LATENT CLASS CLUSTER ANALYSIS

Jeroen K. Vermunt
Tilburg University
Jay Magidson
Statistical Innovations Inc.

INTRODUCTION

Kaufman and Rousseeuw (1990) define cluster analysis as the classification of similar objects into groups, where the number of groups, as well as their forms are unknown. The "form of a group" refers to the parameters of cluster; that is, to its cluster-specific means, variances, and covariances that also have a geometrical interpretation. A similar definition is given by Everitt (1993) who speaks about deriving a useful division into a number of classes, where both the number of classes and the properties of the classes are to be determined. These could also be definitions of exploratory LC analysis, in which objects are assumed to belong to one of a set of K latent classes, with the number of classes and their sizes not known a priori. In addition, objects belonging to the same class are similar with respect to the observed variables in the sense that their observed scores are assumed to come from the same probability distributions, whose parameters are, however, unknown quantities to be estimated. Because of the similarity between cluster and exploratory LC analysis, it is not surprising that the latter method is becoming a more and more popular clustering tool.

In this paper, we want to describe the state-of-art in the field of LC cluster analysis. Most of the work in this field involves continuous indicators assuming (restricted) multivariate normal distributions within classes. Although authors seldom refer to the work of Gibson (1959) and Lazarsfeld and Henry (1968), actually they are using what these authors called latent profile analysis: that is, latent structure models with a single categorical latent variable and a set of continuous indicators. Wolfe (1970) was the first one who made an explicit connection between LC and cluster analysis.

The last decade there was a renewed interest in the application of LC analysis as a cluster analysis method. Labels that are used to describe such a use of LC analysis are: mixture likelihood approach to clustering (McLachlan and Basford 1988; Everitt 1993), model-based clustering (Banfield and Raftery 1993; Bensmail et. al. 1997; Fraley and Raftery 1998a, 1998b), mixture-model clustering (Jorgensen and Hunt 1996; McLachlan et al. 1999), Bayesian classification (Cheeseman and Stutz 1995), unsupervised learning (McLachlan and Peel 1996), and latent class cluster analysis (Vermunt and Magidson 2000). Probably the most important reason of the increased popularity of LC analysis as a statistical tool for cluster analysis is the fact that nowadays high-speed computers make these computationally intensive methods practically applicable. Several software packages are available for the estimation of LC cluster models.

An important difference between standard cluster analysis techniques and LC clustering is that the latter is a model-based clustering approach. This means that a statistical model is postulated for the population from which the sample under study is coming. More precisely,

it is assumed that the data is generated by a mixture of underlying probability distributions. When using the maximum likelihood method for parameter estimation, the clustering problem involves maximizing a log-likelihood function. This is similar to standard non-hierarchical cluster techniques in which the allocation of objects to clusters should be optimal according some criterion. These criteria typically involve minimizing the within-cluster variation and/or maximizing the between-cluster variation. An advantage of using a statistical model is, however, that the choice of the cluster criterion is less arbitrary. Nevertheless, the log-likelihood functions corresponding to LC cluster models may be similar to the criteria used by certain non-hierarchical cluster techniques like k-means.

LC clustering is very flexible in the sense that both simple and complicated distributional forms can be used for the observed variables within clusters. As in any statistical model, restrictions can be imposed on the parameters to obtain more parsimony and formal tests can be used to check their validity. Another advantage of the model-based clustering approach is that no decisions have to be made about the scaling of the observed variables: for instance, when working with normal distributions with unknown variances, the results will be the same irrespective of whether the variables are normalized or not. This is very different from standard non-hierarchical cluster methods, where scaling is always an issue. Other advantages are that it is relatively easy to deal with variables of mixed measurement levels (different scale types) and that there are more formal criteria to make decisions about the number of clusters and other model features.

LC analysis yields a probabilistic clustering approach. This means that although each object is assumed to belong to one class or cluster, it is taken into account that there is uncertainty about an object's class membership. This makes LC clustering conceptually similar to fuzzy clustering techniques. An important difference between these two approaches is, however, that in fuzzy clustering an object's grades of membership are the "parameters" to be estimated (Kaufman and Rousseeuw 1990) while in LC clustering an individual's posterior class-membership probabilities are computed from the estimated model parameters and his observed scores. This makes it possible to classify other objects belonging to the population from which the sample is taken, which is not possible with standard fuzzy cluster techniques.

The remainder of this paper is organized as follows. The next section discusses the LC cluster model for continuous variables. Subsequently, attention is paid to models for sets of indicators of different measurement levels, also known as mixed-mode data. Then we explain how to include covariates in a LC cluster model. After discussing estimation and testing, two empirical examples are presented. The paper ends with a short discussion. An appendix describes computer programs that implement the various kinds of LC clustering methods presented in this paper.

CONTINUOUS INDICATOR VARIABLES

The basic LC cluster model has the form

$$f(\mathbf{y}_i|\theta) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i|\theta_k)\,.$$

Here, $\mathbf{y}_i$ denotes an object's scores on a set of observed variables, K is the number of clusters, and π_k denotes the prior probability of belonging to latent class or cluster k or, equivalently, the size of cluster k. Alternative labels for the y's are indicators, dependent variables, outcome

variables, outputs, endogenous variables, or items. As can be seen, the distribution of $\mathbf{y}_i$ given the model parameters θ, $f(\mathbf{y}_i|\theta)$, is assumed to be a mixture of classes-specific densities, $f_k(\mathbf{y}_i|\theta_k)$.

Most of the work on LC cluster analysis has been done for continuous variables. Generally, these continuous variables are assumed to be normally distributed within latent classes, possibly after applying an appropriate non-linear transformation (Lazarsfeld and Henry 1968; Basfield and Raftery 1993; McLachlan 1988; McLachlan et. al. 1999; Cheeseman and Stutz 1995). Alternatives for the normal distribution are student, Gompertz, or gamma distributions (see, for instance, McLachlan et. al. 1999).

The most general Gaussian distribution of which all restricted versions discussed below are special cases is the multivariate normal model with parameters μ_k and Σ_k. If no further restrictions are imposed, the LC clustering problem involves estimating a separate set of means, variances, and covariances for each latent class. In most applications, the main objective is finding classes that differ with respect to their means or locations. The fact that the model allows classes to have different variances implies that classes may also differ with respect to the homogeneity of the responses to the observed variables. In standard LC models with categorical variables, it is generally assumed that the observed variables are mutually independent within clusters. This is, however, not necessary here. The fact that each class has its own set of covariances means that the y variables may be correlated with clusters, as well as that these correlations may be cluster specific. So, the clusters do not only differ with respect to their means and variances, but also with respect to the correlations between the observed variables.

It will be clear that as the number of indicators and/or the number of latent classes increases, the number of parameters to be estimated increases rapidly, especially the number of free parameters in the variance-covariance matrices, Σ_k. Therefore, it is not surprising that restrictions which are imposed to obtain more parsimony and stability typically involve constraining the class-specific variance-covariance matrices.

An important constraint model is the local independence model obtained by assuming that all within-cluster covariances are equal to zero or, equivalently, by assuming that the variance-covariance matrices, Σ_k, are diagonal matrices. Models that are less restrictive than the local independence model can be obtained by fixing some but not all covariances to zero or, equivalently, by assuming certain pairs of y's to be mutually dependent within latent classes.

Another interesting type of constraint is the equality or homogeneity of variance-covariance matrices across latent classes, i.e., $\Sigma_k = \Sigma$. Such a homogeneous or class-independent error structure yields clusters having the same forms but different locations. Note that these kinds of equality constraints can be applied in combination with any structure for Σ.

Banfield and Raftery (1993) proposed reparameterizing the class-specific variance-covariance matrices by an eigenvalue decomposition:

$$\Sigma_k = \lambda_k D_k A_k D_k^T \, .$$

The parameter λ_k is a scalar, D_k is a matrix with eigenvectors, and A_k is a diagonal matrix whose elements are proportional to the eigenvalues of Σ_k. More precisely, $\lambda_k = |\Sigma_k|^{1/d}$, where d is the number of observed variables, and A_k is scaled such that $|A_k| = 1$.

A nice feature of the above decomposition is that each of the three sets of parameters has a geometrical interpretation: λ_k indicates what can be called the volume of cluster k, D_k its orientation, and A_k its shape. If we think of a cluster as a clutter of points in a multidimensional space, the volume is the size of the clutter, while the orientation and shape parameters indicate

whether the clutter is spherical or ellipsoidal. Thus, restrictions imposed on these matrices can directly be interpreted in terms of the geometrical form of the clusters. Typically, matrices are assumed to be class-independent and/or simpler structures (diagonal or identity) are used for certain matrices. See Bensmail et al. (1997) and Fraley and Raftery (1998b) for overviews of the many possible specifications.

Rather than by a restricted eigenvalue decomposition, the structure of the Σ_k matrices can also be simplified by means of a covariance-structure model. Several authors have proposed using latent class models for dealing with unobserved heterogeneity in covariance-structure analysis (Arminger and Stein 1997; Dolan and Van der Maas 1997; Jedidi et. al. 1997). The same methodology can be used to restrict the error structure in LC cluster analysis with continuous indicators. An interesting structure for Σ_k, that is related to the eigenvalue decomposition described above, is a factor analytic model (Yung 1997; McLachlan and Peel 1998); that is,

$$\Sigma_k = \Lambda_k \Phi_k \Lambda_k + U_k \,. \tag{1}$$

Here, Λ_k is a matrix with factor loadings, Φ_k is the variance-covariance matrix of the factors, and U_k is a diagonal matrix with unique variances. Restricted versions can be obtained by limiting the number of factors (for instance, to one) and/or fixing some factor loading to zero. Such specifications make it possible to describe the correlations between the y variables within clusters or, equivalently, the structure of local dependencies, by means of a small number of parameters.

MIXED INDICATOR VARIABLES

In the previous section, we concentrated on LC cluster models for continuous indicators assuming a (restricted) multivariate normal distribution for $\mathbf{y}_i$ within each of the classes. Often we are, however, confronted with other types of indicators, like nominal or ordinal variables or counts. LC cluster models for nominal and ordinal variables assuming (restricted) multinomial distributions for the items are equivalent to standard exploratory LC models (Goodman 1974; Clogg 1981, 1995). Böckenholt (1993) and Wedel et. al. (1993) proposed LC models for Poisson counts.

Using the general structure of the LC model, it is straightforward to specify cluster models for sets of indicators of different scale types or, as Everitt (1988, 1993) called it, for mixed-mode data (see also Lawrence and Krzanowski 1996; Jorgensen and Hunt 1996; and Vermunt and Magidson 2000: 147-152). Assuming local independence, the LC cluster model for mixed y's is of the form

$$f(\mathbf{y}_i|\theta) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} f_k(y_{ij}|\theta_{jk}) \,, \tag{2}$$

where J denotes the total number of indicators and j a particular indicator.

Rather than specifying the joint distribution of $\mathbf{y}_i$ given class membership using a single multivariate distribution, we now have to specify the appropriate univariate distribution function for each element y_{ij} of $\mathbf{y}_i$. Possible choices for continuous y_{ij} are univariate normal, student, gamma, and log-normal distributions. A natural choice for discrete nominal or ordinal variables is the (restricted) multinomial distribution. Suitable distributions for counts are, for instance, Poisson, binomial, or negative binomial.

In the above specification, we assumed that the y's are conditional independent within latent classes. This assumption can easily be relaxed by using the appropriate multivariate rather than univariate distributions for sets of locally dependent y variables. It is not necessary to present a separate formula for this situation. We can just think of the index j in equation (2) to denote a set of indicators rather than a single indicator. For sets of continuous variables, we can again work with a multivariate normal distribution. A set of nominal/ordinal variables can combined into a (restricted) joint multinomial distribution. Correlated counts could be modeled with a multivariate Poisson model. More difficult is the specification of the mixed multivariate distributions. Krzanowski (1983) described two possible ways of modeling the relationship between a nominal/ordinal and a continuous y: via a conditional Gaussian or via a conditional multinomial distribution, which means either using the categorical variable as a covariate in the normal model or the continuous one as a covariate in the multinomial model. Lawrence and Krzanowski (1996) and Hunt and Jorgensen (1999) used the conditional Gaussian distribution in LC clustering with combinations of categorical and continuous variables. Local dependencies with a Poisson variable could be dealt with in the same way, i.e., by allowing its mean to dependent on the relevant continuous or categorical variable(s).

The possibility to include local dependencies between indicators is very important when using LC analysis as a clustering tool. First, it prevents that one ends with a solution that contains too many clusters. Often, a simpler solution with less clusters is obtained by including a few direct effects between y variables. It should be stressed that there is also a risk of allowing for within-cluster associations: direct effects may hide relevant clusters.

A second reason for relaxing the local independence assumption is that it may yield a better classification of objects into clusters. Saying that two variables are locally dependent is conceptually the same as saying that they contain some overlapping information that should not be used when determining to which class an object belongs. Consequently, if we omit a significant bivariate dependency from a LC cluster model, the corresponding locally dependent indicators get a too high weight in the classification formula (see equation (3)) compared to the other indicators.

COVARIATES

The LC cluster modeling approach described above is quite general: It deals with mixed-mode data and it allows for many different specification of the (correlated) error structure. An important extension of this model is the inclusion of covariates to predict class membership. Conceptually, it makes very much sense to distinguish (endogenous) variables that serve as indicators of the latent variable from (exogenous) variables that are used to predict to which cluster an object belongs. This idea is, in fact, the same as in Clogg's (1981) LCM with external variables.

Note that in certain situations we may want to use the latent cluster variable as a predictor of an observed response variable rather than as a dependent variable. For such situations, we do not need special arrangements like the ones needed with covariates. A model in which the cluster variable serves as predictor can be obtained by using the response variable as one of the y variables.

Using the same basic structure as in equation (2), this yields the following LC cluster model:

$$f(\mathbf{y}_i|\mathbf{z}_i,\theta) = \sum_{k=1}^{K} \pi_{k|\mathbf{z}_i} \prod_{j=1}^{J} f_k(y_{ij}|\theta_{jk}) \,.$$

Here, $\mathbf{z}_i$ denotes object i's covariate values. Alternative terms for the z's are concomitant variables, grouping variables, external variables, exogenous variables, and inputs. To reduce the number of parameters, the probability of belonging to class k given covariate values $\mathbf{z}_i$, $\pi_{k|\mathbf{z}_i}$, will generally be restricted by a multinomial logit model; that is, a logit model with "linear effects" and no higher order interactions.

An even more general specification is obtained by allowing covariates to have direct effects on the indicators, which yields

$$f(\mathbf{y}_i|\mathbf{z}_i,\theta) = \sum_{k=1}^{K} \pi_{k|\mathbf{z}_i} \prod_{j=1}^{J} f_k(y_{ij}|\mathbf{z}_i,\theta_{jk}) \,.$$

The conditional mean of the y variables can now be directly related to the covariates. This makes it possible to relax the implicit assumption in the previous specification that the influence of the z's on the y's goes completely via the latent variable. For an example, see Vermunt and Magidson (2000: 155).

The possibility to have direct effects of z's on y's can also be used to specify direct effects between indicators of different scale types by means of a simple trick: one of the two variables involved should be used both as covariate (not influencing class membership) and as indicator. We will use this trick below in our second example.

ESTIMATION

The two main methods to estimate the parameters of the various types of LC cluster models are maximum likelihood (ML) and maximum posterior (MAP). Wallace and Dowe (forthcoming) proposed a minimum message length (MML) estimator, which in most situations is similar of MAP. The log-likelihood function required in ML and MAP approaches can be derived from the probability density function defining the model. Bayesian MAP estimation involves maximizing the log-posterior distribution, which is the sum of the log-likelihood function and the logs of the priors for the parameters.

Although generally there is not much difference between ML and MAP estimates, an important advantage of the latter method is that it prevents the occurrence of boundary or terminal solutions: probabilities and variances cannot become zero. With a very small amount of prior information, the parameter estimates are forced to stay within the interior of the parameter space. Typical priors are Dirichlet priors for multinomial probabilities and inverted-Wishart priors for the variance-covariance matrices in multivariate normal models. For more details on these priors see Vermunt and Magidson (2000: 164-165)

Most software packages, use the EM algorithm or some modification of it to find the ML or MAP estimates. In our opinion, the ideal algorithm is starting with a number of EM iterations and when close enough to the final solution, switching to Newton-Raphson. This is a way to combine the advantages of both algorithms, that is, the stability of EM even when far away from the optimum and the speed of Newton-Raphson when close to the optimum.

A well-known problem in LC analysis is the occurrence of local solutions. The best way to prevent ending with a local solution is to use multiple sets of starting values. Some computer programs for LC clustering have automated the search for good starting values using several sets of random starting values, as well as solutions obtained with other cluster methods.

In the application of LC analysis to clustering, we are not only interested in the estimation of the model parameters. Another important "estimation" problem is classification of objects into clusters. This can be based on the posterior class membership probabilities

$$\pi_{k|\mathbf{y}_i,\mathbf{z}_i,} = \frac{\pi_{k|\mathbf{z}_i} \prod_j f_k(y_{ij}|\mathbf{z}_i,\theta_{jk})}{\sum_k \pi_{k|\mathbf{z}_i} \prod_j f_k(y_{ij}|\mathbf{z}_i,\theta_{jk})} \,. \tag{3}$$

The standard classification method is modal allocation, which amounts to assigning each object to the class with the highest posterior probability.

MODEL SELECTION

The model selection issue is one of the main research topics in LC clustering. Actually, there are two issues: the first one concerns the decision about the number of clusters, the second one concerns the form of the model given the number of clusters. For an overview on this topic see Celeux et. al. (1997).

Assumptions with respect to the forms of the clusters given their number can be tested using standard likelihood-ratio tests between nested models, for instance, between a model with an unrestricted covariance matrix and a model with a restricted covariance matrix. Wald tests and Lagrange multiplier tests can be used to assess the significance of certain included or excluded terms, respectively. It is well-known that these kinds of chi-squared tests cannot be used to determine the number of clusters.

The most popular set of model selection tools in LC cluster analysis are information criteria like AIC, BIC, and CAIC (Fraley and Raftery 1998b). The most recent development is the use of computationally intensive techniques like parametric bootstrapping (McLachlan, et. al. 1999) and Markov Chain Monte Carlo methods (Bensmail et. al. 1997) to determine the number of clusters and their forms. Cheeseman and Stutz (1995) proposed a fully automated model selection method using approximate Bayes factors (different from BIC).

Another set of methods for evaluating LC cluster models is based on the uncertainty of classification or, equivalently, the separation of the clusters. Besides the estimated total number of misclassifications, Goodman-Kruskal lambda, Goodman-Kruskal tau, or entropy based measures can be used to indicate how well the indicators predict class membership. Celeux et. al. (1997) described various indices that combine information on model fit and information on classification errors; two of them are the classification likelihood (C) and the approximate weight of evidence (AWE).

TWO EMPIRICAL EXAMPLES

Below LC cluster modeling is illustrated by means of two empirical examples. The analyses are performed with the LCA program Latent GOLD (Vermunt and Magidson, 2000), which implements both ML and MAP estimation with Dirichlet and inverted-Wishart priors for multinomial probabilities and error variance-covariance matrices, respectively. A feature of the program that

was extensively used in the analyses described below is the possibility to add local dependencies using information on bivariate residuals. Model selection was based on BIC, where it should be noted that the BIC we use is computed using the log-likelihood value and the number of parameters rather than using the L^2 value and the number of degrees of freedom.

Diabetes data

The first empirical example concerns a three-dimensional data set involving 145 observations used for diabetes diagnosis (Reaven and Miller 1979). The three continuous variables are labeled glucose (y_1), insuline (y_2), and sspg (y_3). The data set also contains information on the clinical classification in three groups (normal, chemical diabetes, and overt diabetes), which makes it possible to compare the clinical classification with the classification obtained from the cluster model. The substantive question of interest is whether the three indirect diagnostic measures yield a reliable diagnosis; that is, whether they yield a classification that is close to the clinical classification.

This data set comes with the MCLUST program and is also used by Fraley and Raftery (1998a, 1998b) to illustrate their model-based cluster analysis based on the eigenvalue decomposition described in equation (1). The final model they selected on the basis of the BIC criterion was the unrestricted three-class model, which means that none of the restrictions that can be specified with their approach holds for this data set.

We used six different specifications for the variance-covariance matrices: class-dependent and class-independent unrestricted, class-dependent and class-independent diagonal, as well as class-dependent and class-independent with only the y_1-y_2 error covariance free. With unrestricted we that all covariances are free and with diagonal that all covariances are assumed to be zero. The models with only the y_1-y_2 error covariance free were used because the bivariate residuals of both diagonal models indicated that there was only a local dependency between these two variables. Moreover, the results from the unrestricted models indicated that the y_1-y_3 and y_2-y_3 covariances did not differ significantly from zero.

[INSERT TABLE 1 ABOUT HERE]

Table 1 reports the BIC values for the estimated one to five class models. The 3-class model that only includes the error covariance between y_1 and y_2 and with class-dependent variances and covariances has the lowest BIC value. Its BIC value is slightly lower than of the class-dependent unrestricted three-class model, Fraley and Raftery's final model for this data set. The BIC values in table 1 show clearly that models with too restrictive error structures for a particular data set overestimate the number of clusters. Here, this applies to the models with class-independent error variances and the class-dependent diagonal model. Therefore, it is important to be able to work with different types of error structures. Note that the most restrictive model that we used – the model with class-independent diagonal error structure – can be seen as a probabilistic variant of k-means cluster analysis (McLachlan and Basford 1988).

[INSERT TABLE 2 ABOUT HERE]

Table 2 reports the parameters estimates for the three-class model with class-dependent variance-covariance matrices and with only a local dependence between y_1 and y_2. These parameters are the cluster sizes (π_k), the cluster-specific means (μ_{jk}), the cluster-specific variances

(σ^2_{jk}), as well as the cluster-specific covariance between y_1 and y_2 (σ_{12k}). The overt diabetes group (cluster 3), has much higher means on glucose and insuline and a much lower mean on sspg than the normal group (cluster 1). The chemical diabetes group (cluster 2) has somewhat lower means on glucose and insuline and a much lower mean on sspg than the normal group. The reported error variances show that the overt diabetes cluster is much more heterogeneous with respect to glucose and insuline and much more homogeneous with respect to sspg than the normal cluster. The chemical diabetes group is the most homogeneous cluster on all three measures. The error covariances are somewhat easier to interpret if we transform them to correlations. Their values are .69, .21, and .93 for cluster 1, 2 and 3, respectively. This indicates that in the overt diabetes group there is a very strong association between glucose and insuline, while in the chemical diabetes group this association is very low, and even not significantly different from zero ($\widehat{\sigma}_{12k}/\mathrm{SE}_{\widehat{\sigma}_{12k}} = 1.60$). Note that the within-cluster correlation of .93 is very high, which indicates that, in fact, the two measures are equivalent in cluster 3.

[INSERT TABLE 3 ABOUT HERE]

Not only the BIC of our final model is somewhat better than Fraley and Raftery's, also our classification is more in agreement with the clinical classification: our model "misclassifies" 13.1 percent of the patients while the unrestricted models misclassifies 14.5 percent. Table 3 reports the cross-tabulation of the clinical and the LC cluster classification based on the posterior class-membership probabilities. As can be seen, some normal patients are classified as cases with chemical diabetes and vice versa. The other type of error is that some overt diabetes cases are classified as normal.

Prostate cancer data

Our second example concerns the analysis of a mixed-mode data set with pre-trial covariates from a prostate cancer clinical trial. Jorgensen and Hunt (1996) and Hunt and Jorgensen (1999) used this data set containing information on 506 patients to illustrate the use of the LC cluster model implemented in their MULTIMIX program. The eight continuous indicators are age (y_1), weight index (y_2), systolic blood pressure (y_5), diastolic blood pressure (y_6), serum haemoglobin (y_8), size of primary tumor (y_9), index of tumor stage and histolic grade (y_{10}), and serum prostatic acid phosphatase (y_{11}). The four categorical observed variables are performance rating (y_3; 4 levels), cardiovascular disease history (y_4; 2 levels), electrocardiogram code (y_7, 7 levels), and bone metastases (y_{12}, 2 levels). The research question of interest is whether on the basis of these pre-trial covariates it is possible to identify subgroups that differ with respect to the likelihood of success of the medical treatment of prostate cancer.

The categorical variables are treated as nominal and for the continuous variables we assumed normal distributions with class-specific variances. We estimated models from one to four latent classes. The first model for each number of classes assumes local independence. The other four specifications are obtained by subsequently adding the direct relationships between y_5 and y_6, y_2 and y_8, y_8 and y_{12}, and y_{11} and y_{12}. This exploratory improvement of the model fit was guided by Latent GOLD's bivariate residuals information, as well as the results reported by Hunt and Jorgensen (1999).

To give an indication about the computation time needed for these kinds of models: all two-class models took less than 5 seconds to converge and all four class models less than 20 seconds on a Pentium II 350 Mhz. Note that here we have a data set with almost 500 cases

and 12 indicators. The estimation time increases linearly with the number of cases and, as long as we do not include too many local dependencies, also almost linearly with the number of indicators.

[INSERT TABLE 4 ABOUT HERE]

Table 4 presents the BIC values for the estimated models. As can be seen, the two-class model that includes all four direct relationships has the lowest BIC. Comparison of the various models given a certain number of classes shows that inclusion of the direct relationship between y_5 and y_6 (the two blood pressure measures) improves the fit in all situations. The other bivariate terms improve the fit in the one-, two-, and three-class models, but not in the four-class model. If we compare the models with different number of classes for a given error structure, the four-class model performs best when assuming local independence, the three-class model when including the y_5 and y_6 covariance, and the two-class model when including additional bivariate terms. Thus, if we are willing to include the y_5-y_6 effect, a model with no more than three classes should be selected. If we are willing to include more direct effects, the two-class model is the preferred one. This shows again that the possibility to work with more local dependencies may yield a simpler final model.

[INSERT TABLE 5 ABOUT HERE]

Table 5 reports the parameters estimates for the two-class model containing all four direct effects. Wald tests for the difference of the means and probabilities between classes indicate that only the mean ages (μ_{1k}) are not significantly different between classes. Cluster 2 turns out to have somewhat higher means on weight (μ_{2k}), blood pressure (μ_{5k} and μ_{6k}), and serum haemoglobin (μ_{8k}), and lower means on size of tumor (μ_{9k}), index of tumor stage (μ_{10k}), and serum prostatic acid phosphatase (μ_{11k}). If we look at the nominal indicators, we see a large difference between the two classes in the distribution of bone metastases (y_{12}), somewhat smaller differences in performance rating (y_3) and cardiovascular disease history (y_4), and a very small difference in electrocardiogram code (y_7). The direct effects between the indicators are quite strong. They all have a positive sign except for the effect of y_{12} on y_{11}.

To investigate the usefulness of the applied technique, Jorgensen and Hunt (1996) and Hunt and Jorgensen (1999) investigated the strength of the relationship between the obtained classification and the outcome of the medical trial. They showed that their two-class solution, which is similar to the two-class model with local dependencies obtained here, predicted very well the success of the medical treatment.

CONCLUSIONS

This paper described the state-of-art in the field of cluster analysis using LC models. Two important recent developments are the possibility to use various kinds of meaningful restrictions on the covariance structure in mixtures of multivariate normal distributions and the possibility to work with mixed-mode data.

The first example demonstrated the use of different types of specifications for the covariance structure. It showed that too restrictive models may yield too many latent classes. The second example illustrated LC clustering with mixed-mode data using models with and without local dependencies.

REFERENCES

Arminger, G., and Stein, P. 1997. "Finite mixture of covariance structure models with regressors: loglikehood function, distance estimation, fit indices, and a complex example." *Sociological Methods and Research* 26: 148-182.

Banfield, J.D., and Raftery, A.E. 1993. "Model-based Gaussian and non-Gaussian clustering." *Biometrics* 49: 803-821.

Bensmail, H., Celeux, G., Raftery, A.E., and Robert, C.P. 1997. "Inference in model based clustering." *Statistics and Computing* 7: 1-10.

Byar, D.P., and Green, S.B. 1980. "The choice of treatment for cancer patients based on covariate information: Application to prostate cancer." *Bulletin of Cancer* 67: 477-490.

Böckenholt, U. 1993. "A latent class regression approach for the analysis of recurrent choices." *British Journal of Mathematical and Statistical Psychology* 46: 95-118.

Celeux, G., Biernacki, C., and Govaert, G. 1997. *Choosing models in model-based clustering and discriminant analysis.* Technical Report. Rhone-Alpes: INRIA.

Cheeseman, P., and Stutz, J. 1995. "Bayesian classification (Autoclass): Theory and results." In *Advances in knowledge discovery and data mining.* edited by U.M.Fayyad, G.Piatetsky-Shapiro, P.Smyth and R.Uthurusamy. Menlo Park: The AAAI Press.

Clogg, C.C. 1981. "New developments in latent structure analysis." Pp. 215-246, in *Factor analysis and measurement in sociological research*, edited by D.J. Jackson and E.F. Borgotta. Beverly Hills: Sage Publications.

Clogg, C.C. 1995. "Latent class models." Pp. 311-359, in *Handbook of statistical modeling for the social and behavioral sciences*, edited by G.Arminger, C.C.Clogg, and M.E.Sobel. New York: Plenum Press.

Dolan, C.V., and Van der Maas, H.L.J. 1997. "Fitting multivariate normal finite mixtures subject to structural equation modeling." *Psychometrika* 63: 227-253.

Everitt, B.S. 1988. "A finite mixture model for the clustering of mixed-mode data." *Statistics and Probability Letters* 6: 305-309.

Everitt, B.S. 1993), *Cluster analysis.* London: Edward Arnold.

Fraley, C., and Raftery, A.E. 1998a. *MCLUST: Software for model-based cluster and discriminant analysis.* Department of Statistics, University of Washington: Technical Report No. 342.

Fraley, C., and Raftery, A.E. 1998b. *How many clusters? Which clustering method? - Answers via model-based cluster analysis.* Department of Statistics, University of Washington: Technical Report no. 329.

Gibson, W.A. 1959. "Three multivariate models: Factor analysis, latent structure analysis, and latent profile analysis." *Psychometrika* 24: 229-252.

Goodman, L.A. 1974. "Exploratory latent structure analysis using both identifiabe and unidentifiable models." *Biometrika* 61: 215-231.

Hunt, L, and Jorgensen, M. 1999. "Mixture model clustering using the MULTIMIX program." *Australian and New Zeeland Journal of Statistics* 41: 153-172.

Jedidi, K., Jagpal, H.S., and DeSarbo, W.S. 1997. "Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity." *Marketing Science* 16: 39-59.

Jorgensen, M., and Hunt, L. 1996. "Mixture model clustering of data sets with categorical and continuous variables." Pp 375-384, in *Proceedings of the Conference ISIS '96, Australia 1996*.

Kaufman, L., and Rousseeuw, P.J. 1990. *Finding groups in data: An introduction to cluster analysis*. New York: John Wiley and Sons, Inc..

Krzanowski, W.J. 1983. "Distance between populations using mixed continuous and categorical variables." *Biometrika* 70: 235-243.

Lawrence C.J., Krzanowski, W.J. 1996. "Mixture separation for mixed-mode data." *Statistics and Computing* 6: 85-92.

Lazarsfeld, P.F., and Henry, N.W. 1968. *Latent structure analysis*. Boston: Houghton Mill.

McLachlan, G.J., and Basford, K.E. 1988. *Mixture models: inference and application to clustering*. New York: Marcel Dekker.

McLachlan, G.J., and Peel, D. 1996. "An algorithm for unsupervised learning via normal mixture models." Pp. 354-363, in *Information, statistics and induction in science*, edited by D.L.Dowe, K.B.Korb, and J.J.Oliver. Singapore: World Scientific Publishing.

McLachlan, G.J., and Peel, D. 1999. *Modelling nonlinearity by mixtures of factor analysers via extension of the EM algorithm*. Technical Report. Australia: Center for Statistics, University of Queensland.

McLachlan, G.J., Peel, D., Basford, K.E., and Adams, P. 1999. "The EMMIX software for the fitting of mixtures of normal and t-components." *Journal of Statistical Software* 4, No. 2.

Moustaki, I. 1996. "A latent trait and a latent class model for mixed observed variables." *The British Journal of Mathematical and Statistical Psychology* 49: 313-334.

Muthen, B., and Muthen, L., 1998. *Mplus: User's manual*. Los Angeles: Muthen and Muthen.

Reaven, G.M., and Miller, R.G. 1979. "An attempt to define the nature of chemical diabetes using multidimensional analysis." *Diabetologia* 16: 17-24.

Vermunt, J.K. 1997. *LEM: A general program for the analysis of categorical data. User's manual.* Tilburg University, The Netherlands.

Vermunt, J.K., and Magidson, J. 2000. *Latent GOLD's User's Guide*. Boston: Statistical Innovations Inc..

Wallace, C.S., and Dowe, D.L. Forthcoming. "MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions." *Statistics and Computing*.

Wedel, M., DeSarbo, W.S., Bult, J.R., and Ramaswamy, V. 1993. "A latent class Poisson regression model for heterogeneous count data with an application to direct mail." *Journal of Applied Econometrics* 8: 397-411.

Wolfe, J.H. 1970. "Pattern clustering by mulltivariate cluster analysis." *Multivariate Behavioral Research* 5: 329-350.

Yung, Y.F. 1997. "Finite mixtures in confirmatory factor-analysis models." *Psychometrika* 62: 297-330.

Table 1: BIC values for diabetes example

	Number of clusters				
Model	1	2	3	4	5
1. Class-dependent unrestricted Σ_k	5138	4819	4762	4788	4818
2. Class-independent unrestricted Σ_k	5138	5014	4923	4869	4858
3. Class-dependent diagonal Σ_k	5530	4957	4833	4805	4815
4. Class-independent diagonal Σ_k	5530	5170	4999	4938	4895
5. Class-dependent Σ_k with only σ_{12k} free	5156	4835	4756	4761	4784
6. Class-independent Σ_k with only σ_{12k} free	5156	5008	4920	4862	4859

Table 2: Parameter estimates for diabetes example

	Cluster					
	1= Normal		2 = Chemical		3 = Overt	
Parameter	Estimate	S.E.	Estimate	S.E.	estimate	S.E.
π_k	0.27	0.05	0.54	0.05	0.19	0.03
μ_{1k}	104.00	2.85	91.23	1.06	234.76	14.87
μ_{2k}	495.06	22.74	359.22	6.63	1121.09	58.70
μ_{3k}	309.43	28.06	163.13	6.37	76.98	9.47
σ^2_{1k}	230.09	62.96	76.48	12.93	5005.91	1414.43
σ^2_{2k}	14844.55	3708.65	2669.75	506.55	73551.09	22176.29
σ^2_{3k}	22966.52	5395.90	2421.45	476.65	2224.50	616.43
σ_{12k}	1279.92	420.93	96.46	60.30	17910.71	5423.37

Table 3: Clinical versus LC cluster classification in diabetes example

Clinical	LC cluster classification			
classification	normal	chemical	overt	total
normal	26	10	0	36
chemical	4	72	0	76
overt	5	0	28	33
total	35	82	28	145

Table 4: BIC values for cancer example

	Number of clusters			
Model	1	2	3	4
1. Local independence	23762	23112	23089	23088
2. Model 1 + σ_{56k}	23529	22889	22883	22887
3. Model 2 + σ_{28k}	23502	22872	22875	22893
4. Model 3 + $\beta_{8.12}$	23473	22861	22866	22895
5. Model 4 + $\beta_{11.12}$	23322	22845	22855	22888

Table 5: Parameter estimates for prostate cancer example

	Cluster 1		Cluster 2	
Parameter	Estimate	S.E.	Estimate	S.E.
π_k	0.45	0.03	0.55	0.03
μ_{1k}	71.38	0.51	71.70	0.43
μ_{2k}	97.51	0.98	100.26	0.83
$\pi_{1,3k}$	0.85	0.02	0.94	0.02
$\pi_{2,3k}$	0.09	0.02	0.05	0.01
$\pi_{3,3k}$	0.05	0.02	0.01	0.01
$\pi_{4,3k}$	0.01	0.01	0.00	0.00
$\pi_{1,4k}$	0.65	0.03	0.49	0.03
$\pi_{2,4k}$	0.35	0.03	0.51	0.03
μ_{5k}	14.18	0.16	14.54	0.16
μ_{6k}	8.00	0.09	8.29	0.10
$\pi_{1,7k}$	0.35	0.03	0.33	0.030
$\pi_{2,7k}$	0.05	0.02	0.05	0.01
$\pi_{3,7k}$	0.14	0.02	0.07	0.02
$\pi_{4,7k}$	0.04	0.01	0.06	0.02
$\pi_{5,7k}$	0.30	0.03	0.31	0.03
$\pi_{6,7k}$	0.12	0.02	0.17	0.02
$\pi_{7,7k}$	0.00	0.00	0.00	0.00
μ_{8k}	128.01	1.38	132.21	1.80
μ_{9k}	4.11	0.12	2.88	0.08
μ_{10k}	12.02	0.11	8.88	0.08
μ_{11k}	4.00	0.12	2.11	0.11
$\pi_{1,12k}$	0.65	0.03	0.99	0.01
$\pi_{2,12k}$	0.35	0.03	0.01	0.01
σ^2_{1k}	52.35	5.36	43.97	4.15
σ^2_{2k}	186.60	19.82	166.73	15.89
σ^2_{5k}	4.98	0.50	6.60	0.59
σ^2_{6k}	1.79	0.18	2.40	0.21
σ^2_{8k}	355.82	35.44	325.52	29.47
σ^2_{9k}	2.91	0.29	1.40	0.14
σ^2_{10k}	2.05	0.21	1.25	0.13
σ^2_{11k}	2.56	0.25	0.25	0.03
σ_{28k}	61.98	19.14	47.56	15.12
σ_{56k}	1.82	0.25	2.52	0.30
$\beta_{8.12}$	5.76	1.35	5.76	1.35
$\beta_{11.12}$	-0.49	0.11	-0.49	0.11

Table 6: Computer programs and their most important features

Name	Multivar. normal	Mixed-mode	Covar.	Estimation method	Algorithm	System / source
NORMIX	yes	no	no	ML	EM	DOS
EMMIX	yes	no	no	ML	EM	DOS + Fortran code
MCLUST	yes	no	no	ML	EM	S-plus
LEM	no	yes	yes	ML	EM +NR	DOS + Windows
Classmix	no	yes	no	ML	EM	unknown
Autoclass	yes	yes	no	MAP	EM	DOS + C code
MULTIMIX	yes	yes	no	ML	EM	Fortran code
Mplus	yes	yes[1]	yes	ML	EM	DOS
Latent GOLD	yes	yes	yes	ML + MAP	EM + NR	Windows

1. In MPLUS, categorical indicators must be dichotomous

SOFTWARE

Several computer programs are available for estimating the various types of LC cluster models discussed in this paper. Table 6 lists the most important packages and gives information on the types of cluster models they implement (multivariate normal distributions and/or mixed-mode data); whether they allow users to include covariates in the model; the estimation method they use; the algorithm (EM or NR=Newton-Raphson) they use; and the system for with an executable version and/or the type of source code that is available.

[INSERT TABLE 6 ABOUT HERE]

We will not repeat all the information listed in table 6 but describe the main special features of some of the programs. NORMIX (Wolfe, 1970), EMMIX (McLachlan et. al., 1999), and MCLUST (Fraley and Raftery, 1998a) are programs for LC clustering with continuous variables using multivariate normal distributions. Special features of EMMIX are that it uses of multiple sets of starting values to prevent local solutions and that it performs likelihood-ratio tests for the number of clusters using parametric bootstrapping. MCLUST allows users to restrict the class-specific variance-covariance matrices using the eigenvalue decomposition described in equation (1).

LEM (Vermunt, 1997) and Classmix (Moustaki, 1996) are LC analysis programs that can be used for clustering with mixed-mode data. LEM cannot only deal with (ordinal) categorical and continuous variables, but also with Poisson counts. In LEM, it is possible to include local dependencies between categorical variables.

MULTIMIX (Hunt and Jorgensen, 1999), Mplus (Muthen and Muthen, 1998), Autoclass (Cheeseman and Stutz, 1995), and Latent GOLD (Vermunt and Magidson, 2000) can deal with multivariate normal distributions, as well as with mixed-mode data. MULTIMIX allows users to specifying local dependencies between categorical and continuous variables using conditional Gaussian distributions. Both Mplus and Latent GOLD are very flexible with respect to the specification of the structure of the error-covariance matrices: any covariance can be included or excluded from the model. Two weak points of Mplus are that the categorical variables should be dichotomous and that the user has to provide starting values for all parameters. Autoclass is a program that has automatized model selection using multiple sets of starting values (also for the number of classes). Latent GOLD is the only fully Windows based program, which make it very easy to use. Like LEM, it cannot only deal with (ordinal) categorical and continuous variables, but also with Poisson counts. Its multiple sets of random starting values help users to prevent ending with a local solution and its bivariate residual measures make it easy to detect local dependencies to be included in the model.

SYMBOLS

K	number of classes or clusters
J	number of indicator variables
i	index to denote a particular case
j	index to denote a particular indicator variable
k	index to denote a particular class or cluster
$\mathbf{y}$	vector of indicator variables
y	value of an indicator variable
$\mathbf{z}$	covariate vector
$f(..)$	density function
π	probability
θ	parameter vector
μ	mean vector
Σ	variance-covariance matrix
σ_j^2	variance of variable j
$\sigma_{j\ell}$	covariance between variables j and ℓ

FURTHER READING

Further reading on cluster analysis by means of latent class or finite mixture models can be done with McLachlan and Basford (1988) and Everitt (1993).