Structural Equation Models: Mixture Models

Jeroen K. Vermunt
Department of Methodology and Statistics
Tilburg University

Jay Magidson
Statistical Innovations Inc.

1 Introduction

This article discusses a modelling framework that links two well-known statistical methods: structural equation modelling (SEM) and latent class or finite mixture modelling. This hybrid approach was proposed independently by Arminger and Stein [1], Dolan and Van der Maas [4], and Jedidi, Jagpal and DeSarbo [5]. Here, we refer to this approach as mixture SEM or latent class SEM.

There are two different ways to view mixture SEM. One way is as a refinement of multivariate normal (MVN) mixtures, where the within-class covariance matrices are smoothed according to a postulated SEM structure. MVN mixtures have become a popular tool for cluster analysis [6] [10], where each cluster corresponds to a latent (unobservable) class. Names that are used when referring to such a use of mixture models are latent profile analysis, mixture-model clustering, model-based clustering, probabilistic clustering, Bayesian classification, and latent class clustering. Mixture SEM restricts the form of such latent class clustering, by subjecting the class-specific mean vectors and covariance matrices to a postulated SEM structure such as a one-factor, a latent-growth, or an autoregressive model. This results in MVN mixtures that are more parsimonious and stable than models with unrestricted covariance structures.

The other way to look at mixture SEM is as an extension to standard SEM similar to multiple group analysis. However, an important difference between this and standard multiple group analysis is that in mixture SEM group membership is not observed. By incorporating latent classes into a SEM model, various forms of unobserved heterogeneity can be detected. For example, groups that have identical (unstandardized) factor loadings but different error variances on the items in a factor analysis or groups that show different patterns of change over time. Dolan and Van der Maas [4] describe a nice application from developmental psychology in which as a result of the existence of qualitative development stages, children that do not master a certain type of tasks have a mean and covariance structure that differs from the one for children that master the tasks.

Below, we first introduce standard MVN mixtures. Then, we show how the SEM framework can be used to restrict the means and covariances. Subsequently, we discuss parameter estimation, model testing, and software. We end with an empirical example.

2 Multivariate normal mixtures

Let $\mathbf{y}_i$ denote a P-dimensional vector containing the scores for individual i on a set of P observed continuous random variables. Moreover, let K be the number of mixture components, latent classes, or clusters, and π_k the prior probability of belonging to latent class or cluster k or, equivalently, the size of cluster k, where $1 \leq k \leq K$. In a mixture model, it is assumed that the density of $\mathbf{y}_i$, $f(\mathbf{y}_i|\boldsymbol{\theta})$, is a mixture or a weighted sum of K class-specific densities $f_k(\mathbf{y}_i|\boldsymbol{\theta}_k)$ [4] [10]. That is,

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

Here, $\boldsymbol{\theta}$ denotes the vector containing all unknown parameters and $\boldsymbol{\theta}_k$ the vector of the unknown parameters of cluster k.

The most common specification for the class-specific densities $f_k(\mathbf{y}_i|\boldsymbol{\theta}_k)$ is multivariate normal, which means that the observed variables are assumed to be normally distributed within latent classes, possibly after applying an appropriate non-linear transformation. Denoting the class-specific mean vector by $\boldsymbol{\mu}_k$ and the class-specific covariance matrix by $\boldsymbol{\Sigma}_k$, we obtain the following class specific densities:

$$f_k(\mathbf{y}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = (2\pi)^{-P/2}|\boldsymbol{\Sigma}_k|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu}_k)' \, \boldsymbol{\Sigma}_k^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_k)\right\}.$$

In the most general specification, no restrictions are imposed on $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ parameters; that is, the model-based clustering problem involves estimating a separate set of means, variances, and covariances for each latent class.

Although in most clustering applications the main objective is finding classes that differ with respect to their means or locations, in the MVN mixture model clusters may also have different shapes.

An unrestricted MVN mixture model with K latent classes contains $(K-1)$ unknown class sizes, $K \cdot P$ class-specific means, $K \cdot P$ class-specific variances and $K \cdot P \cdot (P-1)/2$ class-specific covariances. As the number of indicators and/or the number of latent classes increases, the number of parameters to be estimated may become quite large, especially the number of free parameters in $\mathbf{\Sigma}_k$. Thus, to obtain more parsimony and stability, it is not surprising that restrictions are typically imposed on the class-specific covariance matrices.

Prior to using SEM models to restrict the covariances, a standard approach to reduce the number of parameters is to assume local independence. Local independence means that all within-cluster covariances are equal to zero or, equivalently, that the covariance matrices, $\mathbf{\Sigma}_k$, 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 approach to reduce the number of parameters is to assume the equality or homogeneity of variance-covariance matrices across latent classes; i.e., $\mathbf{\Sigma}_k = \mathbf{\Sigma}$. Such a homogeneous or class-independent error structure yields clusters having the same forms but different locations. This type of constraint is equivalent to the restrictions applied to the covariances in linear discriminant analysis. Note that this between-class equality constraint can be applied in combination with any structure for $\mathbf{\Sigma}$.

Banfield and Raftery [2] proposed reparameterizing the class-specific covariance matrices by an eigenvalue decomposition:

$$\mathbf{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k' \,.$$

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

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, $\mathbf{D}_k$ its orientation, and $\mathbf{A}_k$ 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. Typical restriction are to assume matrices to be equal across classes or to have the forms of diagonal or identity matrices [3].

3 Mixture SEM

As an alternative to simplifying the $\mathbf{\Sigma}_k$ matrices using the eigenvalue decomposition, the mixture SEM approach assumes a covariance-structure model. Several authors [1] [4] [5] have proposed using such a mixture specification for dealing with unobserved heterogeneity in SEM. As explained in the introduction, this is equivalent to restricting the within-class mean vectors and covariance matrices by a SEM. One interesting SEM structure for $\mathbf{\Sigma}_k$ that is closely related to the eigenvalue decomposition described above is a factor-analytic model [6] [11]. Under the factor-analytic structure, the within-class covariances are given by:

$$\mathbf{\Sigma}_k = \mathbf{\Lambda}_k \mathbf{\Phi}_k \mathbf{\Lambda}'_k + \mathbf{\Theta}_k \, .$$

Assuming that there are Q factors, $\mathbf{\Lambda}_k$ is a $P \times Q$ matrix with factor loadings, $\mathbf{\Phi}_k$ is a $Q \times Q$ matrix containing the variances of and the covariances between the factors, and $\mathbf{\Theta}_k$ is a $P \times P$ diagonal matrix containing the unique variances. Restricted covariance structures are obtained by setting $Q < P$ (for instance, $Q = 1$), equating factor loadings across indicators, or fixing some factor loading to zero. Such specifications make it possible to describe the covariances between the y variables within clusters by means of a small number of parameters.

Alternative formulations can be used to define more general types of SEM models. Here, we use the Lisrel submodel that was also used by Dolan and Van der Maas [4]. Other alternatives are the full Lisrel [5], the RAM [8], or the conditional mean and covariance structure [1] formulations.

In our Lisrel submodel formulation, the SEM for class k consists of following two (sets of) equations:

$$\begin{aligned} \mathbf{y}_i &= \boldsymbol{\nu}_k + \mathbf{\Lambda}_k \boldsymbol{\eta}_{ik} + \boldsymbol{\varepsilon}_{ik} \\ \boldsymbol{\eta}_{ik} &= \boldsymbol{\alpha}_k + \mathbf{B}_k \boldsymbol{\eta}_{ik} + \boldsymbol{\varsigma}_{ik}. \end{aligned}$$

The first equation concerns the measurement part of the model in which the observed variables are regressed on the latent factors $\boldsymbol{\eta}_{ik}$. Here, $\boldsymbol{\nu}_k$ is a vector of intercepts, $\mathbf{\Lambda}_k$ a matrix with factor loadings and $\boldsymbol{\varepsilon}_{ik}$ a vector with residuals. The second equation is the structural part of the model, the path model for the factors. Vector $\boldsymbol{\alpha}_k$ contains the intercepts, matrix $\mathbf{B}_k$ the path coefficients and vector $\boldsymbol{\varsigma}_{ik}$ the residuals. The implied mean and covariance structures for latent class k are

$$\begin{aligned}
\boldsymbol{\mu}_k &= \boldsymbol{\nu}_k + \mathbf{\Lambda}_k(\mathbf{I}-\mathbf{B}_k)^{-1}\boldsymbol{\alpha}_k \\
\mathbf{\Sigma}_k &= \mathbf{\Lambda}_k(\mathbf{I}-\mathbf{B}_k)^{-1}\mathbf{\Phi}_k(\mathbf{I}-\mathbf{B}_k')^{-1}\mathbf{\Lambda}_k' + \mathbf{\Theta}_k,
\end{aligned}$$

where $\mathbf{\Theta}_k$ and $\mathbf{\Phi}_k$ denote the covariance matrices of the residuals $\boldsymbol{\varepsilon}_{ik}$ and $\boldsymbol{\varsigma}_{ik}$. These equations show the connection between the SEM parameters and the parameters of the MVN mixture model.

4 Covariates

An important extension of the mixture SEM described above is obtained by including covariates to predict class membership, with possible direct effects on the item means. Conceptually, it makes sense to distinguish (endogenous) variables that are used to identify the latent classes from (exogenous) variables that are used to predict to which cluster an individual belongs.

Using the same basic structure as in Equation 1, this yields the following mixture model:

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

Here, $\mathbf{z}_i$ denotes person 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

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

The conditional means of the y variables are now directly related to the covariates as proposed by Arminger and Stein [1]. 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 classes (see, for example, [9]).

5 Estimation, testing and software

Estimation

The two main estimation methods in mixture SEM and other types of MVN mixture modelling are maximum likelihood (ML) and maximum posterior (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 the latent class probabilities and inverted-Wishart priors for the covariance matrices. For more details on these priors, see Vermunt and Magidson [9].

Most mixture modelling software packages use the EM algorithm or some modification of it to find the ML or MAP estimates. In our opinion, the ideal algorithm starts with a number of EM iterations and when close enough to the final solution, switches to Newton-Raphson. This is a way to combine the advantages of both algorithms – 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 mixture modelling 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 mixture modelling have automated the search for good starting values using several sets of random starting values.

When using mixture SEM for clustering, we are not only interested in the estimation of the model parameters, but also in the classification of individual into clusters. This can be based on the posterior class membership probabilities

$$\pi_k(\mathbf{y}_i, \mathbf{z}_i, \boldsymbol{\pi}, \boldsymbol{\theta}) = \frac{\pi_k(\mathbf{z}_i)\, f_k(\mathbf{y}_i|\mathbf{z}_i, \boldsymbol{\theta}_k)}{\sum_{k=1}^{K} \pi_k(\mathbf{z}_i)\, f_k(\mathbf{y}_i|\mathbf{z}_i, \boldsymbol{\theta}_k)} \, .$$

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 mixture-model clustering. Actually, there are two issues involved: the first concerns the decision about the number of clusters, the second concerns the form of the model given the number of clusters. For an extended overview on these topics, see McLachlan and Peel [6].

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. However, these kinds of chi-squared tests can not be used to determine the number of clusters.

The approach most often used for model selection in mixture modelling is to use information criteria, such as AIC, BIC, and CAIC. The most recent development is the use of computationally intensive techniques like parametric bootstrapping [6] and Markov Chain Monte Carlo methods [3] to determine the number of clusters, as well as their forms.

Another approach for evaluating mixture 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.

Software

Several computer programs are available for estimating the various types of mixture models discussed in this paper. Mplus [7] and Mx [8] are syntax-based programs that can deal with a very general class of mixture SEMs. Mx is somewhat more general in terms of model possible constraints. Latent

GOLD [9] is a fully Windows based program for estimating MVN mixtures with covariates. It can be used to specify restricted covariance structures, including a number of SEM structures such as a one-factor model within blocks of variables and a compound-symmetry (or random-effects) structure.

6 An empirical example

To illustrate mixture SEM we use a longitudinal data set made available by Patrick J. Curran at "http://www.duke.edu/ curran/". The variable of interest is a child's reading recognition skill measured at 4 two-year intervals using the Peabody Individual Achievement Test (PIAT) Reading Recognition subtest. The research question of interest is whether a one-class model with its implicit assumption that a single pattern of reading development holds universally is correct, or whether there are different types of reading recognition trajectories among different latent groups. Besides information on reading recognition, we have information on the child's gender, the mother's age, the child's age, the child's cognitive stimulation at home, and the child's emotional support at home. These variables will be used as covariates. The total sample size is 405, but only 233 children were measured at all assessments. We use all 405 cases in our analysis assuming that the missing data is missing at random (MAR). For parameter estimation, we used the Latent GOLD and Mx programs.

One- to three-class models (without covariates) were estimated under five types of SEM structures fitted to the within-class covariance matrices. These SEM structures are local independence (LI), saturated (SA), random effects (RE), autoregressive (AR), and one factor (FA). The BIC values reported in Table 1 indicate that two classes are needed when using a SA, AR, or FA structure.[1] As is typically the case, working with a misspecified covariance structure (here, LI or RE), yields an overestimation of the number of classes. Based on the BIC criterion, the two-class AR model (Model D2) is the model that is preferred. Note that this model captures the dependence between the time-specific measures with a single path coefficient since the coefficients associated with the autoregressive component of the model is assumed to be equal for each pair of adjacent time points.

[1] BIC is defined as minus twice the log-likelihood plus $\ln(N)$ times the number of parameters, where N is the sample size (here 450).

Subsequently, we included the covariates in the model. Child's age was assumed to directly affect the indicators in order to assure that the encountered trajectories are independent of the child's age at the first occasion. Child's gender, mother's age, child's cognitive stimulation, and child's emotional support were assumed to affect class membership. According to the BIC criterion, this model (Model F) is much better than the model without covariates (Model D2).

According to Model F, Class 1 contains 61 and class 2 39 percent of the children. The estimated means for class 1 are 2.21, 3.59, 4.51, and 5.22, and for class 2 3.00, 4.80, 5.81, and 6.67. These results show that class 2 starts at a higher level and grows somewhat faster than class 1. The estimates of the class-specific variances are 0.15, 0.62, 0.90 and 1.31 for class 1, and 0.87, 0.79, 0.94, and 0.76 for class 2. This indicates that the within-class heterogeneity increases dramatically within class 1 while it is quite stable within class 2. The estimated values of the class-specific path coefficients are 1.05 and 0.43, respectively, indicating that even with the incrementing variance the autocorrelation is larger in latent class 1 than in latent class 2.[2]

The age effects on the indicators are highly significant. As far as the covariate effects on the log-odds of belonging to class 2 instead of class 1 is concerned, only the mother's age is significant. The older the mother, the higher the probability of belonging to latent class 2.

References

[1] 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.

[2] Banfield, J.D., and Raftery, A.E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics,* 49, 803-821.

[3] Bensmail, H., Celeux, G., Raftery, A.E., and Robert, C.P. (1997). Inference in model based clustering. *Statistics and Computing,* 7, 1-10.

[2] The autocorrelation is a standardized path coefficient that can be obtained as the product of the unstandardized coefficient and the ratio of the standard deviations of the independent and the dependent variable in the equation concerned. For example, the class 1 autocorrelation between time points 1 and 2 equals $1.05\frac{\sqrt{0.15}}{\sqrt{0.62}}$.

[4] 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.

[5] 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.

[6] McLachlan, G.J., and Peel, D. (2000). *Finite Mixture Models.* New York: John Wiley & Sons.

[7] Muthén, B., and Muthén, L., (1998). *Mplus: User's manual.* Los Angeles: Muthén & Muthén.

[8] Neale, M.C., Boker, S.M., Xie, G., and Maes, H.H. (2002). *Mx: Statistical Modeling.* Richmond, VA: VCU, Department of Psychiatry.

[9] Vermunt, J.K., and Magidson, J. (2000). *Latent GOLD's User's Guide.* Boston: Statistical Innovations.

[10] Vermunt, J.K., and Magidson, J. (2002). Latent class cluster analysis. J.A. Hagenaars and A.L. McCutcheon (eds.), Applied latent class analysis, 89-106. Cambridge: Cambridge University Press.

[11] Yung, Y.F. (1997). Finite mixtures in confirmatory factor-analysis models. *Psychometrika,* 62, 297-330.

Table 1: Test results for the child's reading recognition example

Model		log-likelihood	# parameters	*BIC*
A1.	1-class LI	-1977	8	4003
A2.	2-class LI	-1694	17	3490
A3.	3-class LI	-1587	26	3330
B1.	1-class SA	-1595	14	3274
B2.	2-class SA	-1489	29	3151
B3.	3-class SA	-1459	44	3182
C1.	1-class RE	-1667	9	3375
C2.	2-class RE	-1561	19	3237
C3.	3-class RE	-1518	29	3211
D1.	1-class AR	-1611	9	3277
D2.	2-class AR	-1502	19	3118
D3.	3-class AR	-1477	29	3130
E1.	1-class FA	-1611	12	3294
E2.	2-class FA	-1497	25	3144
E3.	3-class FA	-1464	38	3157
F.	D2 + covariates	-1401	27	2964