Using Latent Variables in Model Based Clustering: An E-Government Application

Isabella Morlini

Abstract

Besides continuous variables, binary indicators on ICT (Information and Communication Technologies) infrastructures and utilities are usually collected in order to evaluate the quality of a public company and to define the policy priorities. In this chapter, we confront the problem of clustering public organizations with model-based clustering, and we assume each observed binary indicator to be generated from a latent continuous variable. The estimates of the scores of these variables allow us to use a fully Gaussian mixture model for classification.

1 Introduction

In order to handle mixed continuous and binary variables for classification purposes, in this work we assume each observed categorical variable to be generated from a latent continuous variable. For estimating the scores of these latent variables, we use the method proposed in Morlini [3]. In economics, the latent variables may be interpreted as utility functions. The assumption is that the responses (e.g., the presence or the absence of a public or private service) are determined by the crossing of certain thresholds in these functions. The advantages of using the scores of the latent variables in place of the original categories and then specifying a full Gaussian model are threefold. (1) Classification is possible also for a large number of variables, while most of the models currently used for variables with mixed scale types [2, 5] are feasible only with few categorical indicators. (2) Many forms of restrictions can be imposed on the variances and the covariances, to obtain parsimony. (3) Local dependencies can be specified not only for couples of

I. Morlini (✉)
Department of Economics, University of Modena & Reggio Emilia, Italy
e-mail: isabella.morlini@unimore.it

P.E. Oliveira et al. (eds.), *Recent Developments in Modeling and Applications in Statistics*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-32419-2_1, © Springer-Verlag Berlin Heidelberg 2013

continuous variables, but also for couples of (original) categorical variables and for a continuous and a categorical variable.

As shown by Vermunt and Magidson [5], the possibility to include local dependencies among the indicators may prevent the possibility of ending with a solution that has too many clusters since, often, a simpler solution with less groups may be obtained by including some direct effects between the indicators. Moreover, relaxing the local independence assumption may yield a better classification of objects since omitting a significant bivariate dependency from a latent class cluster model leads to too high weights of the indicators in the classification.

In this work, we propose an e-government application with the data collected for the UNDERSTAND project (European Regions UNDER way towards STANDard indicators for benchmarking information society) of the Emilia–Romagna Region (Italy). The data consist of a set of categorical indicators and continuous variables on ICT, comparable at European level. The chapter is organized as follows. In Sect. 2, we summarize the method used for estimating the latent variables scores. In Sect. 3 we briefly describe the data set and we report results on the application.

2 Estimation of the Latent Variables Scores

Let consider a general set up in which the values of p binary attributes and q quantitative variables are collected for n objects and let $g = p + q$. We indicate with x_k $(k = 1, \ldots, p)$ the p binary attributes and with y_j $(j = 1, \ldots, q)$ the q quantitative variables. For each object i $(i = 1, \ldots, n)$, the p-dimensional vector $\mathbf{x}_i = [x_{i1} \ldots x_{ip}]$ contains the values of the binary attributes and the q-dimensional vector $\mathbf{y}_i = [y_{i1} \ldots y_{iq}]$ contains the values of the quantitative variables. We suppose that the binary values are generated from latent continuous variables ξ_k $(k = 1, \ldots, p)$, and we obtain a new $(n \times g)$ matrix of quantitative variables by estimating the score ξ_{ik} for each object i $(i = 1, \ldots, n)$ and each latent variable k $(k = 1, \ldots, p)$. The score ξ_{ik} is associated with the observed categorical value x_{ik} as follows: $x_{ik} = 1$ if $\xi_{ik} \geq T_k$ and $x_{ik} = 0$ if $\xi_{ik} < T_k$, where T_k is the threshold, obtained from the data, for the k-th latent variable. The method includes the following consecutive steps:

1. Estimate the threshold T_k of each latent variable and the tetrachoric correlation coefficient r_{kl} between each pair $\{k, l\}$ of latent variables.
2. Perform a principal component analysis on the matrix of the tetrachoric correlations and obtain the eigenvectors and the eigenvalues.
3. Estimate the score of each principal component for each object, given the eigenvectors and the eigenvalues.
4. Estimate the score of each latent variable for each object, given the scores of the principal components.

We construct a contingency table for each pair of variables x_l and x_k $(l, k = 1, \ldots, p)$, with the following cell frequencies: The estimated value for the threshold generating the variable x_l is the value T_l satisfying $\Phi(T_l) = (a_{kl} + c_{kl})/n$. For

	$x_l = 0$	$x_l = 1$	Tot.
$x_k = 0$	a_{kl}	b_{kl}	$a_{kl} + b_{kl}$
$x_k = 1$	c_{kl}	d_{kl}	$c_{kl} + d_{kl}$
Tot.	$a_{kl} + c_{kl}$	$b_{kl} + d_{kl}$	n

variable x_k, it is the value T_k satisfying $\Phi(T_k) = (a_{kl} + b_{kl})/n$, where Φ is the standard normal cumulative distribution function. We then estimate the matrix of tetrachoric correlations $\mathbf{R} = (r_{kl})$ $(k, l = 1, \ldots, p)$ conditional on the thresholds T_l and T_k, via maximum likelihood. The tetrachoric correlation, introduced by Pearson [4], is the correlation coefficient r_{kl} that satisfies

$$\frac{d_{kl}}{n} = \int_{T_l}^{\infty} \int_{T_k}^{\infty} \phi(\xi_k, \xi_l, r_{kl})\, \mathrm{d}\xi_k \mathrm{d}\xi_l, \tag{1}$$

where $\phi(\xi_k, \xi_l, r_{kl})$ is the bivariate normal density function:

$$\phi(\xi_k, \xi_l, r_{kl}) = \frac{1}{2\pi\sqrt{1 - r_{kl}^2}} \exp\left[-\frac{1}{2(1 - r_{kl}^2)}(\xi_k^2 - 2r_{kl}\xi_k\xi_l + \xi_l^2)\right]. \tag{2}$$

The solution may be found iteratively or by using one of the analytic formula proposed in the seminal work of Pearson [4]. Since the thresholds and the tetrachoric correlation coefficient are identifiable if no frequency in the contingency table is equal to zero, we replace the zero by one half. We perform a principal component analysis on the matrix $\mathbf{R}$ and consider the following model:

$$t_{ih} = \alpha_{h1}\xi_{i1} + \alpha_{h2}\xi_{i2} + \ldots + \alpha_{hk}\xi_{ik} + \ldots + \alpha_{hp}\xi_{ip}, \tag{3}$$

where t_{ih} $(h = 1, \ldots, p, \quad i = 1, \ldots, n)$ is the score of the h-th principal component t_h for object i, α_{hk} $(k = 1, \ldots, p)$ are the loadings, with $\sum_{h=1}^{p} \alpha_{hk}^2 = 1$, and ξ_{ik} is the score for object i relative to the k-th latent variable. $\mathbf{t} \sim N(\mathbf{0}, \Lambda)$ where Λ is a diagonal matrix with elements $\lambda_h^2 = \sum_{k=1}^{p} \alpha_{hk}^2$, since the principal components are orthogonal. The variance of each component t_h and the coefficients α_{hk} $(h = 1, \ldots, p, \quad k = 1, \ldots, p)$ are estimated through the eigenvalues and the eigenvectors, respectively, of the matrix $\mathbf{R}$, without making any assumption about the distribution of the latent variables ξ_k. Given these values, we estimate the score of the principal components by expected a posteriori (EAP) estimates. This analysis does not require previous smoothing if the matrix is not positive definite. However, for the identifiability of the score estimates, all eigenvalues must be positive and a smoothing procedure is required if the matrix is positively semi-definite but not definite. We use the procedure implemented in Matlab, which adds a regularization term to the matrix. Different regularization terms lead to slightly different solutions. The EAP estimator of the h-th principal component score is the mean of the posterior distribution of t_h, which is expressed by:

$$E(t_h|\mathbf{x}_i, \mathbf{w}) = \int t_h f(t_h|\mathbf{x}_i, \mathbf{w}) \mathrm{d}t_h = \frac{\int t_h f(\mathbf{x}_i|t_h, \mathbf{w})\phi(t_h|\mathbf{w})\mathrm{d}t_h}{\int f(\mathbf{x}_i|t_h, \mathbf{w})\phi(t_h|\mathbf{w})\mathrm{d}t_h}, \tag{4}$$

where $f(\cdot)$ indicates the probability density function, $\mathbf{w}$ is the vector of known parameters (the thresholds and the eigenvectors, estimated geometrically by the principal component analysis on $\mathbf{R}$), and ϕ is the Gaussian distribution. In the following equations, for parsimony, $\mathbf{w}$ will be omitted. For every object i ($i = 1, \ldots, n$), the probability of the k-th binary attribute to be equal to 1, given the h-th principal component score, can be formalized as follows:

$$P(x_{ik} = 1|t_h) = \frac{1}{\sigma_{hk}\sqrt{2\pi}} \int_{T_k}^{\infty} \mathrm{e}^{-\frac{(t_{ih} - \alpha_{hk}\xi_k)^2}{2\sigma_{hk}^2}} \mathrm{d}\xi_k. \tag{5}$$

where $\sigma_{hk}^2 = \lambda_h^2 - \alpha_{hk}^2 = \sum_{l \neq k} \alpha_{hl}^2$. Introducing the change in the variable:

$$P(x_{ik} = 1|t_h) = \frac{1}{\alpha_{hk}\sqrt{2\pi}} \int_{-\infty}^{\frac{t_{ih} - \alpha_{hk}T_k}{\sigma_{hk}}} \mathrm{e}^{\frac{-z^2}{2}} \mathrm{d}z, \qquad \text{when} \qquad \alpha_{hk} > 0, \tag{6}$$

$$P(x_{ik} = 1|t_h) = \frac{1}{-\alpha_{hk}\sqrt{2\pi}} \int_{\frac{t_{ih} - \alpha_{hk}T_k}{\sigma_{hk}}}^{\infty} \mathrm{e}^{\frac{-z^2}{2}} \mathrm{d}z, \qquad \text{when} \qquad \alpha_{hk} < 0. \tag{7}$$

Letting $z_{hk} = (t_{ih} - \alpha_{hk}T_k)/\sigma_{hk}$, we may define the following quantities:

$$\begin{aligned}
F_{hk}(t_h) &= |\alpha_{hk}|^{-1}\varPhi(z_{hk}), && \text{when} \quad \alpha_{hk} > 0 \quad \text{and} \quad x_{ik} = 1 \\
&\quad \text{or} \quad \alpha_{hk} < 0 && \text{and} \quad x_{ik} = 0, \\
F_{hk}(t_h) &= |\alpha_{hk}|^{-1}[1 - \varPhi(z_{hk})], && \text{when} \quad \alpha_{hk} < 0 \quad \text{and} \quad x_{ik} = 1 \\
&\quad \text{or} \quad \alpha_{hk} > 0 && \text{and} \quad x_{ik} = 0,
\end{aligned}$$

where $\varPhi$ is the standard normal cumulative function. Assuming the independence of the binary attributes x_k ($k = 1, \ldots, p$) conditionally on each component t_h, we obtain $f(\mathbf{x}_i|t_h) = \prod_{k=1}^{p} F_{hk}(t_h)^{x_{ik}}[1 - F_{hk}(t_h)]^{1-x_{ik}}$. This assumption may be thought of as rather unrealistic, since at least one latent variable generating a binary attribute is dependent from the other latent variables. In fact, formally this is a weak point of our procedure, which allows for simple and fast computation. Considering S quadrature points, we estimate the scores as follows:

$$\tilde{t}_{ih} = \sum_{s=1}^{S} t_{sh}^q \frac{\phi(t_{sh}) \prod_{k=1}^{p} F_{hk}(t_h)^{x_{ik}}[1 - F_{hk}(t_h)]^{1-x_{ik}}}{\sum_{s=1}^{S} \phi(t_{sh}) \prod_{k=1}^{p} F_{hk}(t_h)^{x_{ik}}[1 - F_{hk}(t_h)]^{1-x_{ik}}}, \tag{8}$$

where t_{sh}^q are equally spaced points in $[-z_h, z_h]$ with $\varPhi(-z_h/\lambda_h) = 0.001$, and $\phi(t_{sh}^q)$ are the density functions of these points in the $N(0, \lambda_h^2)$ curve times the interval size.

Given the estimates $\tilde{t}_{ih}$, the EAP estimates $\tilde{\xi}_{ik}$ of the latent variables may be reached through analogous steps. The EAP estimator of the k-th variable score is the mean of the posterior distribution of ξ_k, which is expressed by $E(\xi_k|x_k) = \int \xi_k f(\xi_k|x_k)\mathrm{d}\xi_k$. Let $\xi_k^+ = \xi_k$ if $\xi_k \geq T_k$ and $\xi_k^- = \xi_k$ if $\xi_k < T_k$. Then:

$$\begin{aligned}
f(\xi_k|x_{ik}=1,\tilde{t}_{ih}) &= \frac{1}{\alpha_{hk}\sqrt{2\pi}}\int_{-\infty}^{\frac{\tilde{t}_{ih}-\alpha_{hk}\xi_k^+}{\sigma_{hk}}} \mathrm{e}^{\frac{-z^2}{2}}\mathrm{d}z, \qquad && \text{if } \alpha_{hk}>0,\\
f(\xi_k|x_{ik}=1,\tilde{t}_{ih}) &= \frac{1}{|\alpha_{hk}|\sqrt{2\pi}}\int_{\frac{\tilde{t}_{ih}-\alpha_{hk}\xi_k^+}{\sigma_{hk}}}^{\infty} \mathrm{e}^{\frac{-z^2}{2}}\mathrm{d}z, \qquad && \text{if } \alpha_{hk}<0,\\
f(\xi_k|x_{ik}=0,\tilde{t}_{ih}) &= \frac{1}{|\alpha_{hk}|\sqrt{2\pi}}\int_{-\infty}^{\frac{\tilde{t}_{ih}-\alpha_{hk}\xi_k^-}{\sigma_{hk}}} \mathrm{e}^{\frac{-z^2}{2}}\mathrm{d}z, \qquad && \text{if } \alpha_{hk}<0,\\
f(\xi_k|x_{ik}=0,\tilde{t}_{ih}) &= \frac{1}{\alpha_{hk}\sqrt{2\pi}}\int_{\frac{\tilde{t}_{ih}-\alpha_{hk}\xi_k^-}{\sigma_{hk}}}^{\infty} \mathrm{e}^{\frac{-z^2}{2}}\mathrm{d}z, \qquad && \text{if } \alpha_{hk}>0.
\end{aligned} \tag{9}$$

Let $z^+_{hk} = \frac{\tilde{t}_{ih}-a_{hk}\xi^+_{ik}}{\sigma_{hk}}$, $z^-_{hk} = \frac{\tilde{t}_{ih}-a_{hk}\xi^-_{ik}}{\sigma_{hk}}$,

$$\begin{aligned}
F^+_{hk}(\xi_k) &= (\alpha_{hk})^{-1}\Phi(z^+_{hk}), && \text{when } \alpha_{hk}>0,\\
F^+_{hk}(\xi_k) &= |\alpha_{hk}|^{-1}(1-\Phi(z^+_{hk})), && \text{when } \alpha_{hk}<0,\\
F^-_{hk}(\xi_k) &= |\alpha_{hk}|^{-1}\Phi(z^-_{hk}), && \text{when } \alpha_{hk}<0,\\
F^-_{hk}(\xi_k) &= (\alpha_{hk})^{-1}(1-\Phi(z^-_{hk})), && \text{when } \alpha_{hk}>0.
\end{aligned} \tag{10}$$

Then $f(\xi_k|x_k) = \sum_{h=1}^{p} F^+_{hk}(\xi_k)^{x_{ik}} F^-_{hk}(\xi_k)^{1-x_{ik}} \times \phi(\tilde{t}_{ih})$. Considering S quadrature points, we estimate the scores as follows:

$$\tilde{\xi}_{ik} = \sum_{s=1}^{S} \xi^q_{sk}\phi(\xi_{sk}) \sum_{h=1}^{p} (F^+_{hk}(\xi_s)^{x_{ik}} F^-_{hk}(\xi_s)^{1-x_{ik}} \times \phi(\tilde{t}_{ih})), \tag{11}$$

where ξ^q_{sk} are equally spaced points in $[-z_k, T_k]$ when $x_{ik} = 0$, in $[T_k, z_k]$ when $x_{ik} = 1$, with $\Phi(-z_k) = 0.001$, $\phi(\xi^q_{sk})$ being the density functions of these points in the $N(0,1)$ curve times the interval size.

3 An E-Government Application

In this section, we present a cluster analysis of the Emilia–Romagna municipalities, based on a set of back office and front office indicators. The indicators aim at establishing to what extent e-government is working within the region. The data have been collected using an online questionnaire, a printed questionnaire sent by post, filled in face to face or obtained over the phone. The aim of this study is to obtain an insight into how municipalities are affected by ICTs and ICT-enabled developments. ICTs have opened up new possibilities for municipalities to overcome traditional disadvantages deriving from remoteness and distance. But instead of increasing the quality of service everywhere, they have been shown to exacerbate disparities. This is due to the difference in the speed and intensity of the adoption of ICTs, and also to the degree that these technological innovations are utilized. Regional investment in infrastructure related to the Information Society have increased over the past years, and regional decision makers are increasingly committed to the development of ICT in society. As a consequence, policy-makers need to be able to identify areas in which public investments and political support are most likely to be successful.

In order to cluster the 268 municipalities and identify the number of areas with a different ICT development level, we consider 20 binary features and 3 continuous variables. The binary indicators indicate the presence of the following online facilities: x_1: online resolutions of the public administration; x_2: call for bids; x_3: e-procurement platform; x_4: service delivery information; x_5: informative e-mails; x_6: telephone and e-mail index; x_7: web site organization for life events; x_8: web site organization for personalization; x_9: web site organization for subjects and/or offices; x_{10}: online questionnaires or forum related to the municipality activities; x_{11}: possibility to enter into the home page with call centers or sms or wap; x_{12}: SUAP; x_{13}: dynamic map; x_{14}: information on the government body; x_{15}: e-mail of the elected representative leadership; x_{16}: information on the possibility to access the restricted area; x_{17}: service chart; x_{18}: interactive site map; x_{19}: pages written in a foreign language; x_{20}: quality approved by W3C Markup Validator. The continuous variables are: y_{21}: percentage of employees with a digital signature; y_{22}: percentage of employees dedicated to ICT; y_{23}: percentage of employees that have received ICT training.

We estimate models from 2 to 5 groups with the Latent Gold package [6], considering a data set with all continuous variables. In this data set, the categorical values $x_1, \ldots, x_{20}$ are substituted with the latent variables scores $y_1, \ldots, y_{20}$ and all variables are treated as Normal in the mixture models. The first model for each number of classes assumes local independence. The other specifications are obtained by subsequently adding the direct relationship between couples of variables, on the basis of the Latent Gold's bivariate residuals information. The bivariate residuals computed by the package indicate how similar the estimated and the observed bivariate associations are. These residuals can be interpreted as lower bound estimates for the improvement in fit in the likelihood when the corresponding local independence constraints are relaxed and, in each model, is added the local dependency with the highest Latent Gold's bivariate residual in the previous model. For assessing and comparing the models, we use the BIC criterion. Table 1 reports the BIC values and the number of parameters. Variables from 1 to 20 are the latent variables scores, and variables y_{21}, y_{22}, and y_{23} are the continuous variables in

Table 1 BIC values and number of estimated parameters (par.)

Model	Description	2 clusters		3 clusters		4 clusters		5 clusters	
		BIC	par.	BIC	par.	BIC	par.	BIC	par.
1	Local independence	26591	70	25629	94	25230	118	25291	142
2	Model 1 + $\sigma_{y_{22}y_{19}}$	25336	72	24023	97	24239	122	23918	147
3	Model 2 + $\sigma_{y_{21}y_{18}}$	23710	74	22739	100	22519	126	22548	152
4	Model 3 + $\sigma_{y_{19}y_{14}}$	24994	76	22894	103	22361	130	21870	157
5	Model 4 + $\sigma_{y_{23}y_{20}}$	23394	78	22656	106	22790	134	22580	162
6	Model 5 + $\sigma_{y_{15}x_6}$	23229	80	22828	109	22383	138	21837	167
7	Model 6 + $\sigma_{y_7y_1}$	23582	82	22325	112	22478	142	21555	172
8	Model 7 + $\sigma_{y_9y_2}$	23066	84	22654	115	22951	146	21925	177
9	Model 8 + $\sigma_{y_{14}y_2}$	22996	86	22073	118	22135	150	22005	182
10	Model 9 + $\sigma_{y_8y_7}$	23187	88	21602	121	21974	154	21472	197
11	Model 10 + $\sigma_{y_{16}y_4}$	22849	90	22315	124	22240	158	21497	202

the original data set. σ_{kj} is the covariance between variables k and j. For each number of groups, the models with more local dependencies have the lowest BIC values. The accuracy of fit in all situations is improved with inclusions of direct relationships between variables: the local independence model always performs worst. The fact that working with more local dependencies may yield a simpler final model with less clusters is evident: model 11 with 2 clusters performs better than model 4 with 3 clusters and model 8 with 4 clusters. Table 2 reports the relative frequencies of category 1 in the binary variables $(x_1, \ldots, x_{20})$ and the mean values of the three continuous variables (y_{21}, y_{22}, y_{23}) in each group, in the 4-clusters partitions with all considered bivariate dependencies (model 11). The classification is in agreement with the criterion of segment addressability suggested by Chaturvedi et al. [1] and related to the degree according to which a clustering solution can be explained by variables that can be controlled by policy makers. Indeed, in group 3, the most densely populated municipalities are clustered, with the most efficient public nets that allow both the distribution of nearly all of the interactive services by the Public Administration and the development of other telecommunication services for citizens (call centers, sms, ...). This cluster is homogeneous with respect to the presence of online information and facilities like: resolutions of the public administration (x_1), call for bids (x_2), informative e-mails (x_5), telephone and e-mail index (x_6), web site organization for life events (x_7), web site organization for subjects and/or offices (x_9), information on the governing body (x_{14}), e-mail of the elected representative leadership (x_{15}), interactive site map (x_{18}), and pages written in a foreign language (x_{19}). This group has a small percentage of employees with a digital signature but higher percentages of employees dedicated to ICT support and employees that have received ICT training. In group 2, the smallest municipalities are clustered. These are usually mountain communities, not in tourist areas. These units are the least technologically advanced and are in areas where it is

Table 2 Cluster means for models 11 with 4 groups

Cluster	1	2	3	4	tot
x_1	0.49	0.00	0.73	0.36	0.44
x_2	0.88	0.34	1.00	0.82	0.82
x_3	0.01	0.00	0.27	0.00	0.03
x_4	0.46	0.06	0.93	0.55	0.44
x_5	0.78	0.66	1.00	1.00	0.78
x_6	0.78	0.22	0.87	0.55	0.71
x_7	0.09	0.00	0.67	0.00	0.10
x_8	0.12	0.00	0.47	0.18	0.13
x_9	0.93	0.31	1.00	0.73	0.85
x_{10}	0.06	0.03	0.33	0.00	0.07
x_{11}	0.05	0.00	0.40	0.18	0.07
x_{12}	0.17	0.00	0.80	0.09	0.18
x_{13}	0.52	0.16	0.93	0.55	0.50
x_{14}	1.00	0.00	1.00	0.91	0.87
x_{15}	0.72	0.13	0.93	0.45	0.65
x_{16}	0.31	0.00	0.87	0.36	0.31
x_{17}	0.02	0.00	0.27	0.00	0.03
x_{18}	0.12	0.00	0.93	0.00	0.15
x_{19}	0.00	0.00	1.00	1.00	0.10
x_{20}	0.14	0.00	0.47	0.00	0.13
y_{21}	6.28	3.72	3.13	10.3	6.05
y_{22}	0.95	0.84	1.89	0.79	1.00
y_{23}	8.44	2.77	22.7	10.7	8.71
n_c	210	32	15	11	268
pop	8706	2702	105060	15924	13678

Last column (“tot”) reports the means in the sample. Last row (“pop”) reports the average population of municipalities belonging to the cluster

practically impossible to build an optic fibers net. Due to this technological barrier, the online services offered are only the basic ones. The percentage of employees devoted to ICT support is comparable with the values in the other groups: this aspect denotes that the absence of front office services is due to the absence of a broad band internet connection and of communication infrastructures as opposed to local political will. In group 4, tourist places are clustered. These municipalities are characterized by the presence of online services that are more useful for tourists rather than citizens. Indeed, this cluster is perfectly homogeneous with respect to x_5 (presence of informative e-mails) and x_{19} (presence of pages written in a foreign language). The percentage of employees with a digital signature is much higher than in the other groups. Cluster 1 is the largest one and, obviously, the least homogeneous. It groups municipalities equipped with a public net able to support most interactive services that have not achieved “excellence” and may improve the opportunities for citizens.

References

1. Chaturvedi, A.D., Green, P.E., Carrol, J.D.: K-modes clustering. J. Classification **18**, 35–55 (2001)
2. Lawrence, C.J., Krzanowski, W.J.: Mixture separation for mixed-mode data. Stat. Comput. **6**, 85–92 (1996)
3. Morlini, I.: A latent variables approach for clustering mixed binary and continuous variables within a Gaussian mixture model. ADAC **6**(1), 5–28 (2012)
4. Pearson, K.: Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. Phil. Trans. Roy. Soc. Lond. Ser. A **195**, 1–47 (1900)
5. 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)
6. Vermunt, J.K., Magidson, J.: Technical Guide for Latent GOLD 4.0: Basic and Advanced. Statistical Innovations Inc., Belmon, MA (2005)