehat{\eta}_{ijn}$ is η_{ij} evaluated at $\widehat{\boldsymbol{\beta}}_n$; and η^*_{ij} is η_{ij} evaluated at $\boldsymbol{\beta}^*$.

Next, define two auxiliary random variables that are useful in establishing the results. Let $W^*_i(\mathbf{y})$ be a discrete random variable having frequency function

$$\Pr\bigl(W^*_i(\mathbf{y}) = w|\mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\phi}^*\bigr) = \begin{cases} \eta^*_{ij} & \text{if } w = \prod_{m=1}^{M}\prod_{k=1}^{K}(p^*_{imkj})^{y_{mk}},\ j = 1, \ldots, J, \\ 0 & \text{otherwise,} \end{cases}$$

where $\boldsymbol{\phi}^* = ((\beta^*_{pj}, \gamma^*_{mkj}, \alpha^*_{lmk}),\ \text{for all } p, j, m, k, l)$. Further, let $W_{in}(\mathbf{y}) = \Pr(\mathbf{Y}_i = \mathbf{y}|C_{in}, \mathbf{x}_i, \mathbf{z}_i)$. With equalities (A.1) and (A.2), we now want to show the convergence in distribution of $W_{in}(\mathbf{y})$ to $W^*_i(\mathbf{y})$. Notice that

$$\bigl|\Pr\bigl(W_{in}(\mathbf{y}) \le w_0\bigr) - \Pr\bigl(W^*_i(\mathbf{y}) \le w_0\bigr)\bigr| = \Biggl|\sum_{j\in A_{in}(w_0)} \Pr(C_{in} = j|\mathbf{x}_i) - \sum_{j\in A^*_i(w_0)} \eta^*_{ij}\Biggr|, \tag{A.3}$$

where

$$A_{in}(w_0) = \bigl\{j \text{ s.t. } \Pr(\mathbf{Y}_i = \mathbf{y}|C_{in} = j, \mathbf{z}_i) \le w_0\bigr\},$$

$$A^*_i(w_0) = \Biggl\{j \text{ s.t. } \prod_{m=1}^{M}\prod_{k=1}^{K}\bigl(p^*_{imkj}\bigr)^{y_{mk}} \le w_0\Biggr\}.$$

Since (A.1) holds, $A_{in}(w_0)$ and $A^*_i(w_0)$ are equal as $n \to \infty$. Now, because of (A.2), (A.3) converges to 0, and therefore $W_{in}(\mathbf{y}) \xrightarrow{\mathcal{L}} W^*_i(\mathbf{y})$.

Since $W_{in}(\mathbf{y})$ is uniformly integrable,

$$\begin{aligned} W_{in}(\mathbf{y}) \xrightarrow{\mathcal{L}} W^*_i(\mathbf{y}) \Rightarrow \lim_{n\to\infty} \mathrm{E}\bigl[W_{in}(\mathbf{y})\bigr] &= \mathrm{E}\bigl[W^*_i(\mathbf{y})\bigr] \\ &= \sum_{j=1}^{J} \eta^*_{ij} \prod_{m=1}^{M}\prod_{k=1}^{K}\bigl(p^*_{imkj}\bigr)^{y_{mk}}. \end{aligned}$$

Notice that $\mathrm{E}[W_{in}(\mathbf{y})] = \sum_{j=1}^{J}\{\Pr(\mathbf{Y}_i = \mathbf{y}|C_{in} = j, \mathbf{z}_i)\Pr(C_{in} = j|\mathbf{x}_i)\} = \Pr(\mathbf{Y}_i = \mathbf{y}|\mathbf{x}_i, \mathbf{z}_i)$, concluding the proof.

References

Albert, P.S., McShane, L.M., & Shih, J.H. (2001). Latent class modeling approaches for assessing diagnostic error without a gold standard: with applications to p53 immunohistochemical assays in bladder tumors. *Biometrics*, *57*, 610–619.

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

Brusco, M.J., & Cradit, J.D. (2001). A variable selection heuristic for k-means clustering. *Psychometrika*, *66*, 249–270.

Bryant, P., & Williamson, J.A. (1978). Asymptotic behavior of classification maximum likelihood estimates. *Biometrika*, *65*, 273–281.

Celeux, G., & Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, *14*, 315–332.

Chang, C.J., Chen, W.J., Liu, S.K., Cheng, J.J., Ou Yang, W.C., Chang, H.J., Lane, H.Y., Lin, S.K., Yang, T.W., & Hwu, H.G. (2002). Morbidity risk of psychiatric disorders among the first degree relatives of schizophrenia patients in Taiwan. *Schizophrenia Bulletin*, *28*, 379–392.

Chen, W.J., Liu, S.K., Chang, C.J., Lien, Y.J., Chang, Y.H., & Hwu, H.G. (1998). Sustained attention deficit and schizotypal personality features in nonpsychotic relatives of schizophrenic patients. *American Journal of Psychiatry*, *155*, 1214–1220.

Cheng, J.J., Ho, H., Chang, C.J., Lane, S.Y., & Hwu, H.G. (1996). Positive and Negative Syndrome Scale (PANSS): establishment and reliability study of a Mandarin Chinese language version. *Taiwanese Journal Psychiatry*, *10*, 251–258.

Clogg, C.C. (1995). Latent class models. In Arminger, G., Clogg, C.C., & Sobel, M.E. (Eds.) *Handbook of statistical modeling for the social and behavioral sciences* (pp. 311–360). New York: Plenum.

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

Dayton, C.M., & Macready, G.B. (1998). 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.

Dudoit, S., Fridlyand, J., & Speed, T.P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, *97*, 77–87.

Friedman, J.H., & Meulman, J.J. (2004). Clustering objects on subsets of attributes. *Journal of the Royal Statistical Society. Series B*, *66*, 815–849.

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

Huang, G.H. (2005). Selecting the number of classes under latent class regression: a factor analytic analogue. *Psychometrika*, *70*, 325–345.

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

Hughes, T.R., Mao, M., Jones, A.R., Burchard, J., Marton, M.J., Shannon, K.W., Lefkowitz, S.M., Ziman, M., Schelter, J.M., Meyer, M.R., Kobayashi, S., Davis, C., Dai, H., He, Y.D., Stephaniants, S.B., Cavet, G., Walker, W.L., West, A., Coffey, E., Shoemaker, D.D., Stoughton, R., Blanchard, A.P., Friend, S.H., & Linsley, P.S. (2001). Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature Biotechnology*, *19*, 342–347.

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.

Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, *88*, 365–411.

Liu, S.K., Hwu, H.G., & Chen, W.J. (1997). Clinical symptom dimensions and deficits on the continuous performance test in schizophrenia. *Schizophrenia Research*, *25*, 211–219.

Lubke, G.H., Carey, G., Lessem, J., & Hewitt, J. (2008). Using observed genetic variables to predict latent class membership: a comparison of two methods. *Behavior Genetics*, *38*, 612–653.

Lux, V., & Kendler, K.S. (2010). Deconstructing major depression: a validation study of the DSM-IV symptomatic criteria. *Psychological Medicine*, *40*, 1679–1690.

Marriott, F.H.C. (1975). Separating mixtures of normal distributions. *Biometrics*, *31*, 767–769.

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.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543.

Mohr, P.E., Cheng, C.M., Claxton, K., Conley, R.R., Feldman, J.J., Hargreaves, W.A., Lehman, A.F., Lenert, L.A., Mahmoud, R., Marder, S.R., & Neumann, P. (2004). The heterogeneity of schizophrenia in disease states. *Schizophrenia Research*, *71*, 83–95.

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. (2007). *Mplus user's guide* (5th ed.). Los Angeles: Muthén & Muthén.

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

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Kopenhagen: Nielsen & Lydiche.

Rosvold, H.E., Mirsk, A.F., Sarason, I., Bransome, E.D. Jr., & Bech, L.H. (1956). A continuous performance test of brain damage. *Journal of Consulting Psychology*, *20*, 343–350.

Titterington, D.M., Smith, A.F., & Makov, U.E. (1985). *Statistical analysis of finite mixture distributions*. New York: Wiley.

van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R., & Friend, S.H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, *415*, 530–536.

Manuscript Received: 10 JUN 2010
Final Version Received: 21 MAR 2011
Published Online Date: 12 OCT 2011