Mixed Mode Data Clustering: An Approach Based on Tetrachoric Correlations

Isabella Morlini

Abstract In this paper we face the problem of clustering mixed mode data by assuming that the observed binary variables are generated from latent continuous variables. We perform a principal components analysis on the matrix of tetrachoric correlations and we then estimate the scores of each latent variable and construct a data matrix with continuous variables to be used in fully Guassian mixture models or in the k-means cluster analysis. The calculation of the expected a posteriori (EAP) estimates may proceed by simply considering a limited number of quadrature points. Main results on a simulation study and on a real data set are reported.

1 Introduction

One possible approach to cluster analysis is the mixture maximum likelihood method, in which the data to be clustered are assumed to come from a finite mixture of populations. The method has been well developed and much used for the case of normal populations. A main advantage in using Gaussian distributions is that a number of possible restrictions on the covariance matrices has been proposed in literature (e.g., [1, 3]) to deal with different local dependencies and, at the same time, to alleviate the problem of the rapidly growing of the parameters with the data dimension and with the number of clusters. A large range of Gaussian models are available, from the simple spherical one to the least parsimonious where all elements of the covariance matrix are allowed to vary across clusters. Practical applications, however, often involve mixture of categorical and continuous variables. Everitt [4] and Everitt and Merette [5] extended the normal model to deal with mixed mode data but the computation involved in their model is so extensive that is only feasible for data with very few categorical variables. Lawrence and Krzanowski [7] and Vermunt & Magidson [12] propose conditional Gaussian models with local

I. Morlini (✉)
Dipartimento di Scienze Sociali, Cognitive e Quantitative, Università di Modena e Reggio Emilia, 42100 Reggio Emilia, Italy,
email: isabella.morlini@unimore.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_9,

independence structure. Local dependencies are specified only between pairs of categorical variables and between pairs of continuous variables and are dealt via joint multinomial and multivariate normal distributions. In the "Latent Gold" package [11] the dependence between a categorical and a continuous variable may be dealt with a sort of "trick", by doubling the categorical variable and treating the variable also as a covariate. The estimated dependence, however, may not vary between groups. The mixture model for large data sets implemented in the package SPSS is also based on joint multinomial and gaussian distributions and postulate the hypothesis of local independence between a categorical and a continuous variable.

Here we face the problem of clustering data with different scales and allowing local dependencies also between a categorical and a continuous variable by assuming that each observed categorical variable is generated from a latent continuous variable and by estimating the scores of these latent variables. In economics, these variables are called utility functions and the assumption is that the response (which may be, for example, the presence or the absence of a public service or a public utility) are determined by the crossing of certain thresholds in these functions (see, among others, [8]). Heckman [6] models whether or not American states have introduced fair-employment legislation and describes the corresponding latent response as the "sentiment" favoring fair-employment legislation. In genetics, the latent response is interpreted as the "liability" to develop a qualitative trait or phenotype. There are also examples of continuous variables which are sampled as binary (among others, bit data which are originated by electric voltages). Skrondal and Rabe-Hesketh [10], pp. 16–17, report various interpretations of these latent variables and also state that assuming a latent continuous variable may be useful regardless of whether the latent response can be given a real meaning.

This work represents the first step in the construction of fully Gaussian models for classification, in which correlations among variables may vary across groups and also variable selection may be faced differently in each group. Here we estimate the scores of each latent variable and reach a data matrix with all continuous variables to be used in these models. An application shows that some benefits of using a data matrix with all continuous variables instead of a mixed mode data matrix may be reached in the k-means cluster analysis.

2 From Binary Variables to Continuous Variables

The essential feature of the method to be described in this section is that the observed categorical variables are generated from underlying latent continuous variables according to the values of a set of thresholds. Here we formalize results regarding binary variables but the theory may be extended to multinomial variables by estimating the matrix of polychoric correlations. Given p vectors of binary variables observed for a sample of size n, a contingency table for each couple of variables X_k and X_j is constructed, with the following cell frequencies:

	$x_k = 0$	$x_k = 1$
$x_j = 0$	e_{jk}	b_{jk}
$x_j = 1$	c_{jk}	d_{jk}

The estimated value for the threshold generating the variable X_k is the value h_k satisfying $\Phi(h_k) = (e_{jk} + c_{jk})/n$. For variable X_j it is the value h_j satisfying $\Phi(h_j) = (e_{jk} + b_{jk})/n$, where Φ is the standard normal cumulative distribution function. We then estimate the tetrachoric correlation coefficient r_{jk} conditional on these thresholds, via maximum likelihood. The solution may be found iteratively or by using the following approximate analytic solution:

$$r_{jk} = \sin\left(\frac{\pi}{2}\left(1 + \frac{4e_{jk}b_{jk}c_{jk}d_{jk}n^2}{(e_{jk}d_{jk} - b_{jk}c_{jk})^2(e_{jk} + d_{jk})(b_{jk} + c_{jk})}\right)^{-1/2}\right) \quad (1)$$

In tables with zero frequencies, zero values are set to 0.5. In a simulation study with 5000 different data sets of size (100×6) generated from 10 multivariate normal populations, the estimator (1) has been shown to give better results than the other ones based on approximate analytic solutions of the likelihood function. The $(n \times p)$ matrix of the scores of the p latent continuous variables is reached with expected a posteriori (EAP) estimates. In order to reach semi parametric estimates, we consider a model based on principal components rather than on factors (see, for example, [2] and [9], for EAP estimates reached by considering a fully parametric model where also thresholds, eigenvalues and eigenvectors associated with each factor are estimated by maximizing the likelihood function). We perform a principal component analysis on the matrix of tetrachoric correlations (which does not require previous smoothing if the matrix is not positive definite) and consider the following model:

$$t_{ij} = a_{j1}y_{i1} + a_{j2}y_{i2} + \ldots + a_{jk}y_{ik} + \ldots + a_{jp}y_{ip} \quad (2)$$

where t_{ij} is the score of principal component j for case i, a_{jk} are the loadings (eigenvectors) and y_{ik} is the score for case i relative to the k latent variable associated with the observed categorical variable x_k as follows: $x_{ik} = 1$ if $y_{ik} \geq h_k$ and $x_{ik} = 0$ if $y_{ik} < h_k$. As assumed for the thresholds estimates, $\mathbf{y} \sim N(\mathbf{0}, I)$ and $\mathbf{t} \sim N(\mathbf{0}, \Lambda)$ where Λ is a diagonal matrix with elements $\lambda_j^2 = \sum_{k=1}^{p} a_{jk}^2$ equal to the eigenvalues. The EAP estimator of the jth principal component score is the mean of the posterior distribution of t_j, which is expressed by:

$$\tilde{t}_{ij} = E(t_{ij}|\mathbf{x}_i; \mathbf{w}) = \int t_j f(t_j|\mathbf{x}_i; \mathbf{w}) dt_j = \int \frac{t_j f(\mathbf{x}_i|t_j; \mathbf{w}) g(t_j|\mathbf{w})}{\int f(\mathbf{x}_i|t_j; \mathbf{w}) g(t_j|\mathbf{w}) dt_j} dt_j \quad (3)$$

where $\mathbf{w}$ is the vector of known parameters (the thresholds and the eigenvectors). In the following equations, for economy of space, $\mathbf{w}$ will be omitted. Given $\sigma_{jk}^2 = \lambda_j^2 - a_{jk}^2 = \sum_{h \neq k} a_{jh}^2$, then

$$P(x_{ik} = 1|t_j) = \frac{1}{\sigma_{jk}\sqrt{2\pi}} \int_{h_k}^{\infty} e^{\left(-(t_{ij} - a_{jk} y_{ik})^2 / 2\sigma_{jk}^2\right)} dy_{ik} \tag{4}$$

Introducing the change in the variable:

$$P(x_{ik} = 1|t_j) = \frac{1}{a_{jk}\sqrt{2\pi}} \int_{-\infty}^{(t_{ij} - a_{jk} h_k)/\sigma_{jk}} e^{(-z^2/2)} dz \qquad (a_{jk} > 0) \tag{5}$$

$$P(x_{ik} = 1|t_j) = \frac{1}{-a_{jk}\sqrt{2\pi}} \int_{(t_{ij} - a_{jk} h_k)/\sigma_{jk}}^{\infty} e^{(-z^2/2)} dz \qquad (a_{jk} < 0) \tag{6}$$

Letting $z_{jk} = (t_{ij} - a_{jk} h_k)/\sigma_{jk}$ and $F_{jk}(t_j) = (a_{jk})^{-1} \Phi(z_{jk})$ when $a_{jk} > 0$, $F_{jk}(t_j) = |a_{jk}|^{-1}(1 - \Phi(z_{jk}))$ when $a_{jk} < 0$, assuming the independence of the binary variables x_k conditional on each component t_j, it results

$$f(\mathbf{x}_i|t_j) = \prod_{k=1}^{p} F_{jk}(t_j)^{x_{ik}} [1 - F_{jk}(t_j)]^{1 - x_{ik}} \tag{7}$$

We consider S quadrature points and estimate the scores as follows:

$$\tilde{t}_{ij} = \sum_{s=1}^{S} t_{sj} \frac{\phi(t_{sj}) \prod_{k=1}^{p} F_{jk}(t_j)^{x_{ik}} [1 - F_{jk}(t_j)]^{1 - x_{ik}}}{\sum_{s=1}^{S} \phi(t_{sj}) \prod_{k=1}^{p} F_{jk}(t_j)^{x_{ik}} [1 - F_{jk}(t_j)]^{1 - x_{ik}}} \tag{8}$$

where t_{sj} are equally spaced points in $[-z_j, z_j]$ with $\Phi(-z_j/\lambda_j) = 0.001$, $\phi(t_{sj})$ are the density functions of these points in the $N(0, \lambda_j^2)$ curve times the interval size.

Given the estimates $\tilde{t}_{ij}$, the EAP estimates $\tilde{y}_{ik}$ of the latent variables may be then reached through analogous steps. The EAP estimator of the kth variable scores is the mean of the posterior distribution of y_k, which is expressed by:

$$\tilde{y}_{ik} = E(y_{ik}|x_{ik}; \mathbf{t}_i) = \int y_k f(y_k|x_{ik}; \mathbf{t}_i) dy_k = \int \frac{y_k f(x_{ik}|y_k; \mathbf{t}_i) g(y_k)}{\int f(x_{ik}|y_k; \mathbf{t}_i) g(y_k) dy_k} dy_k \tag{9}$$

Let y_{ik}^+ be the values $y_{ik} \geq h_k$ and y_{ik}^- be the values $y_{ik} < h_k$, then

$$P(x_{ik}=1|y_k;\tilde{t}_{ij})=\frac{1}{a_{jk}\sqrt{2\pi}}\int_{-\infty}^{\frac{\tilde{t}_{ij}-a_{jk}y_{ik}^{+}}{\sigma_{jk}}}e^{-z^2/2}dz \qquad (a_{jk}>0) \tag{10}$$

$$P(x_{ik}=1|y_k;\tilde{t}_{ij})=\frac{1}{|a_{jk}|\sqrt{2\pi}}\int_{\frac{\tilde{t}_{ij}-a_{jk}y_{ik}^{+}}{\sigma_{jk}}}^{\infty}e^{-z^2/2}dz \qquad (a_{jk}<0) \tag{11}$$

$$P(x_{ik}=0|y_k;\tilde{t}_{ij})=\frac{1}{|a_{jk}|\sqrt{2\pi}}\int_{-\infty}^{\frac{\tilde{t}_{ij}-a_{jk}y_{ik}^{-}}{\sigma_{jk}}}e^{-z^2/2}dz \qquad (a_{jk}<0) \tag{12}$$

$$P(x_{ik}=0|y_k;\tilde{t}_{ij})=\frac{1}{a_{jk}\sqrt{2\pi}}\int_{\frac{\tilde{t}_{ij}-a_{jk}y_{ik}^{-}}{\sigma_{jk}}}^{\infty}e^{-z^2/2}dz \qquad (a_{jk}>0) \tag{13}$$

Let

$$z_{jk}^{+}=\frac{\tilde{t}_{ij}-a_{jk}y_{ik}^{+}}{\sigma_{jk}} \qquad z_{jk}^{-}=\frac{\tilde{t}_{ij}-a_{jk}y_{ik}^{-}}{\sigma_{jk}} \tag{14}$$

and $F_{jk}^{+}(y_k)=(a_{jk})^{-1}\Phi(z_{jk}^{+})$ when $a_{jk}>0$, $F_{jk}^{+}(y_k)=|a_{jk}|^{-1}(1-\Phi(z_{jk}^{+}))$ when $a_{jk}<0$, $F_{jk}^{-}(y_k)=|a_{jk}|^{-1}\Phi(z_{jk}^{-})$ when $a_{jk}<0$, $F_{jk}^{+}(y_k)=(a_{jk})^{-1}(1-\Phi(z_{jk}^{-}))$ when $a_{jk}>0$. Then $f(x_{ik}|y_k;\mathbf{t}_i)=\sum_{j=1}^{p}F_{jk}^{+}(y_k)^{x_{ik}}F_{jk}^{-}(y_k)^{1-x_{ik}}\times\phi(\tilde{t}_{ij})$. Considering S quadrature points we estimate the scores as follows:

$$\tilde{y}_{ik}=\sum_{s=1}^{S}y_{sk}\frac{\phi(y_{sk})\sum_{j=1}^{p}(F_{jk}^{+}(y_s)^{x_{ik}}F_{jk}^{-}(y_s)^{1-x_{ik}}\times\phi(\tilde{t}_{ij}))}{\sum_{s=1}^{S}\phi(y_{sk})(\sum_{j=1}^{p}F_{jk}^{+}(y_s)^{x_{ik}}F_{jk}^{-}(y_s)^{1-x_{ik}}\times\phi(\tilde{t}_{ij}))} \tag{15}$$

where y_{sk} are equally spaced points in $[-z_j \quad h_k]$ when $x_{ik}=0$, in $[h_k \quad z_j]$ when $x_{ik}=1$, with $\Phi(-z_j)=0.001$, $\phi(y_{sk})$ being the density functions of these points in the $N(0,1)$ curve times the interval size.

3 Main Results on a Simulation Study and on a Real Data Set

A simulation study is used to evaluate the accuracy of the tetrachoric correlations and the scores estimates. From 10 standard multivariate normal populations with correlation matrices $\mathbf{P}$ with equal elements $\rho_{rc}, r\neq c$, out of the main diagonal, ranging form 0.0 to 0.95, we generate 5,000 data sets (500 from each population) of size (100×6). We then dichotomize the 6 variables by imposing random thresholds from a uniform distribution in the interval $[-2+2]$. The mean absolute errors (MAEs) for the thresholds estimates for each variables (averaged over the 5,000 data sets and the 100 observations of each set) are always less than 0.06. Considering "difficult variables", originated by thresholds outside the interval $[-1+1]$, the MAEs increase to 0.11. These less accurate estimates also lead to larger errors for the scores estimates. Using (1), the mean absolute errors (MAEs) obtained for the

different ρ, averaged over the 500 data sets generated with each correlation matrix, the 100 observations of each set and the 15 correlation coefficients, are:

$\rho = 0$	$\rho = 0.2$	$\rho = 0.3$	$\rho = 0.4$	$\rho = 0.5$	$\rho = 0.6$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$	$\rho = 0.95$
0.07	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.05	0.03

Results seem particularly accurate for all values of ρ. Mean errors also decrease as long as the real correlations among variables increase. Boxplots of the MAEs for the eigenvalues of the principal components, calculated between the eigenvalues of the correlation matrix **P** used to generate the data and the correlation matrix **R** of the generated data, are reported in Fig. 1. For values of ρ_{rc} not exceeding 0.8, estimates

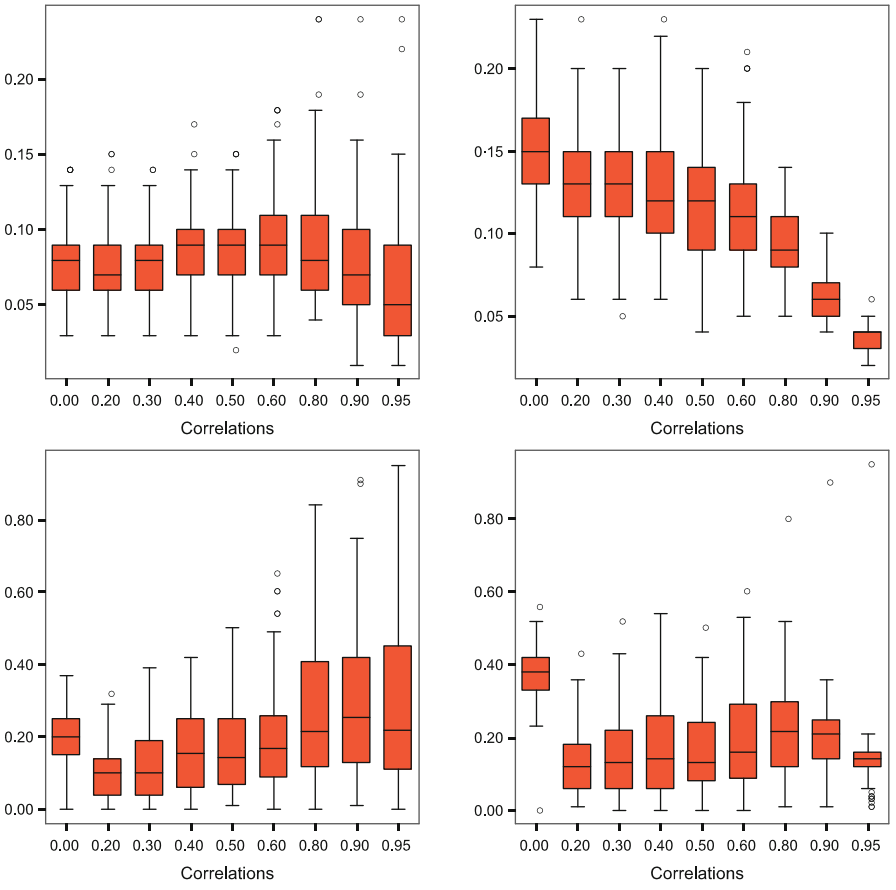

Fig. 1 Boxplots of the mean absolute errors of the eigenvalues plotted along the original correlations ρ_{rc}. In the left-hand boxes, errors are calculated between eigenvalues of the tetrachoric correlation matrix and eigenvalues of the matrix **R** of the generated data. In the right-hand boxes, errors are calculated between eigenvalues of the tetrachoric correlation matrix and eigenvalues of the matrix **P** used to generate the data. In the upper boxes errors are averaged over the six eigenvalues. In the lower boxes, errors are calculated only for the first eigenvalues

of all the eigenvalues better recover the computed correlation matrix, rather than the matrix used to generate the data. This is not true for the first eigenvalue: when this one is large (and the correlations are larger than 0.8) the estimates better recover the first eigenvalues of the matrix **P**. We then estimate the scores of each latent variable and of the principal components. The MAEs, averaged over the 500 data sets generated for each correlation matrix and over the six variables and the 100 observations of each set, are:

	$\rho = 0$	$\rho = 0.2$	$\rho = 0.3$	$\rho = 0.4$	$\rho = 0.5$	$\rho = 0.6$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$	$\rho = 0.95$
MAE $(\tilde{t}_{ij})$	0.87	0.70	0.69	0.65	0.64	0.60	0.57	0.51	0.45	0.42
MAE $(\tilde{y}_{ij})$	0.59	0.59	0.58	0.58	0.58	0.59	0.58	0.58	0.58	0.59

As long as the correlation among variables increases, there is an improvement in the principal components estimates. On the contrary, results regarding the latent variables do not seem to depend on ρ. Estimates of the scores of the latent variables show improvements in average accuracy when the generated thresholds are close to zero, that is are close to the mean (and the median) of the latent variables. When the thresholds are beyond the range $[-1 + 1]$, average errors are significantly greater. Average errors, however, are always less than the variance of each variable and results seem enough accurate. Table 1 reports the MAEs of the latent variables scores obtained in a further study. Here the 6 binary variables are obtained by generating a (5000×6) data set from the same zero-mean multivariate normal populations as before, but with fixed thresholds: -2, -0.5, 0, 0.2, 0.5, 2. Average errors (in the last row) show that the accuracy of the EAP estimates increases as long as the threshold approaches zero. On the other hand, considering the errors computed for different values of the true scores, we note that minima average errors (reported in bold) are obtained for values near the thresholds. The worst fittings are obtained for large positive values when the threshold is -2 and for large negative values when the threshold is $+2$. For variables with thresholds -0.2, 0, 0.2 and 0.5, the correlations between real and estimated scores are 0.74, 0.78, 0.78 and 0.75, respectively.

Table 1 Mean Absolute Errors for the estimates of the 6 latent variables scores, divided into 9 groups. Groups are based on the magnitude of the true score values

	thresh.$= -2$	thresh.$= -0.5$	thresh.$= 0.0$	thresh.$= 0.2$	thresh.$= 0.5$	thresh.$= 2$
scores	MAEs	MAEs	MAEs	MAEs	MAEs	MAEs
< -1.3	0.62	1.02	1.46	1.62	1.89	2.74
$[-1.3 - 0.8)$	**0.20**	0.32	0.77	0.92	1.21	1.99
$[-0.8 - 0.5)$	**0.17**	**0.11**	0.40	0.56	0.84	1.60
$[-0.5 - 0.3)$	0.48	**0.23**	**0.11**	0.27	0.54	1.28
$[-0.3 + 0.0)$	0.75	**0.08**	**0.17**	**0.08**	0.29	1.00
$[+0.0 + 0.3)$	1.02	0.29	**0.15**	**0.23**	**0.08**	0.73
$[+0.3 + 0.5)$	1.30	0.55	**0.13**	**0.09**	**0.19**	0.47
$[+0.5 + 0.8)$	1.61	0.83	0.41	**0.22**	**0.13**	**0.18**
$\geq +0.8$	2.40	1.54	1.14	0.96	0.64	**0.41**
average	0.77	0.53	0.48	0.48	0.50	0.75

We then consider the internet advertisement data set from the UCI machine learning depository (http://archive.ics.uci.edu/ml/). The features encode the geometry of the image as well as phrases occurring in the URL, the image's URL and alt text, the anchor text, and words occurring near the anchor text. The cluster membership of each image is known (clusters are: advertisement or not advertisement). After removing instances with missing values and selecting binary variables with relative frequencies higher than 0.1, we reach a data set with 2,359 instances, 3 continuous variables and 10 binary variables. We perform a k-means cluster analysis and we run the mixture model implemented in SPSS first with mixed mode variables (normalizing continuous variables in the interval [0 1]) and then with all continuous variables (with the estimated scores of the binary ones). The classification error rate decrease from 33 to 30% with k-means and from 35 to 32% with the mixture model.

4 Concluding Remarks

Although it is clearly impossible to generalize from the results presented, it does appear that estimating the scores of the latent continuous variables generating the binary values may improve the clustering results and, above all, it allows fully Gaussian models with different correlations among the variables in each group to be used for classification. This paper describes an initial investigation into the feasibility of estimating the scores of each latent continuous variable. In literature, only EAP estimates of the most relevant factors have been presented, for the different aims of estimating composed items that are assumed to represent a particular set of constructs and for data reduction. Here the aim is to reach a continuous data matrix, of the same dimension of the original one. Possible variations and improvements to the method proposed are relevant topics for future research. Future simulations involve data generated from distributions rather than the normal, to explore whether the EAP estimates work well also in these cases. Indeed, although the threshold estimates are based on the normal distribution and the t_{ij} and the y_{ij} are supposed to be Gaussian, EAP estimates are little affected by the choice of this distribution since loadings and eigenvalues are not estimated by maximum likelihood.

References

1. Banfield, J.D., Raftery, A.E.: Model based Gaussian and non Gaussian clustering. Biometrics **48**, 803–821 (1993)
2. Bock, R.D., Mislevy, R.J.: Adaptive EAP estimation of ability in a microcomputer environment. Appl. Psychol. Meas. **6**(4), 431–434 (1982)
3. Celeux, G., Govaert, G.: Gaussian parsimonious clustering models. Pattern Recognit. **28**: 781–793 (1995)
4. Everitt, B.S.: A finite mixture model for the clustering of mixed mode data. Stat. Probab. Lett. **6**, 305–309 (1988)
5. Everitt, B.S., Merette, C.: The clustering of mixed-mode data: a comparison of possible approaches. J. Appl. Stat. **17**(3), 284–297 (1990)

6. Heckman, J.J.: Dummy endogenous variables in a simultaneous equation system. Econometrica **47**, 153–161 (1978)
7. Lawrence, C.J., Krzanowski, W.J.: Mixture separation for mixed-mode data. Stat. Comput. **6**, 85–92 (1996)
8. Manski, C.: Identification of binary response models. J. Am. Stat. Assoc. **83**, 729–738 (1988)
9. Muraki, E., Engelhard, G.: Full-information item factor analysis: application of EAP scores. Appl. Psychol. Meas. **9**(4), 417–430 (1985)
10. Skrondal, A., Rabe-Hesketh, S.: Generalized Latent Variable Modeling. Chapman & Hall, London (2004)
11. Vermunt, J.K., Magidson, J.: Latent Gold User's Guide. Statistical Innovation, Belmont, MA (2000)
12. Vermunt, J.K., Magidson, J.: Latent class cluster analysis. In: Hagenaars, J.A., McCutcheon, A.L. (eds.) Applied Latent Class Analysis, pp. 89–106. Cambridge University Press, Cambridge (2002)