Hierarchical Mixture Models for Nested Data Structures

Jeroen K. Vermunt[1] and Jay Magidson[2]

[1] Department of Methodology and Statistics, Tilburg University, PO Box 90153, 5000 LE Tilburg, Netherlands
[2] Statistical Innovations Inc., 375 Concord Avenue, Belmont, MA 02478, USA

Abstract. A hierarchical extension of the finite mixture model is presented that can be used for the analysis of nested data structures. The model permits a simultaneous model-based clustering of lower- and higher-level units. Lower-level observations within higher-level units are assumed to be mutually independent given cluster membership of the higher-level units. The proposed model can be seen as a finite mixture model in which the prior class membership probabilities are assumed to be random, which makes it very similar to the grade-of-membership (GoM) model. The new model is illustrated with an example from organizational psychology.

1 Introduction

Social science researchers, as researchers in other fields, are often confronted with nested or hierarchical data structures. Examples are data from employees belonging to the same organizations, individuals living in the same regions, customers of the same stores, repeated measures taken from the same individuals, and individuals belonging to the same primary sampling units in two-stage cluster samples.

This paper introduces an extension of the standard finite mixture model (McLachlan and Peel, 2000) that can take the hierarchical structure of a data set into account. Introducing random-effects in the model of interest is a common way to deal with dependent observations arising from nested data structures. It is well known that the finite mixture model is itself a nonparametric random-effects model (Aitkin, 1999). The solution that is proposed here is to introduce nonparametric random effects within a finite mixture model. That is, on top of a finite mixture model, we build another finite mixture model, which yields a model with a separate finite mixture distribution at each level of nesting.

When using the hierarchical mixture model for clustering, one obtains not only a clustering of lower-level units, but also a clustering of higher-level units. The clusters of higher-level units differ with respect to the prior probabilities corresponding to the lower-level clusters. This is similar to what is done in multiple-group latent class analysis, with the difference that we assume that each group belongs to one of a small number of clusters (latent classes) instead of estimating of a separate latent class distribution for each group. The latter approach would amount to using a fixed-effect instead of a random-effects model.

Because it is not practical to estimate the hierarchical mixture using a standard EM algorithm, we propose a variant of EM that we call the upward-downward algorithm. This method uses the conditional independence assumption of the underlying graphical model for an efficient implementation of the E step.

2 Model formulation

2.1 Standard finite mixture model

Let y_{ik} denote the response of individual i on indicator, attribute, or item k. The number of cases is denoted by N, and the number of items by K. The latent class variable is denoted by x_i, a particular latent class by t, and the number of latent classes by T. Notation $\mathbf{y}_i$ is used to refer to the full response vector for case i. A finite mixture model can be defined as (McLachlan and Peel, 2000)

$$f(\mathbf{y}_i) = \sum_{t=1}^{T} \pi(x_i = t)\, f(\mathbf{y}_i | x_i = t).$$

where $\pi(x_i = t)$ is the prior class membership probability corresponding to class t and $f(\mathbf{y}_i|x_i = t)$ is the class conditional density of $\mathbf{y}_i$. With continuous y_{ik}, we may take $f(\mathbf{y}_i|x_i = t)$ to be multivariate normal. If the indicators y_{ik} are categorical variables, we usually make the additional assumption that responses are independent given class membership (Lazarsfeld and Henry, 1968); that is,

$$f(\mathbf{y}_i|x_i = t) = \prod_{k=1}^{K} \pi(y_{ik}|x_i = t). \tag{1}$$

This assumption is justified if – as in our empirical example – the K items can be assumed to measure a single underlying dimension.

2.2 Hierarchical finite mixture model

For the hierarchical extension of the mixture model, we have to extend our notation to take into account the extra level of nesting. Let y_{ijk} denote the response of lower-level unit i within higher-level unit j on indicator k. The number of higher-level units is denoted by J, the number of lower-level units within higher-level unit j by n_j, and the number of items by K. Notation $\mathbf{y}_{ij}$ is used to refer to the full vector of responses of case i in group j, and $\mathbf{y}_j$ to refer to the full vector of responses for group j.

The latent class variable at the lower level is denoted by x_{ij}, a particular latent class by t, and the number of latent classes by T. The latent class variable at the higher level is denoted by u_j, a particular latent class by m, and the number of latent classes by M.

The hierarchical mixture model consist of two parts. The first part connects the observations belonging to the same group. It has the following form:

$$f(\mathbf{y}_j) = \sum_{m=1}^{M} \pi(u_j = m)\, f(\mathbf{y}_j | u_j = m)$$
$$f(\mathbf{y}_j | u_j = m) = \prod_{i=1}^{n_j} f(\mathbf{y}_{ij} | u_j = m).$$

As can be seen, groups are assumed to belong to one of M latent classes with prior probabilities equal to $\pi(u_j = m)$ and observations within a group are assumed be mutually independent given class membership of the group. Note that this conditional independence assumption is similar to the assumption of the latent class model for categorical variables (see equation 1).

The second part of the model is similar to the structure of a standard finite mixture model, except for the fact that now we are dealing with $f(\mathbf{y}_{ij}|u_j = m)$ instead of $f(\mathbf{y}_i)$; that is, we have to define a density conditional on the class membership of the higher-level unit. This yields

$$f(\mathbf{y}_{ij} | u_j = m) = \sum_{t=1}^{T} \pi(x_{ij} = t | u_j = m)\ f(\mathbf{y}_{ij} | x_{ij} = t). \tag{2}$$

In the case of categorical y_{ijk}, we will again assume that

$$f(\mathbf{y}_{ij} | x_{ij} = t) = \prod_{k=1}^{K} \pi(y_{ijk} | x_{ij} = t).$$

If we compare the standard mixture model with the hierarchical mixture model, we see two important differences: 1] we not only obtain information on class membership of individuals, but also on class membership of groups and 2] groups are assumed to differ with respect to the prior distribution of their members across lower-level latent classes.

It should be noted that the hierarchical mixture model is a graphical model with a tree structure. The upper node is the discrete latent variable at the higher level. The intermediate nodes consist of the n_j discrete latent variables for the lower-level units belonging to higher-level unit j. These x_{ij} are mutually independent given u_j. The lower nodes contain the observed responses y_{ijk}, which in the latent class model are assumed to be mutually independent given x_{ij}.

3 Maximum likelihood estimation by an adapted EM algorithm

If we put the various model parts together, we obtain the following log-likelihood function for the hierarchical mixture model:

$$\begin{aligned}\log L &= \sum_{j=1}^{J} \log f(\mathbf{y}_j) \\ &= \sum_{j=1}^{J} \log \sum_{m=1}^{M} \pi(u_j = m) \prod_{i=1}^{n_j} \left[\sum_{t=1}^{T} \pi(x_{ij} = t | u_j = m) \; f(\mathbf{y}_{ij} | x_{ij} = t) \right].\end{aligned}$$

A natural way to solve the ML estimation problem is by means of the EM algorithm (Dempster, Laird, and Rubin, 1977). The E step of the EM algorithm involves computing the expectation of the complete data log-likelihood, which in the hierarchical mixture model is of the form

$$\begin{aligned}E(\log L_c) = &\sum_{j=1}^{J} \sum_{m=1}^{M} P(u_j = m | \mathbf{y}_j) \log \pi(u_j = m) \\ &+ \sum_{j=1}^{J} \sum_{m=1}^{M} \sum_{i=1}^{n_j} \sum_{x=1}^{T} P(u_j = m, x_{ij} = t | \mathbf{y}_j) \log \pi(x_{ij} = m | u_j = m) \\ &+ \sum_{j=1}^{J} \sum_{m=1}^{M} \sum_{i=1}^{n_j} \sum_{t=1}^{T} P(x_{ij} = t | \mathbf{y}_j) \log f(\mathbf{y}_{ij} | x_{ij} = t).\end{aligned}$$

This shows that, in fact, the E step involves obtaining the posterior probabilities $P(u_j = m, x_{ij} = t | \mathbf{y}_j)$ given the current estimates for the unknown model parameters. In the M step of the EM algorithm, the unknown model parameters are updated so that the expected complete data log-likelihood is maximized (or improved). This can be accomplished using standard complete data algorithms for ML estimation.

The implementation of the E step is more difficult than the M step. A standard implementation would involve computing the joint conditional expectation of the $n_j + 1$ latent variables for higher-level unit j, that is, the joint posterior distribution $P(u_j, x_{1j}, x_{2j}, ..., x_{n_j j} | \mathbf{y}_j)$ with $M \cdot T^{n_i}$ entries. Note that this amounts to computing the expectation of all the "missing data" for a higher-level unit. These joint posteriors would subsequently be collapsed to obtain the marginal posterior probabilities for each lower-level unit i within higher-level unit j. A drawback of this procedure is that computer storage and time increases exponentially with the number of lower-level units, which means that it can only be used with small n_j.

Fortunately, it turns out that it is possible to compute the n_j marginal posterior probability distributions $P(u_j = m, x_{ij} = t | \mathbf{y}_j)$ without going through the full posterior distribution by making use of the conditional independence assumptions implied by the hierarchical mixture model. In that sense our procedure is similar to the forward-backward algorithm that can be used for the estimation of hidden Markov models with large numbers of time points (Baum et al., 1970). In the upward-downward algorithm, first, latent variables are integrated out going from the lower to the higher levels. Subsequently, the relevant marginal posterior probabilities are computed going from the higher to the lower levels. This yields a procedure in which computer storage and time increases linearly with the number of lower-level observations instead of exponentially, as would have been the case with a standard EM algorithm.

The upward-downward algorithm makes use of the fact that

$$\begin{aligned} P(u_j = m, x_{ij} = t|\mathbf{y}_j) &= P(u_j = m|\mathbf{y}_j)P(x_{ij} = t|\mathbf{y}_j, u_j = m) \\ &= P(u_j = m|\mathbf{y}_j)P(x_{ij} = t|\mathbf{y}_{ij}, u_j = m); \end{aligned}$$

that is, given class membership of the group (u_j), class membership of the individuals (x_{ij}) is independent of the information of the other group members. The terms $P(u_j = m|\mathbf{y}_{ij})$ and $P(x_{ij} = t|\mathbf{y}_{ij}, u_j = m)$ are obtained as follows:

$$\begin{aligned} P(x_{ij} = t|\mathbf{y}_{ij}, u_j = m) &= \frac{\pi(x_{ij} = t|u_j = m)f(\mathbf{y}_{ij}|x_{ij} = t)}{f(\mathbf{y}_{ij}|u_j = m)} \\ P(u_j = m|\mathbf{y}_j) &= \frac{\pi(u_j = m)\prod_{i=1}^{n_j} P(\mathbf{y}_{ij}|u_j = m)}{f(\mathbf{y}_j)}, \end{aligned}$$

where $f(\mathbf{y}_{ij}|u_j = m) = \sum_{t=1}^{T} \pi(x_{ij} = t|u_j = m)f(\mathbf{y}_{ij}|x_{ij} = t)$ and $f(\mathbf{y}_j) = \sum_{m=1}^{M} \pi(u_j = m)\prod_{i=1}^{n_j} P(\mathbf{y}_{ij}|u_j = m)$.

In the upward part, we compute $f(x_{ij} = t, \mathbf{y}_{ij}|u_j = m)$ for each individual, collapse these over x_{ij} to obtain $f(\mathbf{y}_{ij}|u_j = m)$, and use these to obtain $P(u_j = m|\mathbf{y}_j)$ for each group. The downward part involves computing $P(u_j = m, x_{ij} = t|\mathbf{y}_{ij})$ for each individual using $P(u_j = m|\mathbf{y}_i)$ and $P(x_{ij} = t|\mathbf{y}_{ij}, u_j = m)$.

A practical problem in the implementation of the above upward-downward method is that underflows may occur in the computation of $P(u_j = m|\mathbf{y}_j)$. Such underflows can, however, easily be prevented by working on a log scale. The algorithm described here will be implemented in version 4.0 of the Latent GOLD program for finite mixture modeling (Vermunt and Magidson, 2000).

4 An empirical example

We will illustrate the hierarchical mixture model using data taken from a Dutch study on the effect of team characteristics on individual work conditions (Van Mierlo, 2003). A questionnaire was completed by 886 employees from 88 teams of two organizations, a nursing home and a domiciliary care organization. Of interest for the illustration of the hierarchical mixture model is that employees are nested within (self-managing) teams, where the total number of observations per team ranged from 1 to 22.

Various aspects of work conditions were measured, one of which was the perceived task variety. The item wording of the five dichotomous items measuring perceived task variety is as follows (translated from Dutch):

1. Do you always do the same things in your work?
2. Does your work require creativity?
3. Is your work diverse?
4. Does your work make enough usage of your skills and capacities?
5. Is there enough variation in your work?

We had 36 cases with missing values on one or more of the indicators, but these cases can be retained in the analysis.

The model we use for these dichotomous response variables is an unrestricted latent class models. Besides a latent class model for the employees, we have to take into account the nested data structure. This is done by allowing teams to belong to clusters of teams that differ with respect to the prior distribution of the task-variety classes of employees. An alternative would have been to adopt a fixed-effects approach in which each team has its own prior latent class distribution. However, given the large number of higher-level units (88), this would yield a model with many parameters.

We fitted models with different numbers of classes of teams and different numbers of classes of employees within classes of teams. Table 1 reports the log-likelihood value, the number of parameters, and the BIC value for the estimated models. In the computation of BIC, we used the total number of employees (886) as the sample size. As can be seen, the very parsimonious model with two classes of teams and two classes of employees (within classes of teams) is the preferred model according to the BIC criterion.

Table 1. Testing results for the estimated models with the task-variety data

Teams	Employees	Log-likelihood	# Parameters	BIC value
1-class	1-class	-2797	5	5628
1-class	2-class	-2458	11	4991
1-class	3-class	-2444	17	5004
2-class	2-class	-2435	13	4958
2-class	3-class	-2419	20	4974
3-class	2-class	-2434	15	4970
3-class	3-class	-2417	23	4991

The estimated probability of giving a response that is in agreement with a high task variety (“no” for item 1 and “yes” for the other 4 indicators) equals

.51, .70, .97, .83, and .93 for the employees in the first latent class and .14, .17, .20, .42, and .17 for the second latent class. Thus, the first latent class can be called the high task-variety class and the second the low task-variety class.

Besides these two classes of employees we encountered two clusters of teams that differ in their team members' prior probability of belonging to the high task-variety class. In the first cluster of teams – containing 66% of the teams – this prior probability equals .79, whereas it is only .39 in the second cluster of teams. This shows that there are large differences between teams with respect to the perceived task variety of their employees. It also shows that the observations belonging to the same group are quite strongly correlated.

Whereas in this application the hierarchical structure arises from the nesting of individuals within groups, the proposed methodology is also useful in longitudinal studies: the higher-level units would then be individuals and the lower-level units measurement occasions or time points.

5 Variants and extensions

This paper presented the simplest form of the hierarchical mixture model. Several extensions and variants can be formulated. One important extension is the use of covariates affecting u_j, x_{ij}, or y_{ijk}. For example, assume that we have a set of P covariates affecting x_{ij} and that z_{ijp} denotes a particular covariate. In that case, we may use the following logit form for $\pi(x_{ij} = t|u_j = m, \mathbf{z}_{ij})$:

$$\pi(x_{ij} = t|u_j = m, \mathbf{z}_{ij}) = \frac{\exp(\gamma_{t0}^{m} + \sum_{p=1}^{P} \gamma_{tp}\, z_{ijp})}{\sum_{r=1}^{T} \exp(\gamma_{r0}^{m} + \sum_{p=1}^{P} \gamma_{rp}\, z_{ijp})}.$$

In equation (2), we implicitly assumed that u_j has no direct effect on $\mathbf{y}_{ij}$. In some applications one may wish to use an alternative structure for this equation. For example,

$$f(\mathbf{y}_{ij}|u_j = m) = \sum_{t=1}^{T} \pi(x_{ij} = t)\; f(\mathbf{y}_{ij}|x_{ij} = t, u_j = m),$$

which can be used for obtaining a three-level extension of the mixture regression model (see Vermunt, 2004). That is, a nonparametric random-coefficients model in which regression coefficients not only differ across clusters of lower-level units, but also across clusters of higher-level units.

The hierarchical mixture model is similar to the grade-of-membership (GoM) model (Manton, Woodbury and Tolley, 1994). As pointed out by Haberman (1995) and Esherova (2003), a GoM model can be defined as a latent class model with multiple exchangeable latent variables, which is exactly the same as is done in the hierarchical mixture model. Whereas we model the variation in prior class membership probabilities by nonparametric random effects, a hierarchical mixture model with parametric random effects would be even more similar to the GoM model. Vermunt (2003) proposed such a variant in which the logit of $\pi(x_{ij} = t|u_j)$ is assumed to be normally distributed, which is the common specification for the random effects in logistic regression models. More specifically,

$$\pi(x_{ij} = t|u_j) = \frac{\exp(\gamma_t + \tau_t \cdot u_j)}{\sum_{r=1}^{T} \exp(\gamma_r + \tau_r \cdot u_j)}$$

with $u_j \sim N(0, 1)$.

Whereas the hierarchical mixture model presented in this paper contains only two levels of nesting, it is straightforward to extend the model and the upward-downward algorithm to three or more levels.

References

AITKIN, M. (1999): A general maximum likelihood analysis of variance components in generalized linear models, *Biometrics, 55, 218-234*.

BAUM, L.E., PETRIE, T., SOULES, G. and WEISS, N. (1970): A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics, 41, 164–171*.

DEMPSTER, A.P., LAIRD. N.M., and RUBIN, D.B. (1977): Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society Series B, 39. 1–38*.

EROSHEVA, E.A. (2003): Partial membership models with application to disability survey data. H. Bozdogan (Ed.): *Statistical data mining and knowledge discovery.* Chapman and Hall/CRC, Boca Raton.

HABERMAN, S.J. (1995), Book review of "Statistical Applications Using Fuzzy Sets' by K.G. Manton, M.A. Woodbury, and H.D. Dennis., *Journal of the American Statistical Association, 90, 1131-1133*.

LAZARSFELD, P.F., and HENRY, N.W. (1968): *Latent Structure Analysis.* Houghton Mifflin, Boston.

MANTON, K.G., WOODBURY, M.A., and TOLLEY H.D. (1994): *Statistical Applications Using Fuzzy sets.* Wiley, New York.

MCLACHLAN, G.J., and PEEL, D. (2000): *Finite Mixture models.* Wiley, New York.

VAN MIERLO, H. (2003). *Self-managing Teams and Psychological Well-being.* Phd. dissertation. Eindhoven University of Technology, The Netherlands.

VERMUNT, J.K. (2003): Multilevel latent class models. *Sociological Methodology, 33, 213–239*.

VERMUNT, J.K. (2004): An EM algorithm for the estimation of parametric and nonparametric hierarchical nonlinear models. *Statistica Neerlandica, 58, 220-233*.

VERMUNT, J.K., and MAGIDSON, J. (2000): *Latent GOLD 2.0 User's Guide.* Statistical Innovations, Belmont, MA.