A Parametric Mixture Model for Clustering Multivariate Binary Data

Ajit C. Tamhane[1], Dingxi Qiu*[2] and Bruce E Ankenman[1]

[1] *Department of Industrial Engineering & Management Sciences, Northwestern University, Evanston, IL 60208, USA*

[2] *Department of Industrial Engineering, University of Miami, Coral Gables, FL 33146, USA*

Received 19 March 2008; revised 10 September 2009; accepted 2 November 2009
DOI:10.1002/sam.10063
Published online 30 December 2009 in Wiley InterScience (www.interscience.wiley.com).

Abstract: The traditional latent class analysis (LCA) uses a mixture model with binary responses on each subject that are independent conditional on cluster membership. However, in many practical applications, the responses are correlated because they are observed on the same subject; this is known as local dependence. In this paper, we extend the LCA model to allow for local dependence in each cluster to improve clustering accuracy. The clustering problem is hard because of its unsupervised learning nature (the true cluster memberships and even the true number of clusters are unknown), the difficulty of estimating a correlation matrix for each cluster and the paucity of information in binary data. Therefore, we follow a parametric approach in which we fit a mixture model whose components follow multivariate Bernoulli distributions (one for each cluster). An extension of a family of parametric models by Oman and Zucker [1] is adopted for this purpose and the maximum likelihood estimation method is used for fitting. The Bayesian information criterion (BIC) due to Schwarz [2] is employed to select the number of clusters. Subjects are classified to clusters using the maximum posterior rule. The proposed method is tested and compared with the LCA method via simulation and by applying both methods to two real data sets. Significant improvement is demonstrated relative to the LCA method. © 2009 Wiley Periodicals, Inc. Statistical Analysis and Data Mining 3: 3–19, 2010

Keywords: Bayes classification rule; Bayesian information criterion (BIC); data mining; EM algorithm; latent class analysis; maximum likelihood estimation

1. INTRODUCTION

In this paper, we study the problem of cluster analysis for multivariate binary data. The problem of cluster analysis is, of course, well-known, and needs no introduction; see Refs [3,4] for overviews and Ref. [5] for a recent perspective. Applications of clustering are numerous. For example, a common problem in database marketing is segmentation of customers so that different marketing strategies can be targeted at different segments.

Much of the work in cluster analysis assumes continuous data. However, the problem of clustering for binary data arises in many applications including biology, education, engineering, marketing, medicine and psychology. Hoff [6] discusses an example dealing with tumor classification. When a cell undergoes tumorigenesis, it accumulates abnormalities at multiple sites in its genome, which can be indicated by 1 if the mutation occurs at a particular site and 0 otherwise. These binary outcomes are not independent. Because different tumors affect different biochemical pathways and hence require different types of treatments, it is desirable to classify these tumors based on their abnormality patterns. In marketing, Experian's behavior data bank has lifestyle indicators. MRI, Scarbourough and Simmons have large banks of questions indicating whether a customer has purchased certain brands, watches certain TV shows, reads certain newspapers/magazines, etc.

In the education field, Bennett and Jordan [7] conducted a survey of 468 teachers in which each teacher was asked 38 yes–no type questions (also referred to as items) about the way they handle their classes. For example, one of the questions was "Do you usually allow your pupils to move around the classroom?" The goal in Bennett and Jordan's analysis was to group the teachers into clusters with similar teaching styles (which turn out to be traditional/disciplinarian teachers and modern/lenient teachers). Toward this end, Aitkin, Anderson and Hinde [8] used the *latent class analysis (LCA)* model, which is a *mixture model* of independent Bernoulli distributions; see Refs [4] (p. 120) and [9] (p. 6). Conditional on the cluster membership, responses are assumed to be independent (hence referred to as *local independence*) and the mixing proportions represent the prior probabilities of the clusters. The

Correspondence to: Dingxi Qiu (dingxi@miami.edu)

local independence assumption is clearly not true in this and many other problems because many questions are related. For example, in the teaching styles study another question was "Do you usually allow your pupils to talk to one another?" The answers to the two questions by the same teacher will be positively correlated even if the teacher type is known. This example is reanalyzed using the proposed model in Section 8.1.

In Section 8.2, we analyze another example using the proposed model which aims to cluster newspaper readers based on their responses to seven questions on whether they read a particular newspaper on each day of the week. Here the clusters turn out to be subscribers and non-subscribers. Once again the responses are highly correlated within each cluster (with different correlation matrices) and so the LCA model does not hold. In both these examples, the proposed model which takes into account the dependencies gives a significantly better fit to the data than the LCA model.

One can delete the obviously redundant questions with very highly correlated responses, but it is not always obvious which responses should be deleted. Also, in less extreme cases deleting items would be wasteful of information. In fact, it would be advantageous to explicitly model *local dependencies* and utilize them in clustering the subjects.

In this paper, we follow a parametric mixture model approach in which local dependencies are modeled using a new multivariate Bernoulli distribution. This approach allows us to fit the model to data having general correlation structures, and exploit the information in the correlations to obtain improved clustering performance [higher correct classification rates (CCRs)].

Here are the salient features of the proposed method.

- The method permits relaxing the local independence assumption implicit in traditional LCA and allows explicit modeling of the correlation structure within each cluster. This latter feature enables a better interpretation of the relationships between variables within each cluster. Also, a significantly better fit to the data is obtained compared with the LCA method.

- No assumptions are required other than the specific correlation model. The method allows for different correlation matrices (from the same family of models) for different clusters.

- The computational burden of the method (caused in large part due to estimation of the different correlation matrices within clusters) limits its applicability to a rather small number of variables—about 10 to 15. This is a definite drawback if the interest is in exploratory studies with a large number of variables; on the other hand, the methodology is highly useful for confirmatory and interpretive purposes for a relatively small number of variables.

- The method was validated for robustness by testing it against the violation of the assumed correlation model using simulated data generated from alternative models. The method was also tested on real data from two case studies mentioned earlier.

The outline of the paper is as follows. Section 2 gives a brief literature review of the approaches used to model local dependencies. Section 3 generalizes a multivariate Bernoulli distribution proposed by Oman and Zucker's [1] model to handle both positive and negative correlations between responses. Section 4 states the mixture model that uses this distribution as its component distribution for each cluster. Section 5 gives the details of the maximum likelihood estimation for the mixture model. Section 6 shows how the fitted model can be used for clustering and also discusses the choice of the number of clusters. Section 7 presents simulation results comparing the proposed method with the traditional LCA method. Section 8 gives two real data examples to illustrate the application of our proposed method to practical situations. Some concluding remarks and directions for future research are outlined in Section 9. Proofs of theoretical results and computational details of the algorithm are given in the Appendix.

2. LITERATURE REVIEW

Qu, Tan and Kutner [10] proposed a latent class mixture model with random effects to handle local dependencies. In this model the component distribution for each cluster was assumed to be multivariate probit (MVP). Specifically, let $X_1, X_2, \ldots, X_m$ denote the manifest Bernoulli responses and let $Y_1, Y_2, \ldots, Y_m$ denote the latent multivariate standard normal random variables (r.v.'s) with $X_i = I(Y_i \leq z(\theta_i))$, where $I(\cdot)$ is an indicator function, i.e. $z(\theta_i)$ is a threshold on Y_i such that $X_i = 1$ if $Y_i \leq z(\theta_i)$ and $X_i = 0$ if $Y_i > z(\theta_i)$. Here $z(\theta_i)$ is the θ_i-quantile of the standard normal distribution so that X_i is Bernoulli with success probability θ_i (denoted by $X_i \sim B(\theta_i)$). The MVP model of Qu *et al.* [10] effectively assumes a product correlation structure for the Y_i's, i.e. $\text{corr}(Y_i, Y_j) = \tau_{ij} = \gamma_i \gamma_j$ where $-1 < \gamma_i < 1$ for all i. It is easy to show that, since $\tau_{ij} > 0$ if both γ_i and γ_j have the same sign and $\tau_{ij} < 0$ otherwise, the X_i's fall into two groups such that the X_i's belonging to the same group are positively correlated and the X_i's belonging to the different groups are negatively correlated. Many real data sets do not have this type of block correlation structure. In addition, the parameter estimation for the MVP model involves evaluations of integrals

by quadrature or Monte Carlo methods. Repeated evaluations of the integrals required for maximum likelihood estimation of the unknown parameters are computationally intensive, especially if the number of variables is large. Emrich and Piedmonte [11] proposed an MVP model for generating multivariate Bernoulli data with a general correlation matrix, but their model is even more computationally intensive not only because there are many more parameters to estimate, but also because multivariate integrals with arbitrary correlations cannot be reduced to univariate integrals as they can be in the product correlation case; see Ref. [12] (p. 374).

An alternative approach is to employ a log-linear formulation of the latent class model [13]. This approach augments the traditional latent class local independence model, which includes marginal probabilities of responses conditional on cluster membership, by incorporating joint response probabilities (also conditional on cluster identity) for selected pairs of items. These joint response probabilities can be modeled either using direct effects ([14,15]) between selected pairs of items or using additional latent variables, called Dfactors, that are associated with those pairs of items. The decision about which two-way interactions to include in the model is based on two-way residual analyses of the manifest variables after fitting the local independence model; more interactions are added iteratively until all residual correlations decrease to an acceptable level.The parameter estimation process is relatively fast and can accommodate many more variables than our proposed method can. Therefore, this method is well-suited for data mining (although requirement of more user input for adding interactions could make the method less automatic) and has been implemented in a commercial software called Latent Gold [16].

3. A MULTIVARIATE BERNOULLI DISTRIBUTION MODEL

Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_m)$ denote a vector of correlated Bernoulli random variables (r.v. 's) on a subject. Marginally each $X_i \sim B(\theta_i)$. We assume that $0 < \theta_i < 1$ for all i. Prentice [17] showed that, due to the binary nature of the X_i's, the correlation coefficient $\rho_{ij} = \text{corr}(X_i, X_j)$ has a limited range , $-\rho_{ij}^* \le \rho_{ij} \le +\rho_{ij}^{**}$, where

$$\rho_{ij}^* = \min\left[\sqrt{\frac{\theta_i\theta_j}{(1-\theta_i)(1-\theta_j)}}, \sqrt{\frac{(1-\theta_i)(1-\theta_j)}{\theta_i\theta_j}}\right] \quad (1)$$

and

$$\rho_{ij}^{**} = \min\left[\sqrt{\frac{\theta_i(1-\theta_j)}{\theta_j(1-\theta_i)}}, \sqrt{\frac{\theta_j(1-\theta_i)}{\theta_i(1-\theta_j)}}\right]. \quad (2)$$

Because of the limited range of ρ_{ij} it will be useful to define the *relative correlation coefficient*, $-1 \le r_{ij} \le 1$, as follows:

$$r_{ij} = \begin{cases} \rho_{ij}/\rho_{ij}^* & \text{if } \rho_{ij} \le 0 \\ \rho_{ij}/\rho_{ij}^{**} & \text{if } \rho_{ij} > 0. \end{cases} \quad (3)$$

Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)$ be a realization of the random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_m)$. Then $\boldsymbol{x}$ can be characterized by its *pattern* of 1's and 0's. We denote a pattern by $P \subseteq M = \{1, 2, \ldots, m\}$, where $x_i = 1\ \forall\ i \in P$ and $x_i = 0\ \forall\ i \notin P$. Each pattern P has a unique *index* p defined by:

$$p = 1 + \sum_{i=1}^{m} 2^{i-1} x_i,$$

where p ranges from 1 (when all $x_i = 0$) to 2^m (when all $x_i = 1$).

Various distributional models have been proposed for correlated binary variables. A standard way to induce correlation is through a common latent variable. The MVP model by Qu *et al.* [10], mentioned in the previous section, uses a $N(0, 1)$ latent variable. Al-Osh and Lee [18] use a discrete (Bernoulli) latent variable, which makes their model rather restrictive. A more flexible model was proposed by Oman and Zucker [1] using an arbitrary *continuous* latent variable. We adopt their model (referred to as the *CLV model*) after extending it to allow for negative correlations as follows.

Let $Z_0, Z_1, \ldots, Z_m$ be i.i.d. continuous r.v.'s with a common known distribution. For convenience and without loss of generality, we will assume that the common known distribution is uniform over $[0, 1]$ (denoted as $U[0, 1]$). Let $V_1, V_2, \ldots, V_m$ be independent Bernoulli r.v.'s with parameters $\beta_1, \beta_2, \ldots, \beta_m$, respectively, and let

$$U_i = V_i Z_0 + (1 - V_i) Z_i. \quad (4)$$

Here Z_0 may be thought of as a subject-specific latent variable common to all items and the Z_i $(1 \le i \le m)$ as item-specific latent variables. Local dependencies between the responses are induced by Z_0. The U_i are positively correlated $U[0, 1]$ r.v.'s. To allow for negative correlations, we introduce independent Bernoulli r.v.'s $W_i \sim B(\gamma_i)$, which allow either not flipping or flipping the sign of U_i (keeping its magnitude the same) with probabilities γ_i and $1 - \gamma_i$, respectively. Let

$$Y_i = U_i W_i + (1 - U_i)(1 - W_i) \text{ and } X_i = I\,(Y_i \le \theta_i)\,, \quad (5)$$

where $I(\cdot)$ is an indicator function. It is easy to see that $Y_i \sim U[0, 1]$ and hence $\Pr(X_i = 1) = \theta_i$. Furthermore,

using straightforward calculations the following expression for ρ_{ij} can be derived:

$$\rho_{ij} = \begin{cases} \beta_i\beta_j\rho_{ij}^{**}[1 - \{\gamma_i(1-\gamma_j) \\ \quad +\gamma_j(1-\gamma_i)\}/\max(\theta_i, 1-\theta_j)] & \text{if } \theta_i \le \theta_j \\ \beta_i\beta_j\rho_{ij}^{**}[1 - \{\gamma_i(1-\gamma_j) \\ \quad +\gamma_j(1-\gamma_i)\}/\max(\theta_j, 1-\theta_i)] & \text{if } \theta_i \ge \theta_j. \end{cases} \tag{6}$$

Thus the sign of ρ_{ij} depends on whether the second term inside the square bracket is > 1 or < 1. In particular, $\rho_{ij} \ge 0$ if and only if

$$\gamma_i(1-\gamma_j) + \gamma_j(1-\gamma_i) \le \begin{cases} 1-\theta_j & \text{if } \theta_i \le \theta_j, \theta_i+\theta_j \le 1 \quad \text{(Region I)} \\ \theta_i & \text{if } \theta_i \le \theta_j, \theta_i+\theta_j \ge 1 \quad \text{(Region II)} \\ \theta_j & \text{if } \theta_j \le \theta_i, \theta_i+\theta_j \ge 1 \quad \text{(Region III)} \\ 1-\theta_i & \text{if } \theta_j \le \theta_i, \theta_i+\theta_j \le 1 \quad \text{(Region IV)} \end{cases}$$

where the four regions are shown in Fig. 1. As the γ's and the θ's can be chosen independently of each other, one can see that this model does not impose a restrictive structure on the correlation matrix.

Although the model defined in Eqs. 4 and 5 is simple, the joint distribution of $\boldsymbol{X} = (X_1, X_2, \ldots, X_m)$ is rather complicated and is given by (the derivation is given in the Appendix):

$$f(p|\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{A\subseteq P}\sum_{B\subseteq Q}\sum_{C\subseteq A\cup B}[\theta^{**}(A,B,C) - \theta^{*}(A,B,C)]^{+} \times \left[\prod_{i\in P\setminus A}\theta_i \prod_{i\in Q\setminus B}(1-\theta_i)\right] \times \left[\prod_{i\in A,B}\beta_i \prod_{i\in P\setminus A, Q\setminus B}(1-\beta_i)\right] \times \left[\prod_{i\in C}\gamma_i \prod_{i\in (A\cup B)\setminus C}(1-\gamma_i)\right], \tag{7}$$

where

$$\theta^{*}(A,B,C) = \max\left\{0, \max_{i\in B\cap C}\theta_i, \max_{i\in A\cap D}(1-\theta_i)\right\}$$

and

$$\theta^{**}(A,B,C) = \min\left\{1, \min_{i\in A\cap C}\theta_i, \min_{i\in B\cap D}(1-\theta_i)\right\},$$

where the sets A, B, C, D are defined in the Appendix.

Note that if all $\beta_i = 0$ then we get the *independence model*: $f(p) = \prod_{i\in P}\theta_i \prod_{i\in Q}(1-\theta_i)$.

The general form of the CLV model has $3m$ parameters. A model with so many parameters is difficult to fit, especially when a separate model must be fitted to each cluster, as we learnt through computational experience. So we decided to make one of the following simplifying assumptions: (i) all $\beta_i \equiv \beta$ but the γ_i are unrestricted (CLV1 Model) or (ii) all $\gamma_i \equiv \gamma$ but the β_i are unrestricted (CLV2 Model). We chose the CLV1 model as it allows a wider range of correlations to be modeled as shown in the following proposition.

PROPOSITION 1: Denote ρ_{ij} for the CLV1 model by $\rho_{ij}^{(1)}$ and that for the CLV2 model by $\rho_{ij}^{(2)}$. Then for fixed θ_i, θ_j, the range of $\rho_{ij}^{(1)}$ is the entire feasible range $[-\rho_{ij}^{*}, \rho_{ij}^{**}]$, whereas the range of $\rho_{ij}^{(2)}$ is $[(1/2)(\rho_{ij}^{**} - \rho_{ij}^{*}), \rho_{ij}^{**}]$, which is only half as wide. If both θ_i, θ_j are either $< 1/2$ or $> 1/2$ then

$$\min \rho_{ij}^{(2)} = \frac{1}{2}(\rho_{ij}^{**} - \rho_{ij}^{*}) > 0,$$

so negative ρ_{ij} cannot be modeled under the CLV2 model in this case.

Proof: See the Appendix. ■

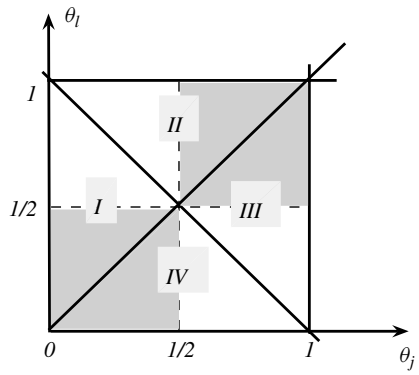

Fig. 1 Shaded regions of the (θ_i, θ_j) space where $\min \rho_{ij}^{(2)} > 0$.

4. MIXTURE MODEL

Consider N subjects on each of whom $m \ge 2$ binary responses are measured. Let $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{im})$ be the vector of responses on the ith subject with pattern index p_i $(1 \le p_i \le 2^m)$ and let $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{im})$ be the

corresponding vector. We want to classify the subjects into $K \geq 1$ disjoint homogeneous clusters, $C_1, C_2, \ldots, C_K$, where provisionally we fix K and assume it to be known (note $K = 1$ means no clusters). For cluster C_k, denote the vector of Bernoulli probabilities by $\boldsymbol{\theta}_k = (\theta_{1k}, \theta_{2k}, \ldots, \theta_{mk})$. Thus $\Pr(X_{ij} = 1 | i \in C_k) = \theta_{jk}$ and $\Pr(X_{ij} = 0 | i \in C_k) = 1 - \theta_{jk}$ (denoted as $X_{ij} \sim B(\theta_{jk})$ conditionally on $i \in C_k$) for $j = 1, 2, \ldots, m$. Under the independence model we have

$$f(p_i|\boldsymbol{\theta}_k) = \prod_{j=1}^{m} \theta_{jk}^{x_{ij}} (1 - \theta_{jk})^{1 - x_{ij}}. \tag{8}$$

In order to account for local dependence, we replace $f(p_i|\boldsymbol{\theta}_k)$ for the independence case by the distribution given by Eq. 7. Here that distribution depends, in addition to $\boldsymbol{\theta}_k$, also on $\boldsymbol{\beta}_k = (\beta_{1k}, \ldots, \beta_{mk})$ and $\boldsymbol{\gamma}_k = (\gamma_{1k}, \ldots, \gamma_{mk})$. For the CLV1 model, $\boldsymbol{\beta}_k$ reduces to a scalar quantity β_k. In that case we denote it by $f(p_i|\boldsymbol{\theta}_k, \beta_k, \boldsymbol{\gamma}_k)$.

Let $\eta_k = \Pr(i \in C_k)$ be the *prior probability* of a randomly chosen subject i belonging to cluster C_k where $\sum_{k=1}^{K} \eta_k = 1$. The η_k are also referred to as *mixing proportions*. Fitting the proposed model involves finding the maximum likelihood estimates (MLEs) of the η_k and $(\boldsymbol{\theta}_k, \beta_k, \boldsymbol{\gamma}_k)$ for $k = 1, 2, \ldots, K$ by maximizing the log-likelihood function

$$\ln L = \sum_{i=1}^{N} \ln \left[\sum_{k=1}^{K} \eta_k f(p_i|\boldsymbol{\theta}_k, \beta_k, \boldsymbol{\gamma}_k) \right].$$

By utilizing the fact that the contribution to the log-likelihood from all subjects with the same pattern index $p_i = p$ is the same and hence can be multiplied by the number of subjects, n_p, having that pattern index ($\sum_{p=1}^{2^m} n_p = N$), the summation over N subjects (which can be quite large) in the above expression is reduced to a summation over 2^m pattern indexes (which can be relatively small) as follows:

$$\ln L = \sum_{p=1}^{2^m} n_p \ln \left[\sum_{k=1}^{K} \eta_k f(p|\boldsymbol{\theta}_k, \beta_k, \boldsymbol{\gamma}_k) \right]. \tag{9}$$

5. MAXIMUM LIKELIHOOD ESTIMATION OF THE MIXTURE MODEL

The expectation-maximization (EM) algorithm of Dempster, Laird and Rubin [19] can be used to compute the MLEs of θ_{jk} and η_k in the LCA model where, at each iteration, closed formulas are available for the current MLEs of the parameters; see [9](pp. 138–139). For the CLV1 model, the log-likelihood function (Eq. 9) is more complex and similar closed formulas for the MLEs do not exist. In fact, from Eq. 7 we see that Eq. 9 is not even differentiable in $\boldsymbol{\theta}$ as it involves the min and max operations on the θ_{jk}. Therefore we used the nonlinear programming (NLP) approach to compute the MLEs of the parameters all of which were constrained to be between 0 and 1. The computational details are given in the Appendix.

6. CLUSTERING PROBLEM

6.1. Classification of Observations

The MLEs, $\widehat{\eta}_k$, $\widehat{\boldsymbol{\theta}}_k$, $\widehat{\beta}_k$, $\widehat{\boldsymbol{\gamma}}_k$, can be used to compute the (estimated) posterior probabilities of the observations belonging to different clusters (also called *responsibilities*). Note that all observations having the same index p have the same posterior probability given by:

$$\widehat{\eta}_k(p) = \frac{\widehat{\eta}_k f(p|\widehat{\boldsymbol{\theta}}_k, \widehat{\beta}_k, \widehat{\boldsymbol{\gamma}}_k)}{\sum_{\ell=1}^{K} \widehat{\eta}_\ell f(p|\widehat{\boldsymbol{\theta}}_k, \widehat{\beta}_k, \widehat{\boldsymbol{\gamma}}_k)}. \tag{10}$$

The Bayes (maximum posterior) rule is used to assign observations to clusters. Specifically, all observations with pattern index p are assigned to that cluster C_k which gives the maximum $\widehat{\eta}_k(p)$. This is equivalent to partitioning the space of 2^m patterns into K partitions such that all patterns in the same partition are assigned to the same cluster. Hastie, Tibshirani and Friedman [20] refer to this as *hard assignment* as opposed to *soft assignment* in which the observations are assigned to different clusters in proportion to their posterior probabilities.

6.2. Number of Clusters

The CLV1 model has $2m + 1$ unknown parameters per cluster (m each of the θ_{jk}'s and γ_{jk}'s and one β_k). Thus for K clusters there are $(2m + 1)K$ unknown parameters. In addition, there are $K - 1$ independent prior probabilities, η_k's. Thus there are $n = 2(m + 1)K - 1$ unknown parameters. In comparison, there are $n = (m + 1)K - 1$ unknown parameters in the LCA model (m each of the θ_{jk}'s and $K - 1$ η_k's).

Effectively, the data in this problem are the pattern frequencies, n_p. There are $2^m - 1$ independent n_p's as they sum to the fixed total sample size N. In order for the model to be estimable we must have the number of parameters to be no more than the number of independent data values, i.e.

$$2(m + 1)K - 1 \leq 2^m - 1 \Longleftrightarrow K \leq K_{\max} = \left\lfloor \frac{2^{m-1}}{m + 1} \right\rfloor, \tag{11}$$

where $K_{\max}$ denotes the maximum number of clusters that can be fitted and $\lfloor x \rfloor$ denotes the integer part of x. The following table gives the $K_{\max}$ values for selected values of m.

m	3	4	5	6	7	8	9	10
$K_{\max}$	1	1	2	4	8	14	25	46

In most applications, $K_{\max} \leq 4$, so $m \geq 6$ is sufficient for clustering purposes.

Determination of the number of clusters is a special case of the model selection problem. Many criteria have been proposed in the literature for model selection. We use Schwarz's [2] Bayesian information criterion (BIC), which is defined as:

$$\text{BIC} = 2 \ln L - n \ln N,$$

where $\ln L$ is the maximized log-likelihood function (Eq.9) with a given number of clusters, $n = 2(m+1)K - 1$ is the total number of parameters and N is the total sample size. The goal is to choose the model that maximizes BIC. This criterion is selected because it takes into account the effect of the sample size in its penalty function and is consistent in the sense that if the true model is among the candidates then the probability of selecting the true model approaches 1 as $N \to \infty$ as shown by Keribin [21]. Note, however, that regardless of which criterion is used, the clusters must be interpretable in the context of the problem. An interpretable solution is preferable to an optimal solution when determining the number of clusters.

7. SIMULATION STUDY

In this section, we compare the performance of the proposed method with the classical LCA method which uses the independence model. We also assess the robustness of the proposed method by generating data using a model different from the CLV1 model.

7.1. Performance Measures

The main performance measure is the *CCR*, which is the proportion of observations that are classified to the correct cluster. The misclassification rate (MCR) equals $1 - \text{CCR}$. For binary data there are lower and upper bounds on CCR (denoted by LCCR and UCCR, respectively) because of the fact that any hard assignment rule classifies each pattern to exactly one cluster. So all observations with that pattern which belong to other clusters are misclassified. One must also remember that cluster labels are arbitrarily assigned; what matters is that the observations belonging to the same cluster are classified together. Therefore CCR must be computed by taking the maximum over all possible cluster labelings.

As an example, suppose that there are 50 observations from each of two clusters which are classified as shown in Table 1. It would appear that CCR is $(15 + 25)/100 = 40\%$. However, if we switch the labels of classified clusters then we see that CCR is $(35 + 25)/100 = 60\%$. This suggests that for two clusters, CCR ≥ 0.5. The maximum achievable upper bound on CCR is found by simply assigning each pattern p to that cluster C_k in which it has the highest frequency. The following proposition gives general lower and upper bounds on CCR.

PROPOSITION 2: Let n_{pk} denote the true (unknown) count of observations having pattern p that come from cluster C_k ($\sum_{k=1}^{K} n_{pk} = n_p$). Then the lower and upper bounds on CCR are given by

$$\text{LCCR} = \frac{1}{K} \text{ and } \text{UCCR} = \frac{\sum_{p=1}^{2^m} \max_k n_{pk}}{N}. \tag{12}$$

Proof: See the Appendix.

Because of the bounds on CCR, it is convenient to use a standardized measure, which we call the *correct classification score (CCS)*, defined as

$$\text{CCS} = \frac{\text{CCR} - \text{LCCR}}{\text{UCCR} - \text{LCCR}}. \tag{13}$$

Note that CCS falls between 0 and 1, and large values of CCS are desirable.

To compare the goodness of fit of the CLV1 model relative to that of the LCA model we propose two measures analogous to R^2 and adjusted R^2 (R^2_{adj}) used in multiple regression. Although these measures are not used in the simulation studies, they are used in the two examples in Section 8, and hence are defined here. Let $\widehat{n}_p = \sum_{k=1}^{K} \widehat{n}_{pk}$ be the estimated (fitted) frequency for pattern p by a particular model where $\widehat{n}_{pk}$ is computed by multiplying the total sample size N by the estimated cluster probability $\widehat{\eta}_k$ and the estimated probability of pattern p conditioned

Table 1. Classification of data into two clusters.

		Classified to Cluster 1	Classified to Cluster 2	
Belong to	Cluster 1	15	35	50
	Cluster 2	25	25	50
		40	60	

Table 2. Low and high parameter values for the CLV1 model used in the simulation study.

Level	Parameter							
	θ_{jk}	β_k	γ_{1k}	γ_{2k}	γ_{3k}	γ_{4k}	γ_{5k}	η_1
Low	0.4	0.50	0.01	0.01	0.01	0.01	0.01	0.4
High	0.6	0.95	0.50	0.99	0.99	0.99	0.99	0.6

on the event that observation belongs to cluster C_k (these probabilities are given by Eq. 8 for the independence model and Eq. 7 for the CLV1 model). Let SS_{e0} and SS_{e1} be the error sums of squares for the independence and CLV1 models, respectively, where

$$\mathrm{SS}_e = \sum_{p=1}^{2^m} (n_p - \widehat{n}_p)^2$$

using the appropriate model to compute $\widehat{n}_p$. Generally one has $\mathrm{SS}_{e1} \leq \mathrm{SS}_{e0}$ as the CLV1 model has more parameters and the independence model is a special case of the CLV1 model. The error degrees of freedom for the two models are $N - 1 - n = N - K(m+1)$ for the independence model and $N - 1 - n = N - 2K(m+1)$ for the CLV1 model. Then

$$R^2 = 1 - \frac{\mathrm{SS}_{e1}}{\mathrm{SS}_{e0}} \text{ and } R^2_{\mathrm{adj}} = 1 - \frac{\mathrm{SS}_{e1}/[N - 2K(m+1)]}{\mathrm{SS}_{e0}/[N - K(m+1)]}. \tag{14}$$

While R^2 gives the straight percentage reduction in SS_e by the CLV1 model compared to the independence model, R^2_{adj} adjusts this for the number of parameters in each model via the corresponding error degrees of freedom.

7.2. Simulation Results

We conducted a simulation study for $K = 2$ clusters, $m = 5$ responses and $N = 500, 5000$ and 50 000. In each case, data were generated using two models: (i) the CLV1 model and (ii) the MVP model with product correlation structure. Data were also generated using the independence model, which is a special case of both these models. The data from the independence and the MVP models were used to test robustness of the proposed method which assumes the CLV1 model.

The parameters for the CLV1 model were chosen as follows. There are a total of 23 parameters in this study (θ_{jk} and γ_{jk} for $j = 1, \ldots, 5, k = 1, 2$; β_1, β_2 and η_1). We chose two levels (high and low) for each parameter as given in Table 2.

Twenty-four different combinations of these parameter values were obtained by using a 24-run Plackett-Burman (PB) design shown in Table 3. The PB design was chosen because it is based on a two-level $(+/-)$ orthogonal array with the minimum number of runs (rows) equal to one more than the number of columns [[22], p. 317]. The run with low values for all parameters for both clusters was replaced with the independence model by setting β_1, β_2 and all γ_{ij} equal to 0, $\theta_{1j} = 0.40, \theta_{2j} = 0.60$ $(1 \leq j \leq 5)$ and $\eta_1 = 0.60$. The runs in Table 3 are arranged according to their values for the weighted average (with weights equal to the cluster mixing proportions) absolute relative correlation (shown in the last column), which is defined as follows:

$$|\overline{r}| = \sum_{k=1}^{K} \eta_k |\overline{r}_k| \text{ where } |\overline{r}_k| = \frac{1}{\binom{m}{2}} \sum_{i<j} |r_{ijk}|,$$

with r_{ijk} being the relative correlation between responses i and j in cluster k. Note that we use $|\overline{r}|$ as a single global measure, but we recognize that no single measure can capture the extent of correlation in the data.

For each run (i.e. each combination of the parameter settings) we performed 20 replications. For each replication both the traditional LCA method and the proposed method were applied. CCR was obtained from which CCS was computed for each method and for each replication. Finally, the CCS values for each method were averaged over 20 replications and their standard deviations, $s/\sqrt{20}$, where s is the sample standard deviation of the CCS values obtained from 20 replications were computed. Table 4 summarizes these results. The results for $N = 5000$ are graphically displayed in Fig. 2. This figure shows the plots of the average CCS values with two standard deviation bars around the average. The plots of the average CCS values for the traditional LCA method are marked with circles, while those for the proposed method are marked with diamonds.

The trends in the CCS results are not very smooth, and the range of variation at different parameter settings is also highly variable for both the proposed method and the LCA method. We believe that this is because the CCS values do not depend on $|\overline{r}|$ alone, but also on the differences between the correlation structures and the θ values of the two clusters. These differences are difficult to quantify in terms of a few simple measures. Nevertheless, there are some general trends as elucidated in the following text.

1. The CCS values for the LCA method show a general decreasing pattern with respect to $|\overline{r}|$. As $|\overline{r}|$ increases, the independence model gives a poorer fit which results in more misclassifications. On the other hand, the average CCS values for the proposed method increase for low values of $|\overline{r}|$ reaching a plateau for medium values of $|\overline{r}|$ and then they decrease for high values of $|\overline{r}|$. Only for the independence case, the proposed method has a significantly lower average CCS value than does the LCA

Table 3. Plackett-Burman design for data generated from the CLV1 model.

Cluster 1											Cluster 2												
θ_1	θ_2	θ_3	θ_4	θ_5	β	γ_1	γ_2	γ_3	γ_4	γ_5	θ_1	θ_2	θ_3	θ_4	θ_5	β	γ_1	γ_2	γ_3	γ_4	γ_5	η_1	$\lvert\bar{r}\rvert$
−	−	−	−	−	0	0	0	0	0	0	+	+	+	+	+	0	0	0	0	0	0	+	0.000
+	+	+	+	+	−	+	−	+	+	−	+	+	−	−	+	−	+	−	−	−	−	−	0.160
−	+	−	−	−	−	+	+	+	+	+	+	−	+	+	−	−	+	+	−	−	+	−	0.161
+	+	+	−	+	−	+	+	−	−	+	−	−	+	−	+	−	−	−	−	+	+	+	0.192
−	+	−	+	+	−	−	+	+	−	−	−	+	−	−	−	−	+	+	+	+	+	+	0.208
+	−	+	−	−	−	−	+	+	+	+	−	+	−	+	+	−	−	+	+	−	−	+	0.240
+	+	−	+	−	+	+	−	−	+	+	−	+	−	+	−	−	−	−	+	+	+	−	0.375
−	−	+	−	+	−	−	−	−	+	+	+	+	−	+	−	+	+	−	−	+	+	+	0.375
−	−	+	+	+	+	+	−	+	−	+	−	−	+	+	−	−	+	−	+	−	−	+	0.412
+	−	+	+	−	−	+	+	−	−	+	+	−	−	−	−	+	+	+	+	+	−	−	0.412
+	+	−	−	+	−	+	−	−	−	−	+	+	+	+	−	+	−	+	+	−	−	+	0.442
−	−	−	+	+	+	+	+	−	+	−	+	−	−	+	+	−	−	+	−	+	−	+	0.442
−	+	+	−	−	+	−	+	−	−	−	+	+	+	+	+	−	+	−	+	+	−	−	0.443
+	−	−	+	+	−	−	+	−	+	−	−	−	+	+	+	+	+	−	+	−	+	−	0.443
−	+	−	+	−	−	−	−	+	+	+	+	−	+	−	+	+	−	−	+	+	−	+	0.489
+	−	+	−	+	+	−	−	+	+	−	+	−	+	−	−	−	−	+	+	+	+	−	0.490
−	+	+	−	−	+	+	−	−	+	−	−	−	−	−	+	+	+	+	+	−	+	+	0.579
+	−	−	−	−	+	+	+	+	+	−	−	+	+	−	−	+	+	−	−	+	−	+	0.579
+	−	−	+	−	+	−	−	−	−	+	+	+	+	−	+	−	+	+	−	−	+	+	0.582
−	−	+	+	−	−	+	−	+	−	−	−	+	+	+	+	+	−	+	−	+	+	−	0.584
+	+	−	−	+	+	−	−	+	−	+	−	−	−	+	+	+	+	+	−	+	−	−	0.692
−	−	−	−	+	+	+	+	+	−	+	+	+	−	−	+	+	−	−	+	−	+	−	0.749
+	+	+	+	−	+	−	+	+	−	−	+	−	−	+	−	+	−	−	−	−	+	+	0.864
−	+	+	+	+	+	−	+	−	+	+	−	+	+	−	−	+	−	+	−	−	−	−	0.867

The symbol + denotes the high level and − denotes the low level; 0 denotes zero values for β_k, γ_{jk}.

method. This is because the proposed method utilizes the information in the correlations and hence results in less misclassifications.

The nonmonotone behavior of the CCS of the proposed method as a function of $|\bar{r}|$ can be explained as follows. The additional information contributed by correlations, as they increase from 0, is utilized by the proposed method thus improving its performance in an absolute sense. As correlations get larger the net amount of information in a fixed number of responses decreases because the responses act as proxies for each other. However, the proposed method is still effective in capturing the correlations; so the performance of the method reaches a plateau. Finally when the correlations get close to 1, the high degree of dependence between the responses means that the effective number of responses becomes less than m. As a result, the CCS values decrease.

2. The standard deviations for the average CCS values for the LCA method are generally much smaller than those for the proposed method. The reason is that the independence model involves fewer parameters (10 θs plus 1 η for a total of 11 parameters instead of 23 parameters in the CLV1 model). As a result, the estimates of these parameters have smaller sampling errors and hence the CCS values have smaller standard deviations.

3. The CCS values for the CLV1 model method increase with the sample size. The independence model method does not exhibit such a monotone behavior with the sample size, and the CCS values appear to fluctuate randomly around a mean value for each parameter configuration.

Simulations for the MVP model were conducted in the same manner as for the CLV1 model. The low and high parameter values for the MVP model are given in Table 5.

In this case there are 21 parameters. We used the first 21 columns of the 24-run PB design shown in Table 6. The run with low values for all parameters for both clusters was replaced with the independence model by setting $\gamma_{ij} = 0$, $\theta_{1j} = 0.40$, $\theta_{2j} = 0.60$ $(1 \le j \le 5)$ and $\eta_1 = 0.60$. This run is identical to the corresponding independence model run for CLV1 data and so was not repeated. The runs in Table 6 are also arranged according to the $|\bar{r}|$ values, which are shown in the last column. Simulation results for the

Table 4. Estimated CCS values and their standard errors[a] for data generated from the CLV1 model.

	$N = 500$		$N = 5000$		$N = 50000$	
$\lvert\bar{r}\rvert$	Proposed method	LCA method	Proposed method	LCA method	Proposed method	LCA method
.000	.420 (.146)	.492 (.243)	.461 (.139)	.867 (.151)	.525 (.251)	.984 (.007)
.160	.301 (.240)	.364 (.219)	.525 (.271)	.375 (.066)	.829 (.131)	.391 (.012)
.161	.365 (.262)	.416 (.143)	.534 (.228)	.449 (.074)	.688 (.329)	.500 (.055)
.192	.379 (.226)	.298 (.192)	.522 (.304)	.349 (.090)	.952 (.107)	.374 (.036)
.208	.402 (.229)	.341 (.117)	.501 (.298)	.404 (.042)	.940 (.100)	.416 (.012)
.240	.419 (.151)	.449 (.107)	.503 (.284)	.459 (.057)	.568 (.421)	.492 (.075)
.375	.498 (.302)	.077 (.059)	.990 (.026)	.027 (.020)	1.000 (.000)	.018 (.011)
.375	.662 (.276)	.192 (.137)	.986 (.022)	.140 (.066)	.999 (.002)	.112 (.016)
.412	.662 (.202)	.064 (.054)	.958 (.069)	.020 1]15 925 (.011)	1.000 (.000)	.015 (.005)
.412	.575 (.268)	.214 (.111)	.916 (.181)	.165 (.024)	.999 (.001)	.155 (.022)
.442	.866 (.167)	.345 (.144)	.998 (.004)	.186 (.096)	1.000 (.000)	.223 (.060)
.442	.615 (.212)	.045 (.048)	.958 (.038)	.026 (.020)	.995 (.010)	.017 (.007)
.443	.417 (.248)	.405 (.177)	.903 (.220)	.379 (.045)	1.000 (.000)	.365 (.017)
.443	.546 (.279)	.100 (.059)	.929 (.084)	.049 (.039)	.998 (.003)	.038 (.015)
.489	.510 (.290)	.147 (.116)	.899 (.176)	.071 (.043)	.814 (.189)	.080 (.004)
.490	.702 (.224)	.184 (.098)	.978 (.090)	.141 (.051)	.920 (.165)	.154 (.008)
.579	.586 (.303)	.489 (.156)	.625 (.292)	.523 (.103)	.759 (.299)	.545 (.015)
.579	.672 (.267)	.218 (.038)	.939 (.068)	.215 (.014)	.986 (.008)	.215 (.006)
.582	.825 (.246)	.159 (.091)	.967 (.119)	.034 (.041)	.881 (.214)	.025 (.007)
.584	.352 (.281)	.239 (.080)	.537 (.425)	.216 (.019)	.919 (.238)	.216 (.005)
.692	.462 (.318)	.268 (.137)	.375 (.325)	.353 (.093)	.724 (.333)	.378 (.009)
.749	.444 (.312)	.051 (.036)	.764 (.197)	.023 (.013)	.854 (.150)	.024 (.004)
.864	.433 (.239)	.268 (.065)	.593 (.260)	.278 (.024)	.564 (.271)	.276 (.007)
.867	.419 (.238)	.053 (.034)	.384 (.354)	.039 (.016)	.682 (.215)	.031 (.006)

[a]The standard errors are given in parentheses.

MVP model are summarized in Table 7. The results for $N = 5000$ are graphically displayed in Fig. 3.

The following interesting results emerge from these simulations.

1. Once again, the CCS values show nonsmooth behavior for the same reasons as explained before. However, there are some general trends as elucidated in the following text.

2. In contrast to the CLV1 data, in this case the CCS values for the proposed method show a random fluctuating pattern with a slightly decreasing trend at high $|\bar{r}|$ values. On the other hand, the CCS values for the LCA method display a generally decreasing trend through the entire range of $|\bar{r}|$ values. Nevertheless, the proposed method does not perform significantly worse than the LCA method for any $|\bar{r}| > 0.368$ barring one exception for $|\bar{r}| = 0.536$.

3. The CCS values using MVP data for the proposed method are uniformly and significantly lower than those for the CLV1 data. Also, in this case, in contrast to the CLV1 data, the performance of the proposed method does not improve with the sample size. This is because the proposed method attempts to fit a wrong model to the data, so having more data does not help improve the fit.

8. EXAMPLES

In this section we analyze two real data sets discussed in the Introduction section. The first data set is from the teaching style study. We chose six binary questions from the survey and attempted to classify the 468 teachers into clusters based on their responses to those six questions. The second data set comes from a research project at the Media Management Center at Northwestern University. Among the many questions asked, we will focus on seven questions that ask the reader if he/she reads a particular newspaper on Monday, Tuesday, . . ., Sunday. Here it is obvious that the responses must be locally dependent.

8.1. Teaching Style Data

We focus attention on the following six questions from the teaching style data.

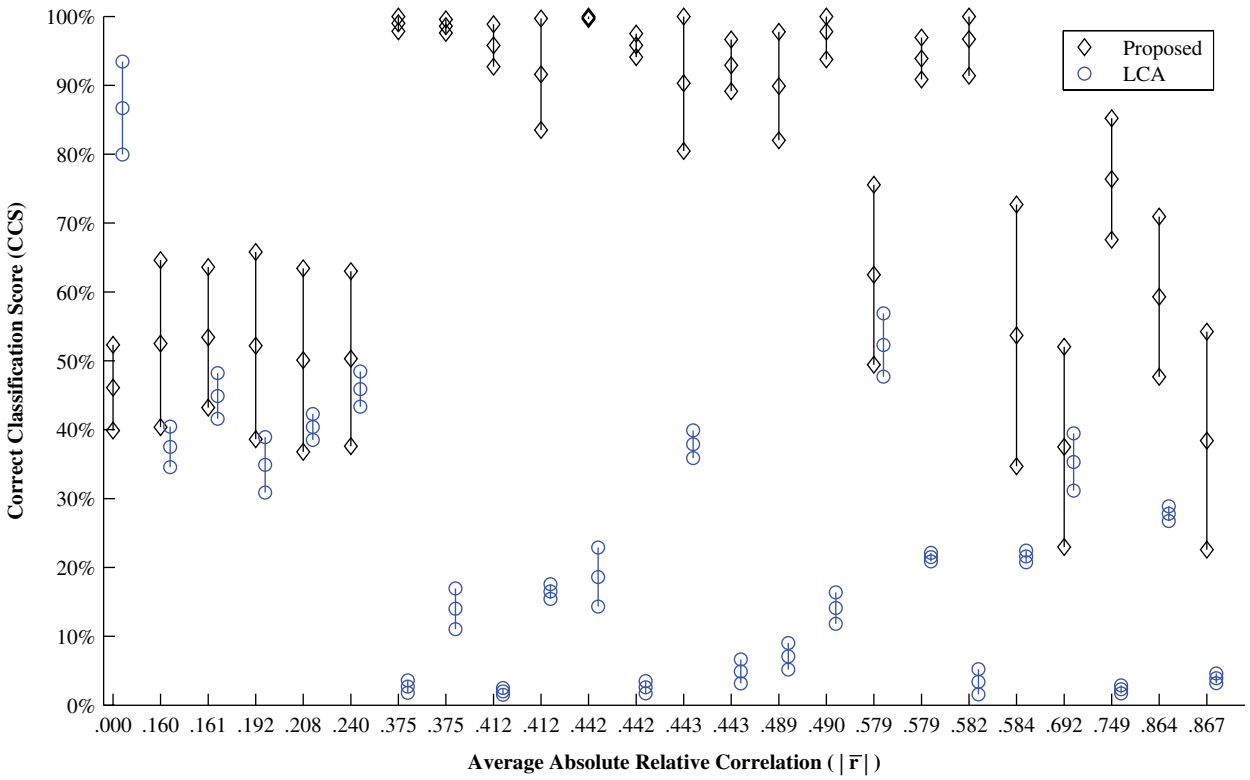

Fig. 2 Average correct classification scores (CCS) using the proposed and the LCA methods [data generated using .the CLV1 model; $N = 5000$)].

Q.1: Pupils not allowed to move around? (Y=1, N=0)

Q.2: Pupils not allowed to talk? (Y=1, N=0)

Q.3: Pupils expected to be quiet? (Y=1, N=0)

Q.4: Explore concepts (1) or develop numerical skills (0)?

Q.5: Emphasis on separate subject teaching? (Y=1, N=0)

Q.6: Emphasis on integrated teaching? (Y=1, N=0)

The proposed mixture model method with the CLV1 distribution as well as the LCA method were applied to these data. The BIC values for $K = 1(1)4$ clusters for the proposed method are shown in Table 8. We see that BIC is maximized for $K = 2$. Hence we selected a two-cluster model. The classification performance of this model was compared with that of the LCA method with two clusters.

The estimates of the θ_k for the two clusters are shown in Table 9. The relative correlation (see Eq. 3) matrices estimated using the proposed method are shown in Table 10.

First, we note that the estimates of the θ_k's obtained by the two methods are similar, but the differences between the two clusters are more evident for the LCA method. Further note that for Cluster 1, $\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_3$ and $\widehat{\theta}_5$ are higher than those for Cluster 2, while the inequality is reversed for $\widehat{\theta}_4$ and $\widehat{\theta}_6$. We see that yes responses to Q.1, Q.2, Q.3 and Q.5 are typical of traditional and disciplinarian teachers, while yes responses to Q.4 and Q.6 are typical of modern and lenient teachers. Thus, both the estimation methods classify teachers into strict and lenient clusters with about 62% in Cluster 1 and 38% in Cluster 2. Although both methods give similar percentages in the two clusters, in fact, 101 out of a total 468 teachers (21.6%) were differentially classified

Table 5. Low and high parameter values for the MVP model used in the simulation study.

	Parameter											
Level	θ_{jk}	γ_{11}	γ_{12}	γ_{13}	γ_{14}	γ_{15}	γ_{21}	γ_{22}	γ_{23}	γ_{24}	γ_{25}	η_1
Low	0.40	0.60	−0.95	0.60	−0.95	0.60	−0.95	0.60	−0.95	0.60	−0.95	0.40
High	0.60	0.95	−0.60	0.95	−0.60	0.95	−0.60	0.95	−0.60	0.95	−0.60	0.60

Table 6. Plackett-Burman design setting for data generated from the MVP model.

Cluster 1										Cluster 2											
θ_1	θ_2	θ_3	θ_4	θ_5	γ_1	γ_2	γ_3	γ_4	γ_5	θ_1	θ_2	θ_3	θ_4	θ_5	γ_1	γ_2	γ_3	γ_4	γ_5	η_1	$\lvert\overline{r}\rvert$
−	−	−	−	−	0	0	0	0	0	+	+	+	+	+	0	0	0	0	0	+	0.000
+	+	+	+	+	−	+	−	+	+	−	+	+	−	−	+	−	+	−	−	−	0.368
−	−	+	+	+	+	+	−	+	−	+	−	−	+	+	−	−	+	−	+	+	0.371
−	+	−	+	−	−	−	−	+	+	+	+	−	+	−	+	+	−	−	+	+	0.392
−	+	−	+	+	−	−	+	+	−	+	−	+	−	−	−	−	+	+	+	−	0.425
+	+	+	−	+	−	+	+	−	−	+	−	−	+	−	+	−	−	−	−	+	0.438
−	+	+	−	−	+	−	+	−	−	−	+	+	+	+	+	−	+	−	+	−	0.458
+	−	+	−	+	+	−	−	+	+	−	+	−	+	−	−	−	−	+	+	−	0.467
−	−	−	−	+	+	+	+	+	−	−	+	+	−	−	+	+	−	−	+	+	0.467
−	+	−	−	−	−	+	+	+	+	−	+	−	+	+	−	−	+	+	−	+	0.472
+	−	+	−	−	−	−	+	+	+	+	−	+	−	+	+	−	−	+	+	+	0.474
+	−	+	+	−	−	+	+	−	−	−	+	−	−	−	−	+	+	+	+	+	0.495
+	+	−	+	−	+	+	−	−	+	−	−	+	−	+	−	−	−	−	+	+	0.502
−	+	+	−	−	+	+	−	−	+	+	−	−	−	−	+	+	+	+	+	−	0.503
+	+	−	−	+	+	−	−	+	−	−	−	−	−	+	+	+	+	+	−	+	0.504
+	−	−	+	+	−	−	+	−	+	−	−	−	+	+	+	+	+	−	+	−	0.504
+	−	−	+	−	+	−	−	−	−	+	+	+	+	−	+	−	+	+	−	+	0.529
−	−	+	−	+	−	−	−	−	+	+	+	+	−	+	−	+	+	−	−	+	0.534
−	−	+	+	−	−	+	−	+	−	−	−	+	+	+	+	+	−	+	−	−	0.536
+	+	−	−	+	−	+	−	−	−	+	+	+	+	+	−	+	−	+	+	−	0.568
+	−	−	−	−	+	+	+	+	+	+	−	+	+	−	−	+	+	−	−	−	0.573
−	−	−	+	+	+	+	+	−	+	+	+	−	−	+	+	−	−	+	−	−	0.621
+	+	+	+	−	+	−	+	+	−	+	+	−	−	+	−	+	−	−	−	−	0.651
−	+	+	+	+	+	−	+	−	+	−	−	+	+	−	−	+	−	+	−	+	0.844

The symbol + denotes the high level; − denotes the low level; 0 denotes zero values for γ_{jk}. xs

by the two methods. Thus, in terms of classification performance, the two methods are significantly different for this data set. Of course, there is no way to tell which method classifies the teachers more accurately.

To compare the goodness of fit of the CLV1 model with the independence model we computed error sums of squares for the two models which were $\mathrm{SS}_{e1} = 127.900$ and $\mathrm{SS}_{e0} = 2781.708$. The error degrees of freedom for the CLV1 model are $2^6 - 2 \times 2 \times (6 + 1) = 36$ and for the independence model are $2^6 - 2 \times (6 + 1) = 50$. Therefore we get

$$R^2 = 1 - \frac{127.9}{2781.708} = 95.40\% \text{ and}$$

$$R^2_{\text{adj}} = 1 - \frac{127.9/36}{2781.708/50} = 93.61\%.$$

Thus the CLV1 model gives a much better fit.

Inspecting the estimated relative correlation matrices we see that, as expected, responses to Q.1, Q.2 and Q.3 are positively correlated with higher relative correlations in Cluster 1 than in Cluster 2. Surprisingly responses to Q.1, Q.2 and Q.3 are negatively correlated with the responses to Q.5 in Cluster 1, but positively correlated in Cluster 2. Finally, responses to Q.6 are negatively correlated with the responses to other questions except Q.4 in both clusters. The negative relative correlation with the responses to Q.5 is especially large (−0.899) in Cluster 2 as teachers who emphasize separate subject teaching are not likely to emphasize integrated teaching.

8.2. Newspaper Reading Survey

The seven questions in the newspaper reading data are Q.i: Do you read (a particular) newspaper on day i? where $i = 1$ for Monday, . . . , $i = 7$ for Sunday. This data set consists of 10 858 responses to a mail survey conducted by the Newspaper Association of America.

The BIC values for $K = 1(1)4$ clusters for the proposed method are shown in Table 11. We see that BIC is maximized for $K = 3$. However, the three-cluster solution was found to be not as readily interpretable as the two-cluster solution (the results are not reported here for the lack of space but are available from the second author). First, one of the clusters had a very low prior probability, and it appeared to be a combination of the other two dominant clusters. Second, the estimated correlation matrices using the CLV1 model also were not interpretable. On the other hand, the two-cluster solution had a nice interpretation, as discussed in the following text, and so was adopted. The classification

Table 7. Estimated CCS values and their standard errors[a] for data generated from the MVP model.

	$N = 500$		$N = 5000$		$N = 50000$	
$\|\bar{r}\|$	Proposed method	LCA method	Proposed method	LCA method	Proposed method	LCA method
.158	.377 (.216)	.804 (.098)	.563 (.301)	.990 (.013)	.736 (.296)	1.000 (.000)
.368	.292 (.156)	.379 (.165)	.310 (.144)	.454 (.084)	.359 (.164)	.486 (.016)
.371	.244 (.171)	.299 (.145)	.284 (.186)	.325 (.091)	.275 (.059)	.330 (.037)
.392	.305 (.193)	.175 (.093)	.243 (.141)	.148 (.045)	.190 (.105)	.145 (.012)
.425	.333 (.225)	.361 (.072)	.324 (.138)	.371 (.030)	.372 (.009)	.383 (.008)
.438	.239 (.171)	.402 (.097)	.281 (.180)	.448 (.038)	.432 (.135)	.451 (.011)
.458	.324 (.179)	.102 (.065)	.384 (.244)	.045 (.031)	.232 (.208)	.054 (.010)
.467	.335 (.221)	.437 (.081)	.426 (.219)	.453 (.022)	.375 (.082)	.461 (.011)
.467	.230 (.153)	.182 (.103)	.218 (.117)	.182 (.046)	.231 (.015)	.190 (.015)
.472	.342 (.224)	.200 (.125)	.272 (.094)	.212 (.051)	.247 (.020)	.215 (.020)
.474	.175 (.145)	.079 (.071)	.099 (.103)	.035 (.028)	.132 (.179)	.030 (.011)
.495	.184 (.192)	.159 (.077)	.122 (.095)	.157 (.025)	.108 (.048)	.153 (.006)
.503	.307 (.124)	.237 (.105)	.235 (.115)	.244 (.027)	.222 (.047)	.234 (.008)
.503	.240 (.132)	.221 (.091)	.198 (.054)	.214 (.032)	.201 (.011)	.218 (.010)
.504	.340 (.180)	.145 (.109)	.319 (.084)	.109 (.043)	.260 (.031)	.108 (.013)
.505	.201 (.151)	.139 (.103)	.070 (.053)	.108 (.046)	.050 (.029)	.114 (.012)
.529	.195 (.156)	.099 (.072)	.147 (.088)	.082 (.032)	.137 (.012)	.077 (.013)
.534	.189 (.141)	.171 (.075)	.253 (.157)	.179 (.042)	.131 (.047)	.178 (.014)
.536	.287 (.250)	.143 (.084)	.101 (.057)	.160 (.026)	.102 (.009)	.166 (.008)
.568	.230 (.145)	.107 (.067)	.268 (.063)	.046 (.029)	.278 (.007)	.027 (.011)
.573	.182 (.121)	.114 (.082)	.117 (.044)	.079 (.037)	.130 (.026)	.073 (.010)
.621	.246 (.157)	.195 (.114)	.268 (.087)	.174 (.049)	.304 (.059)	.170 (.013)
.651	.200 (.141)	.091 (.072)	.179 (.086)	.070 (.039)	.180 (.052)	.053 (.014)
.844	.184 (.157)	.112 (.068)	.168 (.085)	.031 (.025)	.101 (.086)	.035 (.013)

[a]The standard errors are given in parentheses.

performance of this model was compared with that of the LCA method with two clusters. The marginal probability estimates are shown in Table 12. The relative correlation matrices estimated using the proposed method are shown in Table 13.

The results for the two methods are again similar in this case. Cluster 1 marginal probabilities are low for weekdays, but spike to very high values (0.956 using the proposed method and 0.856 using the LCA method) on Day 7 (Sunday). This pattern is consistent with the reading behavior of non-subscribers who tend to purchase the newspaper on weekends, especially on Sundays. On the other hand, both methods give consistently high marginal probabilities for all seven days for cluster 2 (close to 0.9 using the proposed method and close to 1 using the LCA method). This pattern is consistent with the reading behavior of subscribers, who tend to read the newspaper every day. Thus the two clusters can be identified as non-subscribers and subscribers. The percentage of non-subscribers is estimated to be 46% using the CLV1 model and 51% using the LCA method. Although the percentages are somewhat different for the two models, only 690 out of a total 10 858 survey respondents (6.35%) were differentially classified.

To compare the goodness of fit of the CLV1 model with the independence model, we computed error sums of squares for the two models which were $SS_{e1} = 15925.36$ and $SS_{e0} = 577169.5$. The error degrees of freedom for the CLV1 model are $2^7 - 2 \times 2 \times (7 + 1) = 96$ and for the independence model are $2^7 - 2 \times (7 + 1) = 112$. Therefore we get

$$R^2 = 1 - \frac{15925.36}{577169.5} = 97.74\% \text{ and}$$

$$R^2{}_{\text{adj}} = 1 - \frac{15925.36/96}{577169.5/112} = 96.78\%.$$

Thus the CLV1 model gives an even better fit to these data.

Looking at the relative correlation matrices in Table 13, we see that according to the proposed method, the newspaper reading responses over all days of the week are highly correlated for the subscriber group, but for the non-subscriber group, relative correlations are much smaller for weekdays. Especially note that the Sunday response is negatively correlated with all other weekdays. This makes sense as non-subscribers generally do not read the newspaper on weekdays, but often purchase and read it on Sundays. This insight into the relative correlation structure of the data for

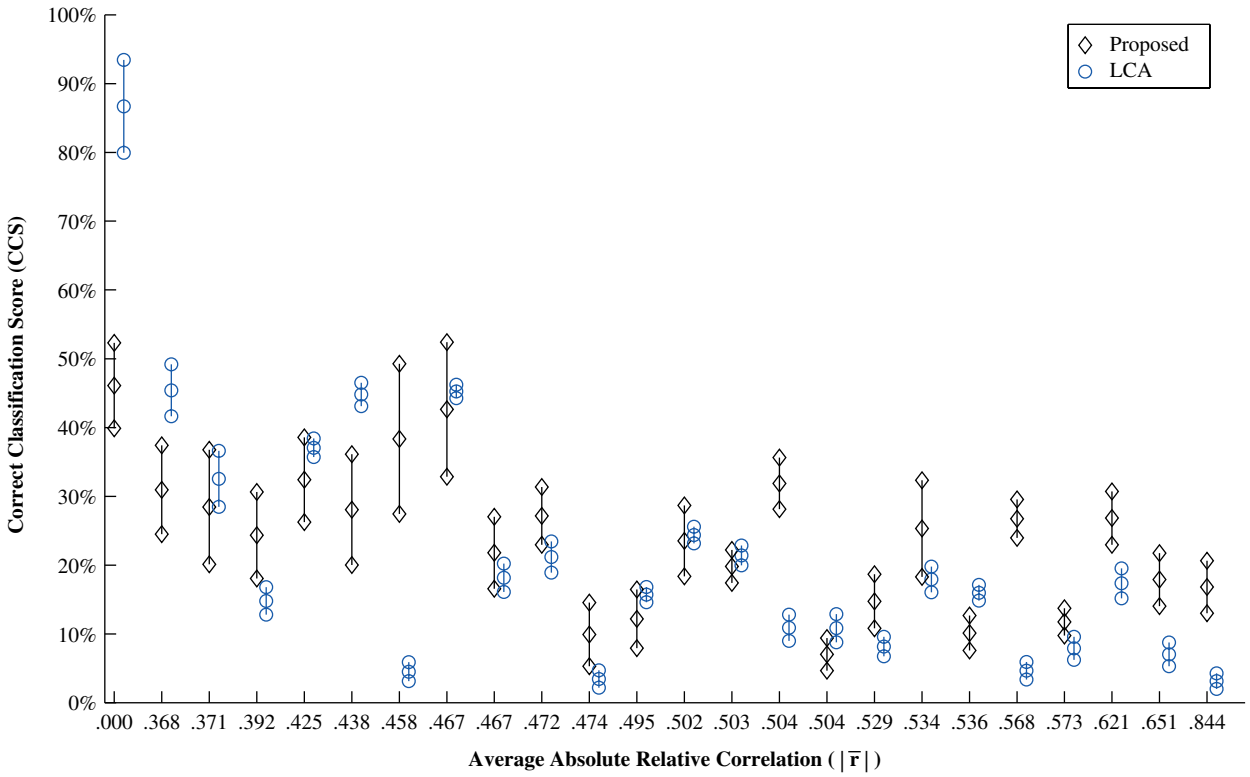

Fig. 3 Average correct classification scores (CCS) using the proposed and the LCA methods [data generated using the MVP model; $N = 5000$)].

Table 8. BIC values for teaching survey data for the proposed method.

K			
1	2	3	4
−3172.67	−3136.77[a]	−3155.49	−3178.0

[a] The maximum BIC value.

each cluster would not be possible without explicit modeling of the correlations in the proposed method.

9. SUMMARY AND CONCLUSIONS

In this paper, we have given a model-based method for clustering of multivariate binary responses. The multivariate Bernoulli distribution used in the mixture model is an

Table 10. Estimated relative correlation matrices for two clusters using the proposed method for teaching style data.

$$\widehat{\boldsymbol{R}}_1 = \begin{bmatrix} 1.000 & 0.956 & 0.602 & -0.090 & -0.433 & -0.017 \\ & 1.000 & 0.605 & -0.089 & -0.620 & -0.015 \\ & & 1.000 & -0.160 & -0.100 & -0.099 \\ & & & 1.000 & -0.351 & 0.291 \\ & & & & 1.000 & -0.533 \\ & & & & & 1.000 \end{bmatrix}$$

$$\widehat{\boldsymbol{R}}_2 = \begin{bmatrix} 1.000 & 0.233 & 0.219 & -0.304 & 0.375 & -0.291 \\ & 1.000 & 0.164 & -0.179 & 0.138 & -0.136 \\ & & 1.000 & -0.160 & 0.188 & -0.185 \\ & & & 1.000 & -0.123 & 0.007 \\ & & & & 1.000 & -0.899 \\ & & & & & 1.000 \end{bmatrix}$$

Table 9. Estimates of the θ's and η's for two clusters using the proposed and the LCA methods for teaching style data.

Method	Cluster	$\widehat{\theta}_1$	$\widehat{\theta}_2$	$\widehat{\theta}_3$	$\widehat{\theta}_4$	$\widehat{\theta}_5$	$\widehat{\theta}_6$	$\widehat{\eta}$
Proposed	1	0.846	0.770	0.675	0.342	0.846	0.277	0.62
	2	0.638	0.571	0.555	0.362	0.429	0.622	0.38
LCA	1	0.858	0.764	0.711	0.285	0.967	0.098	0.61
	2	0.630	0.596	0.520	0.459	0.243	0.918	0.39

Table 11. BIC values for newspaper survey data for the proposed method.

K			
1	2	3	4
−54822.8	−51903.3	−51673.6[a]	−51677.9

[a]The maximum BIC value.

extension of Oman and Zucker's [1] model. The proposed clustering method generalizes the traditional LCA which assumes local independence.

The maximum likelihood method is used to estimate the model parameters for all clusters and the mixing proportions (prior probabilities). Commercial optimization software is used for parameter estimation. Application of the proposed method to two real data sets indicates that the method is practicable at least for a small number of variables and gives interpretable results. The resulting clusters can be assigned meaningful labels (e.g. subscribers and non-subscribers in the newspaper survey example). Although, the proposed and the LCA methods both give fairly similar results, the proposed method also gives estimates of the relative correlation matrices for the two clusters, which give insight into the relationships between the response variables as seen from the two examples.

Clearly, much remains to be done in this problem. First, as noted before, we need to compare the proposed approach with the log-linear model approach implemented in Latent Gold. Faster computational methods need to be developed to handle larger values of m. Increasing the number of responses, m, allows better discrimination between a fixed number of clusters or fitting more clusters to the data. However, because the number of patterns grows exponentially with m, it would be virtually impossible to handle very large values of m (although for sparse data the number of actual patterns may be much less than 2^m, which may help lighten the computational burden); incidentally, this problem affects the computing time of any method, even those that do not take into account correlations such as the LCA method. Therefore some method of prescreening the variables is needed in order to reduce a large number of responses to a manageable number before the proposed method can be applied.

Table 13. Estimated relative correlation matrices for two clusters using the proposed method for newspaper survey data.

$$\widehat{\boldsymbol{R}}_1 = \begin{bmatrix} 1.000 & 0.275 & 0.249 & 0.263 & 0.204 & 0.151 & -0.277 \\ & 1.000 & 0.292 & 0.267 & 0.180 & 0.056 & -0.335 \\ & & 1.000 & 0.225 & 0.163 & -0.029 & -0.300 \\ & & & 1.000 & 0.218 & 0.206 & -0.266 \\ & & & & 1.000 & 0.299 & -0.175 \\ & & & & & 1.000 & -0.038 \\ & & & & & & 1.000 \end{bmatrix}$$

$$\widehat{\boldsymbol{R}}_2 = \begin{bmatrix} 1.000 & 0.944 & 0.935 & 0.938 & 0.929 & 0.863 & 0.840 \\ & 1.000 & 0.988 & 0.992 & 0.982 & 0.914 & 0.886 \\ & & 1.000 & 0.981 & 0.972 & 0.904 & 0.877 \\ & & & 1.000 & 0.975 & 0.908 & 0.880 \\ & & & & 1.000 & 0.899 & 0.872 \\ & & & & & 1.000 & 0.812 \\ & & & & & & 1.000 \end{bmatrix}$$

Also, the problem of determination of optimum number of clusters needs further research. Our current recommendation is to use the BIC criterion subject to the requirement of interpretability of the resulting clusters. Finally, most real data sets involve a combination of binary and continuous (as well as categorical and ordinal) responses, and it would be desirable to develop clustering methods to deal with such hybrid data sets. In conclusion, this is a fertile area for research with diverse potential applications to clustering and data mining.

Acknowledgments

The authors wish to thank the editor and two referees for their very helpful comments, Professor Edward Malthouse of the Integrated Marketing Communications Program at the Medill School of Journalism and Media Management Center at Northwestern University for allowing access to newspaper survey data and help with interpretation of results, and UK Data Archive of University of Essex for allowing access to the teaching survey data. This research was supported by a grant from National Security Agency Award No. H98230-07-01-0068.

Table 12. Estimates of the θ's and η's for two clusters using the proposed and the LCA methods for newspaper survey data.

Method	Cluster	$\widehat{\theta}_1$	$\widehat{\theta}_2$	$\widehat{\theta}_3$	$\widehat{\theta}_4$	$\widehat{\theta}_5$	$\widehat{\theta}_6$	$\widehat{\theta}_7$	$\widehat{\eta}$
Proposed	1	0.147	0.067	0.215	0.132	0.249	0.289	0.956	0.46
	2	0.888	0.888	0.889	0.888	0.891	0.820	0.858	0.54
LCA	1	0.117	0.045	0.175	0.110	0.239	0.259	0.856	0.51
	2	0.991	0.997	0.997	0.997	0.993	0.906	0.953	0.49

REFERENCES

[1] S. D. Oman and D. M. Zucker, Modeling and generating correlated binary variables, Biometrika 88(1) (2001), 287–290.

[2] G. Schwarz, Estimating the dimension of a model, Ann Stat 6 (1978), 461–464.

[3] J. A. Hartigan, Clustering Algorithms, New York, John Wiley & Sons, 1975.

[4] B. S. Everitt, Cluster Analysis (3rd ed.), New York: Halsted Press, 1993.

[5] J. R. Kettenring, A perspective on cluster analysis, Stat Anal Data Min 1 (2008), 52–53.

[6] P. D. Hoff, Subset clustering of binary sequences, with an application to genomic abnormality data, Biometrics 61 (2005), 1027–1036.

[7] S. N. Bennett and J. Jordan, A typology of teaching styles in primary schools, Br J Educ Psychol 45 (1975), 20–28.

[8] M. Aitkin, D. Anderson, and J. Hinde, Statistical modeling of data on teaching styles, J R Stat Soc Ser A 144 (1981), 419–461.

[9] D. J. Bartholomew and M. Knott, Latent Variable Models and Factor Analysis (2nd ed.), New York: Oxford University Press, 1999.

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

[11] L. A. Emrich and M. R. Piedmonte, A method for generating high-dimensional multivariate binary variates, Am Stat 45 (1991), 302–304.

[12] Y. Hochberg and A. C. Tamhane, Multiple Comparison Procedures, New York: John Wiley & Sons, 1987.

[13] L. A. Goodman, Analyzing Qualitative/Categorical Data: Log-Linear Models and Latent Structure Analysis, Cambridge, MA: Abt Books, 1978.

[14] J. A. Hagenaars, Latent structure models with direct effects between indicators: Local dependence models, Sociol Methods Res 16 (1988), 379–405.

[15] J. S. Uebersax, Probit latent class analysis with dichotomous or ordered category measures: conditional independence/dependence models, Appl Psyschol Meas 23 (1999), 283–297.

[16] J. K. Vermunt and J. Magidson, Technical Guide to Latent GOLD 4.0: Basic and Advanced, Belmont, MA: Statistical Innovations, Inc., 2004,.

[17] R. L. Prentice, Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors, J Am Stat Assoc 81 (1986), 321–327.

[18] M. A. Al-Osh and S. J. Lee, A simple approach for generating correlated binary variates, J Stat Comput Simul 70 (2001), 231–255.

[19] A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J R Stat Soc Ser B 39 (1977), 1–38.

[20] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference and Prediction, New York, Springer, 2001.

[21] C. Keribin, Consistent estimation of the order of mixture models, Indian J Stat Ser A 1 (2000), 49–66.

[22] A. C. Tamhane, Statistical Analysis of Designed Experiments, New York, John Wiley & Sons, 2009.

[23] R. H. Byrd, J. Nocedal, and R. A. Waltz, KNITRO: an integrated package for nonlinear optimization, In Large-scale Nonlinear Optimization, G. di Pillo and M. Roma, eds. New York, Springer Verlag, 2006, 35–59.

[24] R. H. Byrd, M. E. Hribar, and J. Nocedal, An interior point method for large scale nonlinear programming, SIAM J Optim 9 (1999), 877–900.

APPENDIX

Computational Details

We used the algorithm KNITRO-4.0 by Ziena Optimization, Inc. [23] for parameter estimation. The algorithm is described in Byrd, Hribar and Nocedal [24]. It requires only the gradient of the objective function at each step and not the Hessian matrix. Gradients w.r.t. the θ_{jk} at points of non-differentiability were computed by taking an average of the gradients on both sides of those points.

No NLP algorithm can guarantee a global maximum solution for an arbitrary objective function such as ours. Therefore we tried n different starting combinations of the values of n parameters whose MLEs have to be found and applied the algorithm for each starting combination. The best solution that yielded the largest log-likelihood function was taken to be the global maximum. Instead of choosing the n starting combinations at random we chose them in a systematic manner by using an $n \times n$ Latin square. Because the goal here is to cover the parameter space as uniformly as possible so as not to miss the global maximum, we used the simplest Latin square obtained by cyclically permuting the levels of the factors in each of the n runs.

The parameter estimation process was implemented in Microsoft Visual C++ environment with Knitro executable in a dynamic link library. The program was run on a single PC with 2.99 GHz CPU speed and 512 MB of RAM. Because of the high computational demand in the simulation studies, distributed computing was employed across several PCs. The example data and the C++ codes for simulation and parameter estimation can be obtained by contacting the second author or visiting his research website: http://www.ie.miami.edu/qiu/research.html.

Derivation of the Joint Distribution of X in Eq. 7

Consider a pattern P with index p. Let $Q = M \setminus P$, i.e. $X_i = 1 \; \forall \; i \in P$ and $X_i = 0 \; \forall \; i \in Q$. Then

$$
\begin{aligned}
f(p|\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}) &= \Pr\{Y_i \le \theta_i \; \forall \; i \in P; Y_i > \theta_i \; \forall \; i \in Q\} \\
&= \sum_{A \subseteq P} \sum_{B \subseteq Q} \Pr\{Z_0 W_i + (1 - Z_0)(1 - W_i) \\
&\le \theta_i \; \forall \; i \in A; Z_0 W_i + (1 - Z_0)(1 - W_i) \\
&> \theta_i \; \forall \; i \in B; Z_i W_i + (1 - Z_i)(1 - W_i) \\
&\le \theta_i \; \forall \; i \in P \setminus A; Z_i W_i + (1 - Z_i)(1 - W_i) \\
&> \theta_i \; \forall \; i \in Q \setminus B\} \\
&\quad \times \prod_{i \in A, B} \beta_i \prod_{i \in P \setminus A, Q \setminus B} (1 - \beta_i),
\end{aligned}
$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m)$. Let $C = \{i \in A \cup B : W_i = 1\}$ and $D = \{i \in A \cup B : W_i = 0\}$. Then it is readily seen that $\theta^*(A, B, C) \leq Z_0 \leq \theta^{**}(A,B,C)$. Therefore the probability pertaining to Z_0 is $[\theta^{**}(A,B,C) - \theta^*(A,B,C)]^+$.

Next note that for $i \in P \setminus A$, if $W_i = 1$ then $Z_i \leq \theta_i$ and if $W_i = 0$ then $Z_i > 1 - \theta_i$; in either case the probability pertaining to Z_i is θ_i. Similarly, for $i \in Q \setminus B$, if $W_i = 1$ then $Z_i > \theta_i$ and if $W_i = 0$ then $Z_i \leq 1 - \theta_i$; in either case the probability pertaining to Z_i is $1 - \theta_i$. Putting all these pieces together, we get the final expression for the joint distribution of $\boldsymbol{X}$ as Eq. 7.

Proof of Proposition 1

For the CLV1 model, we assume $\beta_j \equiv \beta$ for all j. Hence Eq. (6) becomes

$$\rho_{ij}^{(1)} = \begin{cases} \beta^2 \rho_{ij}^{**}[1 - \{\gamma_i(1-\gamma_j) \\ \quad +\gamma_j(1-\gamma_i)\}/\max(\theta_i, 1-\theta_j)] & \text{if } \theta_i \leq \theta_j \\ \beta^2 \rho_{ij}^{**}[1 - \{\gamma_i(1-\gamma_j) \\ \quad +\gamma_j(1-\gamma_i)\}/\max(\theta_j, 1-\theta_i)] & \text{if } \theta_i \geq \theta_j. \end{cases}$$

To explore the full range of $\rho_{ij}^{(1)}$, let $\beta = 1$. The minimum value of $\gamma_i(1-\gamma_j) + \gamma_j(1-\gamma_i)$ is 0 and the maximum value is 1. Therefore $\max \rho_{ij}^{(1)}$ attains the upper bound ρ_{ij}^{**}. Now we show that $\min \rho_{ij}^{(1)}$ attains the lower bound $-\rho_{ij}^*$.

The values of ρ_{ij}^* and ρ_{ij}^{**} are different in the four regions of the (θ_i, θ_j)-space (see Fig. 1):

Region I: $\theta_i \leq \theta_j, \theta_i + \theta_j \leq 1$ (i.e. $\max(\theta_i, 1-\theta_j) = 1 - \theta_j$)

Region II: $\theta_i \leq \theta_j, \theta_i + \theta_j \geq 1$ (i.e. $\max(\theta_i, 1-\theta_j) = \theta_i$)

Region III: $\theta_i \geq \theta_j, \theta_i + \theta_j \geq 1$ (i.e. $\max(\theta_j, 1-\theta_i) = \theta_j$)

Region IV: $\theta_i \geq \theta_j, \theta_i + \theta_j \leq 1$ (i.e. $\max(\theta_j, 1-\theta_i) = 1 - \theta_i$).

In region I, we have

$$\rho_{ij}^* = \sqrt{\frac{\theta_i\theta_j}{(1-\theta_i)(1-\theta_j)}} \text{ and } \rho_{ij}^{**} = \sqrt{\frac{\theta_i(1-\theta_j)}{\theta_j(1-\theta_i)}}. \quad \text{(A.1)}$$

Therefore

$$\begin{aligned} \min \rho_{ij}^{(1)} &= \rho_{ij}^{**}\left[1 - \frac{1}{\max(\theta_i, 1-\theta_j)}\right] \\ &= \sqrt{\frac{\theta_i(1-\theta_j)}{\theta_j(1-\theta_i)}}\left[1 - \frac{1}{1-\theta_j}\right] \\ &= -\sqrt{\frac{\theta_i\theta_j}{(1-\theta_i)(1-\theta_j)}} \\ &= -\rho_{ij}^*. \end{aligned}$$

Thus $\rho_{ij}^{(1)}$ attains the lower bound.

Next, for the CLV2 model, we assume $\gamma_j \equiv \gamma$ for all j. Hence Eq. 6 becomes

$$\rho_{ij}^{(2)} = \begin{cases} \beta_i\beta_j\rho_{ij}^{**}[1 - \{2\gamma(1-\gamma)\}/ \\ \quad \max(\theta_i, 1-\theta_j)] & \text{if } \theta_i \leq \theta_j \\ \beta_i\beta_j\rho_{ij}^{**}[1 - \{2\gamma(1-\gamma)\}/ \\ \quad \max(\theta_j, 1-\theta_i)] & \text{if } \theta_i \geq \theta_j. \end{cases}$$

To explore the full range of $\rho_{ij}^{(2)}$, let $\beta_i = \beta_j = 1$. The minimum value of $\gamma(1-\gamma)$ is 0 and the maximum value is 1/4. Hence in region I we have

$$\min \rho_{ij}^{(2)} = \rho_{ij}^{**}\left[1 - \frac{1}{2\max(\theta_i, 1-\theta_j)}\right] \text{ and}$$

$$\max \rho_{ij}^{(2)} = \rho_{ij}^{**}.$$

Thus $\rho_{ij}^{(2)}$ attains the upper bound. Furthermore, using the bounds Eq. A.1, we have

$$\begin{aligned} \min \rho_{ij}^{(2)} &= \rho_{ij}^{**} - \frac{1}{2(1-\theta_j)}\sqrt{\frac{\theta_i(1-\theta_j)}{\theta_j(1-\theta_i)}} \\ &= \rho_{ij}^{**} - \frac{1}{2}\sqrt{\frac{\theta_i}{\theta_j(1-\theta_i)(1-\theta_j)}} \\ &= \rho_{ij}^{**} - \frac{1}{2}(\rho_{ij}^* + \rho_{ij}^{**}) \\ &= \frac{1}{2}(\rho_{ij}^{**} - \rho_{ij}^*). \end{aligned}$$

Therefore

$$\begin{aligned} \min \rho_{ij}^{(2)} > 0 &\iff \rho_{ij}^{**} > \rho_{ij}^* \iff \sqrt{\frac{\theta_i(1-\theta_j)}{\theta_j(1-\theta_i)}} \\ &> \sqrt{\frac{\theta_i\theta_j}{(1-\theta_i)(1-\theta_j)}} \iff \theta_j < 1/2. \end{aligned}$$

A similar proof can be given for the other three regions with the following results:

Region I: $\min \rho_{ij}^{(2)} > 0 \iff \theta_j < 1/2$

Region II: $\min \rho_{ij}^{(2)} > 0 \iff \theta_i > 1/2$

Region III: $\min \rho_{ij}^{(2)} > 0 \iff \theta_j > 1/2$

Region IV: $\min \rho_{ij}^{(2)} > 0 \iff \theta_i < 1/2$.

These four subregions are shown shaded in Figure 1.

We see that they can be summarized simply as both θ_i, θ_j are $< 1/2$ or $> 1/2$ thus proving the proposition. ■

Proof of Proposition 2

First consider the upper bound UCCR. In order to maximize the number of observations that are correctly classified, one must assign each pattern p to that cluster which yields the maximum number of observations having that pattern. This proves the upper bound UCCR.

Next consider the lower bound LCCR. Consider a $2^m \times K$ table in which the patterns are the rows and the clusters are the columns. The entries in the table are n_{pk}, which are the number of observations having pattern p that come from cluster C_k. There are $K!$ possible assignments of cluster labels. Let $\sigma = (\sigma(1), \sigma(2), \ldots, \sigma(K))$ be a permutation of the cluster labels. Then for this permuted assignment of the cluster labels to the patterns, the CCR is

$$\text{CCR}_\sigma = \frac{1}{N} \sum_{p=1}^{2^m} \sum_{k=1}^{K} n_{p\sigma(k)} I\left(p \in C_{\sigma(k)}\right), \tag{A.2}$$

where $I\left(p \in C_{\sigma(k)}\right) = 1$ if pattern p is classified to cluster $C_{\sigma(k)}$ and 0 otherwise. As a pattern p can be assigned to exactly one cluster, only one of the indicator variables, $I\left(p \in C_{\sigma(k)}\right)$, equals 1 for $k = 1, 2, \ldots, K$ and others equal zero.

The $K!$ permutations can be divided into $(K-1)!$ groups, each consisting of K permutations, such that if two permutations σ_1 and σ_2 belong to the same group then $\sigma_1(k) \neq \sigma_2(k)$ for $k = 1, \ldots, K$. For example, for $K = 3$, the six permutations divide into two groups: $G_1 = \{(1, 2, 3), (2, 3, 1), (3, 1, 2)\}$ and $G_2 = \{(1, 3, 2), (2, 1, 3), (3, 2, 1)\}$. Within each group CCR_σ sum to 1. To see this first consider a numerical example for $K = 3$ and $m = 2$. Label the four patterns, $(0, 0), (1, 0), (0, 1), (1, 1)$ as $1, 2, 3, 4$. Then $\sum_{p=1}^{4} \sum_{k=1}^{3} n_{pk} = N$. Suppose a clustering rule classifies pattern 1 to cluster 1, pattern 2 to cluster 2, and patterns 3 and 4 to cluster 3. Then CCR for this rule is

$$\text{CCR}_1 = \frac{n_{11} + n_{22} + n_{33} + n_{43}}{N}.$$

But the cluster labels can be permuted to $(2, 3, 1)$ or $(3, 1, 2)$ in the group G_1. CCR for these two permutations are, respectively,

$$\text{CCR}_2 = \frac{n_{12} + n_{23} + n_{31} + n_{41}}{N} \text{ and}$$
$$\text{CCR}_3 = \frac{n_{13} + n_{21} + n_{32} + n_{42}}{N}.$$

Hence,

$$\text{CCR}_1 + \text{CCR}_2 + \text{CCR}_3 = \frac{\sum_{p=1}^{4} \sum_{k=1}^{3} n_{pk}}{N} = 1.$$

More generally, let $\sigma_1, \sigma_2, \ldots, \sigma_K$ denote K permutations in one of these groups. Then

$$\sum_{j=1}^{K} \text{CCR}_{\sigma_j} = \frac{1}{N} \sum_{p=1}^{2^m} \sum_{k=1}^{K} \sum_{j=1}^{K} n_{p\sigma_{j(k)}} I\left(p \in C_{\sigma_{j(k)}}\right).$$

Now for each k there is exactly one j for which $I(p \in C_{\sigma_{j(k)}}) = 1$; for all other j, $I(p \in C_{\sigma_{j(k)}}) = 0$. Denote the corresponding $\sigma_{j(k)} = \ell$. Furthermore, for each such (j, k) combination we have a distinct value of ℓ and hence ℓ runs through 1 to K. Substituting this simplification in the above expression we get

$$\sum_{j=1}^{K} \text{CCR}_{\sigma_j} = \frac{1}{N} \sum_{p=1}^{2^m} \sum_{\ell=1}^{K} n_{p\ell} = 1.$$

Therefore there is at least one assignment, σ_j, of cluster labels in each group such that $\text{CCR}_{\sigma_j} \geq 1/K$. Hence the lower bound on CCR is $1/K$. ■