Finite Mixture Modeling with Mixture Outcomes Using the EM Algorithm

Bengt Muthén

Graduate School of Education and Information Studies and Department of Statistics,
University of California, Moore Hall, Box 951521, Los Angeles, California 90095-1521, U.S.A.
email: bmuthen@ucla.edu

and

Kerby Shedden

Department of Statistics, University of California, Los Angeles

SUMMARY. This paper discusses the analysis of an extended finite mixture model where the latent classes corresponding to the mixture components for one set of observed variables influence a second set of observed variables. The research is motivated by a repeated measurement study using a random coefficient model to assess the influence of latent growth trajectory class membership on the probability of a binary disease outcome. More generally, this model can be seen as a combination of latent class modeling and conventional mixture modeling. The EM algorithm is used for estimation. As an illustration, a random-coefficient growth model for the prediction of alcohol dependence from three latent classes of heavy alcohol use trajectories among young adults is analyzed.

KEY WORDS: Growth modeling; Latent class analysis; Latent variables; Maximum likelihood; Trajectory classes.

1. Introduction

This paper proposes an extended finite mixture model that combines features of Gaussian mixture models and latent class models. Analysis of this model is carried out using maximum-likelihood estimation with the EM algorithm and bootstrap standard errors. The research is motivated by an alcohol study concerned with the longitudinal development of heavy drinking and its relation to alcohol dependence.

Motivating Example

A conventional random-coefficient repeated measurement model was initially used to describe heavy drinking development by a quadratic growth model. For individual i of age a_t at time t, using centering of age, and letting η denote a random coefficient,

$$y_{it} = \eta_{1i} + (a_t - \bar{a})\eta_{2i} + (a_t - \bar{a})^2\eta_{3i} + \epsilon_{it}, \qquad t = 1, 2, \ldots, T, \tag{1}$$

where y_{it} is a measure of heavy drinking for individual i at age a_t, η_1 is the intercept, η_2 is the linear rate, η_3 is the quadratic growth rate, and ϵ is a residual. Using the estimated means of the η coefficients, this analysis showed that heavy drinking among young adults typically accelerates from 18 to 21 years of age and decelerates thereafter, which is in line with normative development found by alcohol researchers. Using η values different from the means, other classes of trajectory shapes recognized in the alcohol literature were also seen, such

as trajectories that are high at age 18 and either stay high or decrease with age and trajectories that increase from age 18 and show no downturn.

Of primary interest is how to model the relation between the shape of the heavy drinking trajectory for an individual in the 18–25-year age range and the probability of alcohol dependence at age 30. This could be done using logistic regression of dependence on the three η coefficients in (1). This is problematic, however, because a given η coefficient assumes different meanings depending on the value of the other η coefficients. A better approach, one that more clearly reflects that the curve shape has predictive value, is provided by an extended finite mixture model. This model allows the joint estimation of (i) a conventional finite mixture growth model where different curve shapes are captured by class-varying random coefficient means and (ii) a logistic regression of alcohol dependence on the classes.

In contrast to this approach, a three-step procedure is needed using conventional modeling techniques: (i) estimating the conventional finite mixture growth model, (ii) estimating each individual's most likely class membership *based on* the posterior probabilities for the classes derived from the estimated model in step (i), and (iii) regressing alcohol dependence on the estimated class membership. This three-step procedure introduces estimation errors in step (ii) by forcing each individual to be classified into a single class,

whereas each individual typically has a nonzero posterior probability for each class.

The extended finite mixture model allows three generalizations. First, the class membership probability is allowed to vary as a function of covariates. This is useful in alcohol studies, where it is known that white males are more likely to have a high level of heavy drinking at age 18. Second, for each class, the values of the η coefficients are allowed to be influenced by covariates. Third, the latent class variable is allowed to predict more than one binary outcome variable.

The case of multiple binary outcome variables predicted by the latent class variable warrants special attention. This model feature relates to latent class analysis (cf., Clogg, 1995). The alcohol study provides an interesting application of multiple binary outcomes. Drawing on the alcohol literature, the age 18–25-year growth curve classes can be viewed as representing different developmental pathways, some of which manifest themselves as deviant behavior already in adolescence. Using adolescent observations of deviant behavior, such as early onset of regular drinking and dropping out of high school, to form multiple binary outcomes, these outcomes can therefore be viewed as early indicators of the latent classes, where alcohol dependence is a later indicator. Together with the heavy drinking y's of equation (1) and the background variables, these binary outcomes contain information about the latent class membership. Given an estimated model, an individual's observations on these outcomes can be used to compute posterior probability estimates for different classes before the individual reaches age 25. This suggests for which individuals an intervention may be beneficial in order to avoid heavy drinking development typical of the nonnormative classes and to reduce the risk of developing alcohol dependence problems.

2. The Extended Finite Mixture Model

The following model incorporates the ideas presented in the Introduction. Consider a p-dimensional vector $\mathbf{y}$ of continuous variables and an r-dimensional vector $\mathbf{u}$ of binary variables, which are related to each other and to a q-dimensional vector $\mathbf{x}$ of covariates. The three sets of observed variables are related to each other via two vectors of unobserved variables, an m-dimensional vector $\boldsymbol{\eta}$ of latent continuous variables, and a K-dimensional vector $\mathbf{c}$ of latent categorical variables. Here, $\mathbf{c}_i = (c_{i1}, \ldots, c_{iK})'$ has a multinomial distribution, where $c_{ik} = 1$ if individual i falls in class k and is zero otherwise.

Consider the set of continuous observed variables $\mathbf{y}$ related to the continuous latent variables $\boldsymbol{\eta}$ for individual i,

$$\mathbf{y}_i = \Lambda_y \boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i, \tag{2}$$

where Λ_y is a $p \times m$ matrix of parameters and $\boldsymbol{\epsilon}_i$ is a residual vector that is uncorrelated with other variables in the model and is normally distributed with mean zero and a diagonal covariance matrix Θ. The continuous latent variables $\boldsymbol{\eta}$ are related to the categorical latent variables $\mathbf{c}$ and to the observed covariate vector $\mathbf{x}$ by the relations

$$\boldsymbol{\eta}_i = \mathbf{A}\mathbf{c}_i + \Gamma_\eta \mathbf{x}_i + \boldsymbol{\zeta}_i, \tag{3}$$

where the $m \times K$ matrix $\mathbf{A}$ contain columns $\boldsymbol{\alpha}_k$, $k = 1, 2, \ldots, K$, of intercept parameters for each c class, Γ_η is an

$m \times q$ parameter matrix, and $\boldsymbol{\zeta}$ is an m-dimensional residual vector that is normally distributed, uncorrelated with other variables, with mean zero and covariance matrix Ψ.

Consider next modeling of $\mathbf{u}$. In line with latent class analysis, the r binary variables u_{ij} are assumed to be conditionally independent given $\mathbf{c}_i$,

$$P(u_{i1}, u_{i2}, \ldots, u_{ir} \mid \mathbf{c}_i)$$
$$= P(u_{i1} \mid \mathbf{c}_i)P(u_{i2} \mid \mathbf{c}_i) \cdots P(u_{ir} \mid \mathbf{c}_i). \tag{4}$$

Defining $\tau_{ij} = P(u_{ij} = 1 \mid \mathbf{c}_i)$, the r-dimensional vector $\boldsymbol{\tau}_i = (\tau_{i1}, \tau_{i2}, \ldots, \tau_{ir})'$, and the r-dimensional vector $\text{logit}(\boldsymbol{\tau}_i) = (\log[\tau_{i1}/(1 - \tau_{i1})], \log[\tau_{i2}/(1 - \tau_{i2})], \ldots, \log[\tau_{ir}/(1 - \tau_{ir})])'$,

$$\text{logit}(\boldsymbol{\tau}_i) = \Lambda_u \mathbf{c}_i, \tag{5}$$

where Λ_u is an $r \times K$ parameter matrix. Define $\tau_{ijk} = P(u_{ijk} = 1 \mid c_{ik} = 1)$.

The categorical latent variables of $\mathbf{c}$ represent mixture components that are related to $\mathbf{x}$ through a multinomial logit regression model for unordered polytomous response. Defining $\pi_{ik} = P(c_{ik} = 1 \mid \mathbf{x}_i)$, the K-dimensional vector $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{iK})'$, and the $(K-1)$-dimensional vector $\text{logit}(\boldsymbol{\pi}_i) = (\log[\pi_{i1}/\pi_{iK}], \log[\pi_{i2}/\pi_{iK}], \ldots, \log[\pi_{i,K-1}/\pi_{iK}])'$,

$$\text{logit}(\boldsymbol{\pi}_i) = \boldsymbol{\alpha}_c + \Gamma_c \mathbf{x}_i, \tag{6}$$

where $\boldsymbol{\alpha}_c$ is a $(K-1)$-dimensional parameter vector and Γ_c is a $(K-1) \times q$ parameter matrix.

In the model of (2), (3), (4), (5), and (6), the finite mixture arises because the conditional distribution of $\mathbf{y}$ and $\mathbf{u}$ given $\mathbf{x}$ is governed by parameters that vary across the categories of $\mathbf{c}$; the mean vector of $\mathbf{y}$ is allowed to vary due to the inclusion of $\mathbf{c}$ in (3); and the probabilities of $\mathbf{u}$ are allowed to vary due to the inclusion of $\mathbf{c}$ in (5).

3. The EM Algorithm

Consider a sample of size n from the above model. The observed-data log likelihood may be expressed as follows. From the proposed model, we obtain $[\boldsymbol{\eta}_i \mid \mathbf{c}_i, \mathbf{x}_i] = N_m(\mathbf{A}\mathbf{c}_i + \Gamma_\eta \mathbf{x}_i, \Psi)$ and $[\mathbf{y}_i \mid \boldsymbol{\eta}_i] = N_p(\Lambda_y \boldsymbol{\eta}_i, \Theta)$ so that

$$[\mathbf{y}_i \mid \mathbf{c}_i, \mathbf{x}_i] = \int [\boldsymbol{\eta}_i \mid \mathbf{c}_i, \mathbf{x}_i][\mathbf{y}_i \mid \boldsymbol{\eta}_i]\partial\boldsymbol{\eta}_i$$
$$= N_p[\Lambda_y(\mathbf{A}\mathbf{c}_i + \Gamma_\eta \mathbf{x}_i), \Lambda_y \Psi \Lambda_y' + \Theta], \tag{7}$$

where $[\mathbf{z}]$ denotes a density or probability distribution for a random variable vector $\mathbf{z}$. The observed-data log likelihood is

$$\log L = \sum_{i=1}^{n} \log[\mathbf{y}_i, \mathbf{u}_i \mid \mathbf{x}_i], \tag{8}$$

where the mixture is defined as

$$[\mathbf{y}_i, \mathbf{u}_i \mid \mathbf{x}_i] = \sum_{k=1}^{K} \pi_{ik} N_p[\Lambda_y(\boldsymbol{\alpha}_k + \Gamma_\eta \mathbf{x}_i), \Lambda_y \Psi \Lambda_y' + \Theta]$$
$$\times [\mathbf{u}_i \mid c_{ik} = 1], \tag{9}$$

where π_{ik} is defined in connection with (6), $\boldsymbol{\alpha}_k$ is the kth column of the matrix $\mathbf{A}$ in (3), and $[\mathbf{u}_i \mid c_{ik} = 1]$ is defined by (4) and (5).

3.1 *The Complete-Data Likelihood*

Maximization of (9) can be simplified by using the EM algorithm (cf., McLachlan and Krishnan, 1997). Here, the continuous latent variable observations $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_n$, and the categorical latent variable observations $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n$ are viewed as missing data. Given the model, the complete-data log likelihood can then be written as

$$\log L_c = \sum_{i=1}^{n} (\log[\mathbf{c}_i \mid \mathbf{x}_i] + \log[\boldsymbol{\eta}_i \mid \mathbf{c}_i, \mathbf{x}_i] + \log[\mathbf{y}_i \mid \boldsymbol{\eta}_i]$$
$$+ \log[\mathbf{u}_i \mid \mathbf{c}_i]), \qquad (10)$$

where

$$\sum_{i=1}^{n} \log[\mathbf{c}_i \mid \mathbf{x}_i] = \sum_{i=1}^{n} \sum_{k=1}^{K} c_{ik} \log \pi_{ik} \qquad (11)$$

and

$$\sum_{i=1}^{n} \log[\mathbf{u}_i \mid \mathbf{c}_i] = \sum_{i=1}^{n} \sum_{j=1}^{r} \sum_{k=1}^{K} c_{ik}[u_{ij} \log \tau_{ijk}$$
$$+ (1 - u_{ij}) \log(1 - \tau_{ijk})], \qquad (12)$$

where τ_{ijk} is given in connection with (5). The remaining terms correspond to normal densities given by the model.

3.2 *The E-Step*

The EM algorithm maximizes the expected complete-data log likelihood (10) given the data on $\mathbf{y}, \mathbf{x}, \mathbf{u}$. In the E-step, we find the conditional expectations of c_{ik}, $\mathbf{S}_{\eta\eta}$, $\mathbf{S}_{cc}$, $\mathbf{S}_{\eta x}$, $\mathbf{S}_{\eta c}$, $\mathbf{S}_{cx}$, and $\mathbf{S}_{y\eta}$ using the notation $\mathbf{S}_{zw} = 1/n \sum_{i=1}^{n} \mathbf{z}_i \mathbf{w}_i'$. We note that

$$[\mathbf{c}_i, \boldsymbol{\eta}_i \mid \mathbf{y}_i, \mathbf{x}_i, \mathbf{u}_i] = [\mathbf{c}_i \mid \mathbf{y}_i, \mathbf{x}_i, \mathbf{u}_i][\boldsymbol{\eta}_i \mid \mathbf{c}_i, \mathbf{y}_i, \mathbf{x}_i, \mathbf{u}_i]. \qquad (13)$$

Corresponding to the first term, the posterior term for $\mathbf{c}_i$, we need the expectation $\mathrm{E}(c_{ik} \mid \mathbf{y}_i, \mathbf{x}_i, \mathbf{u}_i)$,

$$P(c_{ik} = 1 \mid \mathbf{y}_i, \mathbf{x}_i, \mathbf{u}_i)$$
$$= \pi_{ik} N_p[\boldsymbol{\Lambda}_y(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_\eta \mathbf{x}_i), \boldsymbol{\Lambda}_y \boldsymbol{\Psi} \boldsymbol{\Lambda}_y' + \boldsymbol{\Theta}]$$
$$\times [\mathbf{u}_i \mid c_{ik} = 1]/[\mathbf{y}_i, \mathbf{u}_i \mid \mathbf{x}_i], \qquad (14)$$

where $[\mathbf{y}_i, \mathbf{u}_i \mid \mathbf{x}_i]$ is given in (9). Let p_{ik} denote the posterior probability of (14). The resulting E-step quantities are given in the Appendix.

3.3 *The M-Step*

The M-step of the EM algorithm is defined as follows. Inserting the posterior probabilities p_{ik} of (14) in (11), the M-step maximizes

$$\sum_{i=1}^{n} \sum_{k=1}^{K} p_{ik} \log \pi_{ik} \qquad (15)$$

with respect to the parameters of $\boldsymbol{\alpha}_c$, $\boldsymbol{\Gamma}_c$. This may be seen as a multinomial logistic regression with fractional observations p_{ik}. Similarly, inserting p_{ik} in (12), the M-step maximizes

$$\sum_{i=1}^{n} \sum_{j=1}^{r} \sum_{k=1}^{K} p_{ik}[u_{ij} \log \tau_{ijk} + (1 - u_{ij}) \log(1 - \tau_{ijk})] \qquad (16)$$

with respect to the parameters of $\boldsymbol{\Lambda}_u$. This may be seen as a type of multivariate-response logistic regression with observations u_{ij} and weights p_{ik}. We have found that the EM algorithm converges nicely even if only one or two Newton–Raphson steps are taken in the two logistic regressions (see also generalized EM as discussed in McLachlan and Krishnan, 1997). The M-step estimates for the parameter arrays $\boldsymbol{\Psi}$, $\boldsymbol{\Gamma}_\eta$, $\boldsymbol{\Lambda}_y$, and $\boldsymbol{\Theta}$ are obtained in closed form and are given in the Appendix.

3.4 *Comments*

The model is not identified without some parameter restrictions, which have to be determined for a given application. As a well-known example drawing on factor analysis (cf., Lawley and Maxwell, 1971), at least m^2 restrictions need to be imposed on the elements of $\boldsymbol{\Lambda}_y$ and/or $\boldsymbol{\Psi}$. It is difficult to give rules for identification of the general model (cf., Titterington, Smith, and Makov, 1985). A heuristic approach to understanding the identification status of a particular case of the model is to divide the model into its parts. As an example, for the growth model in (1), all elements of $\boldsymbol{\Lambda}_y$ are fixed so that more than m^2 restrictions are imposed. Here, the $\mathbf{y}$, $\boldsymbol{\eta}$, $\mathbf{c}$ part of the model is the mixed-effect mixture model of Verbeke and Lesaffre (1996). The covariance matrices $\boldsymbol{\Psi}$ and $\boldsymbol{\Theta}$ are class invariant, while the random coefficient $\boldsymbol{\eta}$ means of $\mathbf{A}$ vary across classes with class probabilities determined by $\boldsymbol{\alpha}_c$. The identification therefore concerns a multivariate normal mixture for $\mathbf{y}$ with class-invariant covariance matrix and with means that are functions of the class-varying means $\mathbf{A}$ of the reduced dimension m of the underlying $\boldsymbol{\eta}$. Growth curve data with clearly separated growth forms are likely to be able to identify such mean mixtures. Adding $\mathbf{x}$ to the model introduces the parameter arrays $\boldsymbol{\Gamma}_c$ and $\boldsymbol{\Gamma}_\eta$, for which the joint distribution of $\mathbf{y}$ and $\mathbf{x}$ carries information. The elements of $\boldsymbol{\Gamma}_c$ affect the probabilities of $\mathbf{c}$, where $\mathbf{c}$ class membership alters the means $\mathbf{A}$ of $\boldsymbol{\eta}$ and ultimately the means of $\mathbf{y}$. For a certain $\mathbf{x}$ value, a certain mixture of the class-specific means $\mathbf{A}$ of $\boldsymbol{\eta}$ is obtained, while at a different $\mathbf{x}$ value, a different mixture of the means is obtained. The model is identifiable as long as the resulting $\mathbf{y}$ means are different and there are more distinct $\mathbf{x}$ combinations than corresponding parameters. The elements of $\boldsymbol{\Gamma}_\eta$ concern effects of $\mathbf{x}$ on $\boldsymbol{\eta}$ given $\mathbf{c}$. $\boldsymbol{\Gamma}_\eta$ and $\boldsymbol{\Gamma}_c$ affect the model differently in that the elements of $\boldsymbol{\Gamma}_\eta$ affect the $\boldsymbol{\eta}$ means an equal amount for each class, while the elements of $\boldsymbol{\Gamma}_c$ affect the $\boldsymbol{\eta}$ means through a mixture. When the $\mathbf{y}$, $\mathbf{c}$, $\mathbf{x}$ part of the model is identified, the identification of the $\mathbf{u}$, $\mathbf{c}$, $\mathbf{x}$ part of the model concerns only the parameters of $\boldsymbol{\Lambda}_u$. Information on these parameters is obtained from the marginal distribution of $\mathbf{u}$ and from the joint distribution of $\mathbf{u}$ and $\mathbf{x}$. The marginal distribution of $\mathbf{u}$ may not be sufficient to identify $\boldsymbol{\Lambda}_u$ unless there are many u variables relative to the number of classes (cf., Clogg, 1995). Each row of $\boldsymbol{\Lambda}_u$ is, however, identifiable from the joint distribution of each u and $\mathbf{x}$ as in finite mixture logistic regression with class-varying intercepts (see Follman and Lambert, 1989, 1991).

In some applications, it is of interest to generalize the model in (2), (3), (4), (5), and (6) by allowing the inclusion of direct effects from $\mathbf{x}$ to $\mathbf{y}$ and from $\mathbf{x}$ to $\mathbf{u}$. The former corresponds to using time-varying covariates in a growth model setting, while the latter will be shown useful in our application. To include direct effects from $\mathbf{x}$ to $\mathbf{u}$, we extend the logit expression in

(5) to

$$\text{logit}(\tau_i) = \boldsymbol{\Lambda}_u \mathbf{c}_i + \mathbf{K}_u \mathbf{x}_i, \qquad (17)$$

where, more generally, $\mathbf{K}_u$ can also be allowed to vary across k, with $k = 1, 2, \ldots, K$.

We may generalize the model to allow for class-specific effects $\boldsymbol{\Gamma}_\eta$ in (3), e.g., by extending (3) to

$$\boldsymbol{\eta}_i = \mathbf{A}\mathbf{c}_i + \boldsymbol{\Gamma}_\eta^* \mathbf{c}_i^* \mathbf{x}_i + \boldsymbol{\zeta}_i, \qquad (18)$$

where $\boldsymbol{\Gamma}_\eta^* = (\boldsymbol{\Gamma}_{\eta_1} \mid \cdots \mid \boldsymbol{\Gamma}_{\eta_K})$ and $\mathbf{c}_i^* = \mathbf{c}_i \otimes \mathbf{cI}_q$. Class-specific covariance matrices $\boldsymbol{\Psi}$ and $\boldsymbol{\Theta}$ may also be of interest.

Standard errors of the estimates are obtained using a bootstrap procedure, where sampling an observation with replacement n times from the sample of size n is replicated 200 times. Reasonably stable standard error estimates are obtained already after 50 replications. Here, the entire vector $(\mathbf{u}_i', \mathbf{y}_i', \mathbf{x}_i')$ is bootstrapped rather than holding the x part fixed. While $\mathbf{x}_i$ is a covariate vector, in the applications considered here, it is a random vector, for which no model structure is imposed. A comparison of model fit across models with different numbers of classes is achieved by using BIC (Bayesian Information Criterion; Schwarz, 1978).

4. A Growth Curve Application

This section returns to the alcohol research question discussed in the Introduction, i.e., what the influence of membership in different growth curve classes for heavy drinking from ages 18 to 25 is on alcohol dependence at age 30. Alcohol data from the National Longitudinal Survey of Youth (NLSY), a nationally representative household survey of young adults living in the U.S. in 1979, are used. The heavy drinking variable is obtained from the question "How often have you had 6 or more drinks on one occasion during the last 30 days?" The variable is scored 0 (never), 1 (once), 2 (2 or 3 times),

3 (4 or 5 times), 4 (6 or 7 times), 5 (8 or 9 times), and 6 (10 or more times). The vector $\mathbf{y}$ in equation (2) consists of heavy drinking measured at 18, 19, 20, 24, and 25 years for the NLSY cohort born in 1964. The growth model of (1) is expressed in finite mixture form by equations (2) and (3), arranging the vector $\mathbf{y}_i$ as $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{i5})'$ and defining $\boldsymbol{\eta}$ as the 3×1 vector of intercept, linear, and quadratic coefficients and the 5×3 matrix $\boldsymbol{\Lambda}_y$ as having columns of constants 1, $(a_t - \bar{a})$, and $(a_t - \bar{a})^2$. The time scale is centered so that the intercept refers to the approximate peak for the normative trajectory, $\bar{a} = 21.2$. As a starting point, a two-class model $(K = 2)$ is postulated. In equation (3) for $\boldsymbol{\eta}$, the vector $\mathbf{x}$ consists of four covariates scored zero or one: male, Black, Hispanic, and FH123. Here, FH123 is scored one if the respondent has a first-degree relative and a second- or third-degree relative with alcohol problems. In equation (5) for $\mathbf{u}$, three binary u variables are included. The first u, DEP, captures a diagnosis of alcohol dependence at age 30. This diagnosis is based on alcohol-dependent behaviors such as giving up important social or work-related functions in favor of or as a consequence of drinking. The second u, ES, is scored one if the respondent started drinking regularly at or before age 14. The third u, HS, is scored one if the respondent did not complete high school by age 22. In equation (6) for $\mathbf{c}$, the same set of x variables is used as in (3) for $\boldsymbol{\eta}$. The sample size is 935. The y variables are not normally distributed in this application. To investigate the sensitivity of the modeling to this assumption, an analysis of variables transformed as $\log(1 + y)$ was also carried out and gave very similar results.

The two-class mixture model results in two local solutions when using different starting points for the EM algorithm. The estimates for the majority class showing the normative curve is very similar in the two solutions. However, the two solutions find different minority classes, one with a curve that
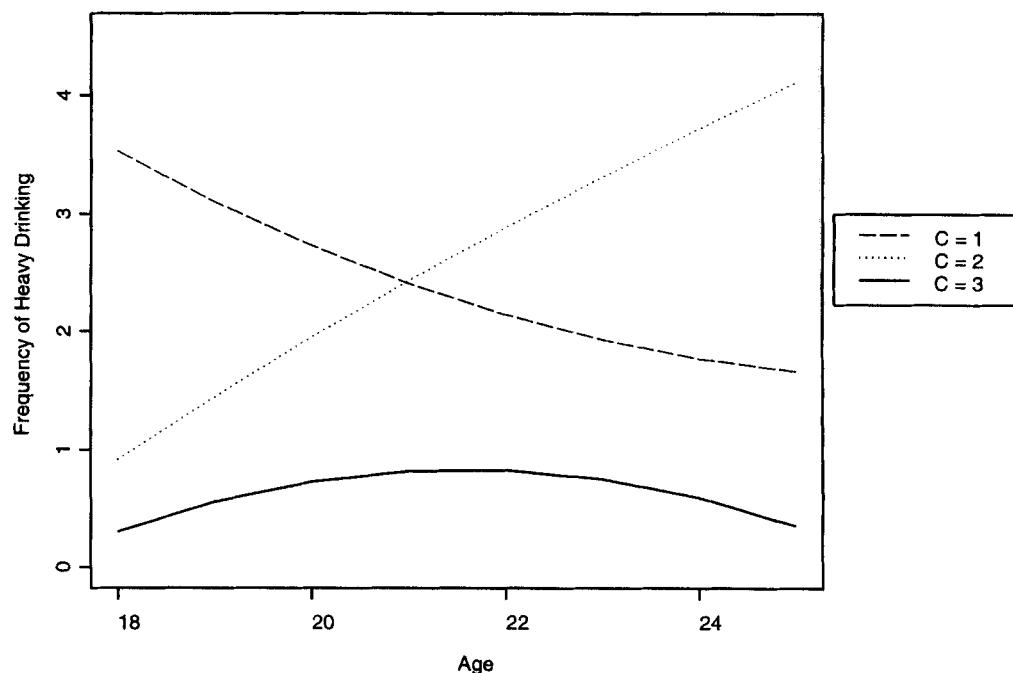

Figure 1.   Estimated three-class curves.

Table 1
Marginal fit of the three-class model

ES	HS	DEP	Observed	Fitted	Frequency
0	0	0	0.652	0.637	610
0	0	1	0.070	0.081	65
0	1	0	0.113	0.126	106
0	1	1	0.027	0.019	25
1	0	0	0.088	0.095	82
1	0	1	0.020	0.017	19
1	1	0	0.025	0.021	23
1	1	1	0.005	0.005	5

Table 2
Estimates for the growth factors of
η *(standard errors in parentheses)*

	Growth factor		
Parameter	Intercept	Linear rate	Quadratic rate
A			
High	2.263 (0.456)	−0.268 (0.053)	0.027 (0.043)
Up	2.655 (0.556)	0.456 (0.062)	−0.012 (0.045)
Norm	0.829 (0.102)	0.007 (0.012)	−0.041 (0.008)
Γ_η			
Male	0.779 (0.133)	0.048 (0.015)	−0.043 (0.010)
Black	−0.762 (0.124)	0.012 (0.014)	0.043 (0.010)
Hisp	−0.675 (0.137)	−0.003 (0.018)	0.041 (0.011)
FH123	0.076 (0.169)	0.019 (0.024)	0.002 (0.014)
Ψ			
Intercept	1.191 (0.206)		
Linear	0.025 (0.020)	0.062 (0.003)	
Quadratic	−0.076 (0.014)	−0.001 (0.002)	0.005 (0.001)

Table 3
Estimates for the latent class membership
of c (standard errors in parentheses)

Parameter	High versus norm	Up versus norm
α_c	−2.510 (0.303)	−3.282 (0.444)
Γ_c		
Male	1.293 (0.250)	1.419 (0.393)
Black	−1.416 (1.054)	−0.539 (0.594)
Hisp	0.015 (0.307)	−0.476 (0.538)
FH123	0.760 (0.353)	0.651 (1.202)

is high at age 18 with a subsequent decrease and one with a curve that accelerates from age 18. A three-class solution captures the three classes represented in both two-class solutions and is therefore preferred. The three estimated curves for the three-class solution are shown in Figure 1 using the estimated mean of η for the subgroup of $x = (0\,0\,0\,0)$, white females with no family history of alcohol problems. The log likelihood value for the three-class solution is −4934.91. In comparison, the log likelihood value for the two-class solution that includes a curve that goes down is −5034.27, while the log likelihood value for the two-class solution that includes a curve that goes up is −5121.72. The corresponding three BIC values are 10,218.69, 10,376.36, and 10,551.26, suggesting a choice of three classes over two classes.

A relatively strong assumption in the model is that the set of u variables depends on the x's only through c as indicated in (5) and (6). A plausible alternative is that there are direct effects from some x's to some u's as in (17), i.e., some u's differ in their probabilities not only as a function of class membership but also as a function of their covariate characteristics for given class membership. In this application, there are 12 such direct effects that could be included. Explorations of these effects show that only five of them are significant using the improvement in the log likelihood as criterion. The direct effects are for male to DEP, for FH123 to DEP, for Black to ES and to HS, and for Hisp to HS and are all of expected signs. For the final three-class model, the log likelihood value is −4909.81, where the improvement in fit relative

to the three-class model without direct effects is significant, corresponding to a likelihood ratio chi-square value of 50.2 with 5 d.f.

A simple empirical approach to checking the local identifiability of the estimated model is to start from a model that is known to be identified and check whether adding a parameter changes the observed-data log likelihood. For the five direct effects, this check was just accomplished by the chi-square difference test. Setting the 12 parameters of Γ_η to zero gives a chi-square difference value of 127.16; setting the eight parameters of Γ_c to zero gives a chi-square difference value of 71.54; and setting the nine parameters of Λ_u to zero gives a chi-square difference test of 1090.5.

For the final three-class model, Table 1 shows that the marginal table for ES, HS, and DEP is well fitted. The y means also fit well. The three curves are almost exactly the same as in Figure 1. The notation High, Up, and Norm will be used for the classes corresponding to these three curve shapes, where High refers to those who are high at age 18, Up refers to those who accelerate their use, and Norm refers to the normative curve. The three classes have estimated proportions 0.120, 0.063, and 0.817.

Table 2 shows the estimates for the growth factors of η corresponding to (3). Table 3 shows the estimates for the latent class membership corresponding to (6). Table 4 shows the estimates for y and u corresponding to (2) and (5) with direct effects $\mathbf{K}_u$ corresponding to (17). In Table 5, these estimates have been translated into conditional probability estimates for the three u variables DEP, ES, and HS given class membership, evaluated at $x = (0\,0\,0\,0)$. The DEP probabilities vary greatly as a function of latent class membership, and ES and HS are indicative of membership in the nonnormative classes. The normative class constitutes about 89% of all individuals with $x = (0\,0\,0\,0)$, while the High and Up classes constitute 7 and 4%, respectively. In contrast, for $x = (1\,0\,0\,1)$, white males with family history, there are only 51% in the normative class and 33 and 16% in the High and Up classes, respectively. For this group, the DEP probabilities still vary greatly across the latent classes: 0.378, 0.560, and 0.194, for the High, Up, and Norm classes, respectively. This analysis illustrates the explanatory power of the latent class variable.

5. Conclusion

The extended finite mixture model proposed here offers a flexible analysis framework. On the one hand, the model may be seen as a generalization of latent class analysis, so that the emphasis of the analysis is not only on the indicators of the

Table 4
Estimates for heavy drinking y and binary outcomes u (standard errors in parentheses)

Parameter					
Λ_y diag(Θ)	Fixed 0.598 (0.130)	1.006 (0.109)	0.980 (0.109)	1.147 (0.138)	0.525 (0.130)
			High	Up	Norm
Λ_u	DEP		−1.903 (0.338)	−1.159 (0.427)	−2.820 (0.202)
	ES		−0.436 (0.266)	−1.243 (0.457)	−2.015 (0.137)
	HS		−1.584 (0.274)	−1.440 (0.383)	−2.205 (0.182)
		Male	Black	Hisp	FH123
$\mathbf{K}_u$	DEP	0.719 (0.251)			0.689 (0.296)
	ES		−0.738 (0.274)		
	HS		0.658 (0.240)	1.153 (0.237)	

Table 5
Estimated probabilities for u^a

	High	Up	Norm
DEP	0.131	0.240	0.056
ES	0.389	0.227	0.117
HS	0.169	0.188	0.099
Class probabilities	0.073	0.035	0.892

[a] Evaluated at $x = (0000)$.

latent classes but also on incorporating other model parts and outcomes. On the other hand, the model may be seen as a generalization of conventional Gaussian mixture modeling where mixture indicators have been added. The EM algorithm was found to be a practical tool for estimation of the type of models considered. As an illustration of the broader analysis potential, the extended finite mixture model was found useful for random-coefficient growth modeling in the presence of several growth classes. Here, a mixture outcome was measured at a later time point and the latent growth classes were used as predictors of this outcome. Repeated measurement random coefficient modeling with mixtures avoids the normality assumption typically used for the random coefficients (see also Verbeke and Lesaffre, 1996). A drawback is that multiple solutions are often found so that multiple starting points are necessary.

ACKNOWLEDGEMENTS

RÉSUMÉ

Dans cet article, nous discutons d'un modèle de mélange fini dans lequel les classes latentes correspondant aux composantes du mélange pour un ensemble des variables observées influencent un second ensemble de variables observées. La motivation de cette recherche provient d'une étude de mesures répétées utilisant un modèle à coefficient aléatoire pour déterminer l'influence de l'appartenance à une classe de trajectoire de croissance latente sur la probabilité de l'issue d'une maladie binaire. Plus généralement, on peut voir ce modèle comme une combinaison de modélisation de classe latente et de modélisation d'un mélange conventionnel. Pour l'estimation, nous utilisons l'algorithme EM. Comme illustration, nous analysons un modèle de croissance à coefficient aléatoire pour prédire la dépendance alcoolique à partir de trois classes latentes de forte consommation alcoolique chez de jeunes adultes.

REFERENCES

Clogg, C. C. (1995). Latent class models. In *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, G. Arminger, C. C. Clogg, and M. E. Sobel (eds), 311–359. New York: Plenum.

Follman, D. A. and Lambert, D. (1989). Generalizing logistic regression by nonparametric mixing. *Journal of the American Statistical Association* **84**, 294–300.

Follman, D. A. and Lambert, D. (1991). Identifiability of finite mixtures of logistic regression models. *Journal of Statistical Planning and Inference* **27**, 375–381.

Lawley, D. N. and Maxwell, A. E. (1971). *Factor Analysis as a Statistical Method*. London: Butterworths.

McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. New York: John Wiley and Sons.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Chichester, U.K.: John Wiley and Sons.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association* **91**, 217–221.

Appendix

The E-Step

Let $\mathbf{p}_i = (p_{i1}, \ldots, p_{iK})'$ and let $\hat{\mathbf{S}}$ denote the conditional expectation of $\mathbf{S}$. This defines the E-step quantities for the missing data involving $\mathbf{c}$,

$$\hat{\mathbf{S}}_{cc} = 1/n \sum_{i=1}^{n} \text{diag}(\mathbf{p}_i) \tag{19}$$

and

$$\hat{\mathbf{S}}_{cx} = 1/n \sum_{i=1}^{n} \mathbf{p}_i \mathbf{x}_i'. \tag{20}$$

The average posterior probability for class k is

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} p_{ik}.$$

Furthermore, for the missing data involving $\boldsymbol{\eta}$,

$$\hat{\mathbf{S}}_{\eta\eta} = \mathbf{V} + \mathbf{B} + \mathbf{C} + \mathbf{D}, \tag{21}$$

where

$$\mathbf{V} = (\boldsymbol{\Lambda}_y' \boldsymbol{\Theta}^{-1} \boldsymbol{\Lambda}_y + \boldsymbol{\Psi}^{-1})^{-1}, \tag{22}$$

$$\mathbf{B} = \mathbf{V}\boldsymbol{\Psi}^{-1}(\boldsymbol{\Gamma}_\eta \mathbf{S}_{xx} \boldsymbol{\Gamma}_\eta' + \boldsymbol{\Gamma}_\eta \hat{\mathbf{S}}_{cx}' \mathbf{A}' + \mathbf{A}\hat{\mathbf{S}}_{cx} \boldsymbol{\Gamma}_\eta' + \mathbf{A}\hat{\mathbf{S}}_{cc} \mathbf{A}')$$
$$\times \boldsymbol{\Psi}^{-1} \mathbf{V}, \tag{23}$$

$$\mathbf{C} = \mathbf{V}\boldsymbol{\Lambda}_y' \boldsymbol{\Theta}^{-1} \mathbf{S}_{yy} \boldsymbol{\Theta}^{-1} \boldsymbol{\Lambda}_y \mathbf{V}, \tag{24}$$

$$\mathbf{D} = \mathbf{V}\boldsymbol{\Psi}^{-1} \boldsymbol{\Gamma}_\eta \mathbf{S}_{xy} \boldsymbol{\Theta}^{-1} \boldsymbol{\Lambda}_y + \mathbf{V}\boldsymbol{\Psi}^{-1} \mathbf{A}\mathbf{S}_{cy} \boldsymbol{\Theta}^{-1} \boldsymbol{\Lambda}_y, \tag{25}$$

and

$$\hat{\mathbf{S}}_{y\eta} = \hat{\mathbf{S}}_{yy} \boldsymbol{\Theta}^{-1} \boldsymbol{\Lambda}_y + (\mathbf{S}_{yx} \boldsymbol{\Gamma}_\eta' + \mathbf{S}_{yp} \mathbf{A}') \boldsymbol{\Psi}^{-1} \mathbf{V}, \tag{26}$$

$$\hat{\mathbf{S}}_{\eta x} = \mathbf{V}(\boldsymbol{\Psi}^{-1}(\boldsymbol{\Gamma}_\eta \mathbf{S}_{xx} + \mathbf{A}\mathbf{S}_{px}) + \boldsymbol{\Lambda}_y' \boldsymbol{\Theta}^{-1} \mathbf{S}_{yx}), \tag{27}$$

$$\hat{\mathbf{S}}_{\eta c} = \mathbf{V}(\boldsymbol{\Psi}^{-1}(\boldsymbol{\Gamma}_\eta \mathbf{S}_{xp} + \mathbf{A}\hat{\mathbf{S}}_{cc}) + \boldsymbol{\Lambda}_y' \boldsymbol{\Theta}^{-1} \mathbf{S}_{yp}). \tag{28}$$

The M-Step

The M-step for $\boldsymbol{\Psi}$ and $\boldsymbol{\Gamma}_\eta$ is obtained using the $\hat{\mathbf{S}}$ matrices from the E-step. Maximizing with respect to the regression coefficients $\boldsymbol{\Gamma}_\eta$ and $\mathbf{A}$ gives

$$(\hat{\boldsymbol{\Gamma}}_\eta \hat{\mathbf{A}}) = (\hat{\mathbf{S}}_{\eta x} \hat{\mathbf{S}}_{\eta c}) \hat{\mathbf{S}}_{x,c}^{-1}, \tag{29}$$

where $\hat{\mathbf{S}}_{x,c}$ is the joint covariance matrix for $(\mathbf{x}, \mathbf{c})$. Maximizing with respect to the covariance matrix $\boldsymbol{\Psi}$ gives

$$\hat{\boldsymbol{\Psi}} = \hat{\mathbf{S}}_{\eta\eta} + \hat{\boldsymbol{\Gamma}}_\eta \hat{\mathbf{S}}_{xx} \hat{\boldsymbol{\Gamma}}_\eta' + \hat{\mathbf{A}}\hat{\mathbf{S}}_{cc}\hat{\mathbf{A}}' - \hat{\mathbf{S}}_{\eta x} \hat{\boldsymbol{\Gamma}}_\eta' - \hat{\boldsymbol{\Gamma}}_\eta \hat{\mathbf{S}}_{x\eta} - \hat{\mathbf{S}}_{\eta c} \hat{\mathbf{A}}'$$
$$- \hat{\mathbf{A}}\hat{\mathbf{S}}_{c\eta} + \hat{\boldsymbol{\Gamma}}_\eta \hat{\mathbf{S}}_{xc} \hat{\mathbf{A}}' + \hat{\mathbf{A}}\hat{\mathbf{S}}_{cx} \hat{\boldsymbol{\Gamma}}_\eta'. \tag{30}$$

The M-step for $\boldsymbol{\Lambda}_y$ and $\boldsymbol{\Theta}$ is obtained as

$$\hat{\boldsymbol{\Lambda}}_y = \hat{\mathbf{S}}_{y\eta} \hat{\mathbf{S}}_{\eta\eta}^{-1}, \tag{31}$$

while, noting that $\boldsymbol{\Theta}$ is assumed to be diagonal,

$$\hat{\boldsymbol{\Theta}} = \text{diag}(\hat{\mathbf{S}}_{yy} - \hat{\mathbf{S}}_{y\eta} \hat{\boldsymbol{\Lambda}}_y' - \hat{\boldsymbol{\Lambda}}_y \hat{\mathbf{S}}_{\eta y} + \hat{\boldsymbol{\Lambda}}_y \hat{\mathbf{S}}_{\eta\eta} \hat{\boldsymbol{\Lambda}}_y'). \tag{32}$$