PSYCHOMETRIKA—VOL. 63, NO. 3, 227–253
SEPTEMBER 1998

FITTING MULTIVARIAGE NORMAL FINITE MIXTURES SUBJECT TO STRUCTURAL EQUATION MODELING

CONOR V. DOLAN

VRIJE UNIVERSITEIT

HAN L. J. VAN DER MAAS

UNIVERSITY OF AMSTERDAM

This paper is about fitting multivariate normal mixture distributions subject to structural equation modeling. The general model comprises common factor and structural regression models. The introduction of covariance and mean structure models reduces the number of parameters to be estimated in fitting the mixture and enables one to investigate a variety of substantive hypotheses concerning the differences between the components in the mixture. Within the general model, individual parameters can be subjected to equality, nonlinear and simple bounds constraints. Confidence intervals are based on the inverse of the Hessian and on the likelihood profile. Several illustrations are given and results of a simulation study concerning the confidence intervals are reported.

Key words: structural equation modeling, multivariate normal mixtures, quasi-Newton, LISREL.

Introduction

A finite mixture distribution is one that consists of a number of distinct components (Everitt & Hand, 1981; Titterington, Smith, & Makov, 1985). Each case, drawn from such a distribution, belongs to one of the component distributions, but to which one it belongs is unknown. In addition to the unknown component membership of the cases, the number and the type of the component distributions may be unknown. In so-called *direct* applications of finite mixture modeling (Titterington, et al., 1985, chap. 2), the aim is to determine the number and the type of components in the mixture, to estimate the unknown parameters, and to assign the cases to their respective components. In *indirect* applications the finite mixture model is employed as a mathematical device, for example, to approximate intractable heavy-tailed distributions. In such applications, the underlying components do not necessarily have a physical interpretation.

Here we are concerned with fitting multivariate normal finite mixtures in direct ap-

Requests for reprints should be sent to Han L. J. van der Maas, Developmental Psychology, Psychology Faculty, University of Amsterdam, Roetersstraat 15, 1018WB Amsterdam, THE NETHERLANDS.

We thank Harrie Vorst for his generous support. We thank Yiu-Fai Yung for kindly making available his thesis (Yung, 1994), and a copy of his, at the time of revising the present paper, unpublished Psychometrika article (Yung, 1997). We thank Petra Stein for making available her paper with Gerhard Arminger (Arminger & Stein, 1997). A total of five reviews provided by anonymous referees, and additional comments by the editors resulted in a good number of improvements. Finally, we thank Peter Molenaar for his critical comments.

This paper was first submitted to Psychometrika in May of 1996, after a period of about 6 months in which we developed and tested our FORTRAN routines. In two rounds of highly constructive reviews, two dissertations (Stein, 1997; Yung, 1994), and several articles were brought to our attention, that were either submitted (Arminger & Stein), or in press (Yung, Jedidi et al.). In the mean time, a number of these papers have appeared (Jedidi et al. 1997a, 1997b), and more are sure to follow (Yung, 1997; Arminger & Stein, 1997). It is clear that the subject of structural equation modeling within normal mixtures has taken off over the past few years (though Blåfield's pioneering thesis appeared in 1980). Although we were unaware of the work that was on-going, or indeed completed, when we embarked on this project, it is with pleasure that we acknowledge the precedence of those, whose work, in press, submitted, or otherwise, has come to our attention since first submitting this paper to Psychometrika.

0033-3123/98/0900-96483$00.75/0

plications subject to structural equation modeling (SEM) of the mean vectors and covariance matrices within each component. The model we employ is a LISREL submodel (Jöreskog & Sörbom, 1993) that includes confirmatory factor and structural equation models. Restrictions on the parameters in normal mixtures fall into two categories. In the first category, restrictions are imposed to avoid local maxima and singularities in the likelihood surface in maximum likelihood estimation (e.g., Everitt & Hand, 1981, p. 39; Titterington, et al., 1985, sec. 4.3.3). These restrictions, which include equality and proportionality constraints on variances (Hathaway, 1985), are based primarily on computational, not substantive, considerations. In the other category, the restrictions are aimed at testing substantive hypotheses concerning the relationship among the variables within the components. These categories are not mutually exclusive: substantively motivated restrictions may be computationally beneficial.

The restrictions in the second category are inspired by substantive hypotheses concerning the relationship between the variables in the mixture distribution. Because of much recent interest, there is at present a considerable body of work concerned with statistical modeling within mixtures. A large part of this work is based on so-called conditional mixture models (Wedel & DeSarbo, 1994, 1995). In conditional mixture models, the mixture is fitted to the conditional distribution of a dependent variable given one or more independent variables. Within each component the parameters that characterize the relationship between the independent and dependent variables assume distinct values. Early studies concerned normal conditional mixtures incorporating the linear regression model (Hosmer, 1974). Extensions and generalizations of this work include time-series models (Hamilton, 1990), binomial and multinomial probit and logit regression models, poisson regression models, and multivariate normal regression models. The reader is referred to Wedel and DeSarbo (1994) for a review of these developments. Wedel and DeSarbo (1995) present a generalized linear regression mixture model, which includes many of the models mentioned as special cases. The conditional distributions that make up the mixture include the common distributions of the exponential family (see Wedel & DeSarbo, 1995, Table 1). A computer program is available to fit Wedel and DeSarbo's conditional mixtures (Wedel, 1995).

Whereas the work mentioned so far concerns conditional mixtures, a growing amount of work has been devoted to covariance and mean structure modeling in unconditional normal multivariate mixtures. In contrast to the conditional mixture models, the emphasis here is on regression between observed and latent variables, and on regression among latent variables.

Blåfield (1980) presents an unconditional multivariate normal mixture model incorporating first and second order confirmatory factor models, but does not include a model for the means. Blåfield relies on a quasi-Newton routine to obtain maximum likelihood parameter estimates. Yung (1994, 1997) presents multivariate normal mixture model incorporating the confirmatory factor model with structured means (Sörbom, 1974). Yung takes into account complex sampling schemes and presents three approaches to the problem of estimation. Arminger and Stein (1997; Stein, 1997) present a mixture model that includes confirmatory factor models and structural regression models among latent variables. Their model includes the possibility to introduce fixed observed regressors (e.g., gender or income; see Stein, 1997; and Arminger & Stein, 1997, for an illustration). The inclusion of such fixed observed regressors allows one to replace the requirement of unconditional normality by the requirement of conditional normality within each component of the mixture. This approach then allows one to specify both conditional and unconditional normal mixtures subject to structural equation modeling. Arminger and Stein adopt a two stage procedure to estimation consisting of the EM algorithm and a weighted least squares (minimal distance) loss function. Jedidi, Jagpal and DeSarbo (1997a, 1997b), finally, present an unconditional normal mixture model that includes confirmatory factors models and a full structural equation model (the full LISREL model). They rely on the EM algorithm, with a iterative procedure in the M phase, to obtain maximum likelihood estimates.

The aim of this paper is to present an approach to covariance and mean structure modeling within unconditional multivariate normal mixtures that includes the models presented by Blåfield (1980) and Yung (1994, 1997) as special cases, and offers the same possibilities as Jedidi et al. to fit more elaborate (unconditional) models. In addition, linear and nonlinear may be imposed on the parameters in the model. As we demonstrate, such constraints are useful to express prior information. We limit our attention to the sampling scheme where the component membership of the cases is unknown. We consider two approaches to maximum likelihood estimation. On the one hand, we use the quasi-Newton algorithm, which incorporate exact gradients, but avoids the calculation of second order partial derivatives. On other hand, we implement a more simple procedure due to Yung (1994, 1997). The latter method is used to generate starting values for the former method. We present three illustrations based on real and simulated data.

In addition, we report results of a simulation study. It is well established that confidence intervals and standard errors are unreliable when components in a normal mixture are poorly separated (Yung, 1994, 1997). The aim of the simulation study is to compare confidence intervals based on the observed information and confidence intervals based on the likelihood profile (Azzalini, 1996; Neale & Miller, in press; Venzon & Moolgavkar, 1988) given two level of separation and varying sample sizes.

Multivariate Normal Mixture Subject to Structural Equation Modeling

Consider the P-dimensional random vector[1] $\mathbf{y}_i$ of subject i ($i = 1, \ldots, N$), which is characterized by the following density function:

$$f(\mathbf{y}_i; \mathbf{p}, \boldsymbol{\Sigma}, \boldsymbol{\mu}) = \sum_{k=1}^{R} p_k g_k(\mathbf{y}_i; \boldsymbol{\Sigma}_k, \boldsymbol{\mu}_k), \tag{1}$$

where the $(P \times R * P)$ matrix $\boldsymbol{\Sigma}$ equals $[\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \ldots, \boldsymbol{\Sigma}_R]$, the $(P \times R)$ matrix $\boldsymbol{\mu}$ equals $[\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_R]$, and the R dimensional vector $\mathbf{p}^T$ equals $[p_1, p_2, \ldots, p_R]$. The mixing proportions in $\mathbf{p}$ determine the proportion of subjects in each component of the mixture. These proportions are subject to the following constraints: $l_k \leq p_k \leq u_k$, where $0 < l_k < u_k < 1$ and $\Sigma\, p_k = 1$. The distribution of $\mathbf{y}_i$, shown in (1), is a mixture of R multivariate normals. The density within each component is ($k = 1, \ldots, R$):

$$g_k(\mathbf{y}_i; \boldsymbol{\Sigma}_k, \boldsymbol{\mu}_k) = (2\pi)^{-P/2}|\boldsymbol{\Sigma}_k|^{-1/2} \exp\,[-1/2(\mathbf{y}_i - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_k)], \tag{2}$$

where $|\boldsymbol{\Sigma}_k|$ denotes the determinant of the $(P \times P)$ covariance matrix $\boldsymbol{\Sigma}_k$, and $\boldsymbol{\mu}_k$ represents the P-dimensional mean vector. Within each component of the mixture, we specify the following LISREL submodel for the observations (Jöreskog & Sörbom, 1993):

$$\mathbf{y}_i = \boldsymbol{\nu}_k + \boldsymbol{\Lambda}_k\boldsymbol{\eta}_{ik} + \boldsymbol{\varepsilon}_{ik} \tag{3}$$

$$\boldsymbol{\eta}_{ik} = \boldsymbol{\alpha}_k + \mathbf{B}_k\boldsymbol{\eta}_{ik} + \boldsymbol{\zeta}_{ik}, \tag{4}$$

where $k = 1, \ldots, R$ and $i = 1, \ldots, N$. The components of the Q_k-dimensional random vector, $\boldsymbol{\eta}_{ik}$, are the common factors scores of the i-th case within the k-th component of the mixture. The components of the P-dimensional random vector $\boldsymbol{\varepsilon}_{ik}$ ideally represent

[1] All vectors are column vectors. Superscripted capital T denotes transposition.

measurement error terms that are distributed as zero-mean multivariate normals. The $(P \times Q_k)$ matrix Λ_k contains the loadings (i.e., regression coefficients) of the observed variables, $\mathbf{y}_i$, on the common factors, $\boldsymbol{\eta}_{ik}$. In (4), linear regressions are specified among the common factors. The $(Q_k \times Q_k)$ matrix $\mathbf{B}_k$ contains the (structural) regression coefficients. The components of the Q_k-dimensional vector $\boldsymbol{\zeta}_{ik}$, residual terms in these structural regressions, are zero mean multivariate normals. The components of the Q_k and P-dimensional vectors, $\boldsymbol{\alpha}_k$ and $\boldsymbol{\nu}_k$, finally, are means and intercepts, respectively.

The model for the observations implies the following model for the covariance and mean structures. The covariance matrix and mean vectors of the common factors are:

$$E[(\boldsymbol{\eta}_{ik} - E[\boldsymbol{\eta}_{ik}])(\boldsymbol{\eta}_{ik} - E[\boldsymbol{\eta}_{ik}])^T] = (\mathbf{I} - \mathbf{B}_k)^{-1}\boldsymbol{\Psi}_k(\mathbf{I} - \mathbf{B}_k^T)^{-1}$$

$$E[\boldsymbol{\eta}_{ik}] = (\mathbf{I} - \mathbf{B}_k)^{-1}\boldsymbol{\alpha}_k,$$

where $\boldsymbol{\Psi}_k$ is the $(Q_k \times Q_k)$ covariance matrix of the residuals $\boldsymbol{\zeta}_{ik}$, and $\mathbf{I}$ is the $(Q_k \times Q_k)$ identity matrix. The covariance matrix and mean vectors of the observed variables are:

$$\boldsymbol{\Sigma}_k = \Lambda_k(\mathbf{I} - \mathbf{B}_k)^{-1}\boldsymbol{\Psi}_k(\mathbf{I} - \mathbf{B}_k^T)^{-1}\Lambda_k^T + \boldsymbol{\Theta}_k \tag{5}$$

$$\boldsymbol{\mu}_k = \boldsymbol{\nu}_k + \Lambda_k(\mathbf{I} - \mathbf{B}_k)^{-1}\boldsymbol{\alpha}_k, \tag{6}$$

where $\boldsymbol{\Theta}_k$ is the $(P \times P)$ covariance matrix of the error terms $\boldsymbol{\varepsilon}_{ik}$. Neither $\boldsymbol{\Theta}_k$ nor $\boldsymbol{\Psi}_k$ is necessarily diagonal.

Let $\boldsymbol{\tau}_k$ represent the vector of free (to be estimated) parameters in Λ_k, $\mathbf{B}_k$, $\boldsymbol{\Psi}_k$, $\boldsymbol{\Theta}_k$, $\boldsymbol{\nu}_k$, and $\boldsymbol{\alpha}_k$. We indicate the parametrization of the covariance matrix and mean vector shown in Eqs. 5 and 6 by $\boldsymbol{\Sigma}_k\{\boldsymbol{\tau}_k\}$ and $\boldsymbol{\mu}_k\{\boldsymbol{\tau}_k\}$, respectively. We denote the mixture density subject to the parametrization within each component of the mixture as:

$$f(\mathbf{y}_i;\, \mathbf{p},\, \boldsymbol{\Sigma}\{\boldsymbol{\tau}\},\, \boldsymbol{\mu}\{\boldsymbol{\tau}\}) = \sum_{k=1}^{R} p_k g_k(\mathbf{y}_i;\, \boldsymbol{\Sigma}_k\{\boldsymbol{\tau}_k\},\, \boldsymbol{\mu}_k\{\boldsymbol{\tau}_k\}), \tag{7}$$

where $\boldsymbol{\tau}^T = [\boldsymbol{\tau}_1^T, \boldsymbol{\tau}_2^T, \ldots, \boldsymbol{\tau}_R^T]$, $\boldsymbol{\Sigma}\{\boldsymbol{\tau}\} = [\boldsymbol{\Sigma}_1\{\boldsymbol{\tau}_1\}, \boldsymbol{\Sigma}_2\{\boldsymbol{\tau}_2\}, \ldots, \boldsymbol{\Sigma}_R\{\boldsymbol{\tau}_R\}]$ and $\boldsymbol{\mu}\{\boldsymbol{\tau}\} = [\boldsymbol{\mu}_1\{\boldsymbol{\tau}_1\}, \boldsymbol{\mu}_2\{\boldsymbol{\tau}_2\}, \ldots, \boldsymbol{\mu}_R\{\boldsymbol{\tau}_R\}]$. And so, we have

$$g_k(\mathbf{y}_i;\, \boldsymbol{\Sigma}_k\{\boldsymbol{\tau}_k\},\, \boldsymbol{\mu}_k\{\boldsymbol{\tau}_k\}) = (2\pi)^{-P/2}|\boldsymbol{\Sigma}_k\{\boldsymbol{\tau}_k\}|^{-1/2} \cdot \exp\,[-1/2(\mathbf{y}_i - \boldsymbol{\mu}_k\{\boldsymbol{\tau}_k\})^T\boldsymbol{\Sigma}_k\{\boldsymbol{\tau}_k\}^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_k\{\boldsymbol{\tau}_k\})]. \tag{8}$$

The LISREL submodel shown in (5) and (6) encompasses a large number of models.[2] These include the confirmatory factor models presented by Yung (1994, 1997), the second order confirmatory factor models presented by Blåfield, (1980), multiple regression models and simultaneous equation models for observed variables (Jedidi, et al. 1996). As in Jedidi et al. (1997a, 1997b; Arminger & Stein, 1997, the present model includes multiple regression models and simultaneous equation models at the level of the latent variables. Yung (1994, 1997) fits confirmatory factor models subject to *form invariance*. This means that the confirmatory factor models in each component has the same set of parameter matrices with the same dimensions and the same location of fixed, free, and constrained parameters. In addition, Yung requires the number of common factors to be equal over the components. We impose neither of these conditions.

Given the parameters estimates, it is possible to assign the cases to the components

[2] The full LISREL model, which includes 8 parameters matrices and 4 parameter vectors (see Jöreskog & Sörbom, 1993) incorporates endogenous and exogenous latent variables. We sacrificed this conceptual distinction in favor of a reduction in programming burden. However, any model that can be fitted using the full model can also be fitted using the present submodel. See Jedidi, Jagpal & DeSarbo (1997a, 1997b) for a mixture approach incorporating the full LISREL model.

by calculating the posterior probability that a given case i belongs to a given class k. (Everitt & Hand, 1981, p. 10). By Bayes' theorem, this posterior probability equals:

$$\omega_{ki} = \frac{p_k g_k(\mathbf{y}_i; \boldsymbol{\Sigma}_k\{\boldsymbol{\tau}_k\}, \boldsymbol{\mu}_k\{\boldsymbol{\tau}_k\})}{f(\mathbf{y}_i; \mathbf{p}, \boldsymbol{\Sigma}\{\boldsymbol{\tau}\}, \boldsymbol{\mu}\{\boldsymbol{\tau}\})}. \tag{9}$$

Estimation

Methods of estimation in fitting mixtures are discussed extensively by Titterington, et al. (1985, chap. 4; see also Everitt & Hand, 1981, sec. 2.3). These include method of moments, graphical methods, and maximum likelihood estimation. In the presentation and development of methods for normal mixtures the focus has mainly been on estimation in the unconstrained case (i.e., without further modeling of covariance and mean structures). Maximum Likelihood (ML) estimation is the dominant method of estimation in fitting mixtures. Here we rely solely on ML estimation (e.g., see Azzalini, 1996).

ML estimates of $\mathbf{p}$ and $\boldsymbol{\tau}$ are obtained by maximizing the loglikelihood function within the admissible range of the parameter values, given the observed data, $\mathbf{Y}$, and the number of components, R:

$$L(\mathbf{p}, \boldsymbol{\tau}; \mathbf{Y}, R) = \sum_{i=1}^{N} \ln [f(\mathbf{y}_i; \mathbf{p}, \boldsymbol{\Sigma}\{\boldsymbol{\tau}\}, \boldsymbol{\mu}\{\boldsymbol{\tau}\})], \tag{10}$$

where $\mathbf{Y}$ is the $(N \times P)$ data matrix, $\mathbf{Y}^T = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N]$.

The likelihood equations associated with (10) cannot be solved in closed form so that some iterative scheme has to be invoked. The most popular methods of maximizing the loglikelihood function *in the unconstrained case* are the Expectation-Maximization (EM) algorithm and the Newton-Raphson (NR) algorithm (Titterington, et al., 1981, p. 84). The EM algorithm is easy to program (e.g., Everitt & Hand, 1981, p. 37) and has the advantage that it is sure to produce a monotone increase in loglikelihood (Everitt & Hand, 1981; Titterington, et al., 1985). The NR algorithm is more difficult to implement, because it requires the gradients and the Hessian. The Hessian can be replaced by the Fisher Information matrix, or an approximation thereof (Wolfe, 1970; Yung, 1994, 1997), or by an approximation based on an updating algorithm (Gill, Murray & Wright, 1981). In the first case the resulting algorithm is referred to as the Method of Scoring, in the latter case, the generic term quasi-Newton algorithm is used. Although these methods are generally faster than the EM algorithm, they are more sensitive to starting values in that they diverge and break-down when started at an infelicitous point.

In contrast to regular multigroup covariance structure analysis, the log-likelihood surface in normal mixtures is known to be "littered with singularities" (Titterington, et al., 1985, p. 83; Yung, 1994, p. 7). As mentioned in the Introduction, appropriate constraints greatly alleviate this problem (Hathaway, 1985). In addition, local maxima (e.g., see Everitt & Hand, 1981, p. 43; Jedidi, Jagpal & DeSarbo, 1997b), to which the EM algorithm and Newton-type algorithms are equally susceptible, may make the generation of starting values and the checking of solutions a tedious business.

In fitting normal mixtures subject to structural equation modeling, the accepted view of the EM algorithm as a simple way of maximizing the loglikelihood requires qualification. Yung (1994, 1997) presents EM algorithms to maximize the log-likelihood subject to confirmatory factor modeling within each component. In this evaluation of the EM algorithm (Yung, 1994, p. 37), he indicates that (a) the preparation required to implement the EM algorithm of a given model requires is great, and 2) slight changes to the specified model (e.g., the introduction of an equality constraint) may require nontrivial modifications of the EM algorithm. Yung's evaluation renders the EM algorithm less than ideal in fitting normal mixtures subject to structural equation modeling. Jedidi, Jagpal and DeSarbo (1997b; Wedel & DeSarbo, 1994, 1995) employ an EM algorithm which incorporates a iterative (conjugate gradient) method in the M step to obtain estimates of the parameters that feature in the structural equation model. This approach avoids the difficulties noted by Yung. Arminger and Stein (1997) employ a two step procedure which consists of the EM algorithm to obtain unconstrained estimates of the covariance matrices and mean vectors followed by a Newton-type algorithm to minimize a weighted least squares loss function.

When relying solely on the Newton-Raphson or related approaches to fit multivariate normals certain problems arise that are absent in ordinary multigroup SEM. The Method of Scoring requires the evaluation of the Fisher information matrix at each iteration. Because this matrix is difficult to obtain analytically (Everitt & Hand, 1981, p. 40; Titterington, et al., 1985, p. 88; Wolfe, 1971), Yung (1994, 1997) employs the Method of Scoring with a simplification of the Information matrix due to Wolfe. This simplification involves setting the posterior probabilities of case i equal to 1.0 in the component where the posterior probability, ω_{ki} ($k = 1, \ldots, R$), is maximal and to zero elsewhere. This simplification results in a tractable expression for the Information matrix. Yung refers to this method as the *Method of Approximate Scoring*.

Besides his EM algorithms for constrained mixtures and the Method of Approximate Scoring, Yung (1994, pp. 35–36) presents a two-stage procedure called *2 stage GLS*. In this approach, the unconstrained means and covariance matrices are estimated using the EM algorithm or the Method of Approximate Scoring. Subsequently, these summary statistics are used as input in a standard multigroup covariance structure modeling program such as LISREL 8 (Jöreskog & Sörbom, 1993), or Mx (Neale, 1995), and are analyzed by minimizing the normal theory generalized least squares, or likelihood ratio function (Jöreskog & Sörbom). The specified sample size depends on the estimated mixing proportions. This two stage procedure is very simple to implement and very flexible, because one can specify any model available in the programs mentioned. As discussed by Yung (1997, p. 37), the disadvantages of the two stage procedure are that it does not produce ML estimates, and that standard errors and χ^2 goodness of fit index produced by the standard programs are not correct. These have to be evaluated using separate programs. When the components in the mixture are well separated, however, the results are similar to those obtained in regular multigroup structural equation modeling. It should be emphasized that Arminger and Stein's two stage procedure does produce correct results. They do not treat the components are independent groups, following the first (EM) stage of their estimation procedure. Rather they calculate a weight matrix for the second part of their procedure (WLS), that takes into account the intercorrelation of the parameters estimated in the various components.

Blåfield (1980) is not concerned with second order partial derivatives. Rather, he employs the Davidon-Fletcher-Powell update within a quasi-Newton algorithm to maximize the loglikelihood using exact gradients. The advantage of this approach is that it is easy to implement. In addition, it produces, as a by-product of the optimization process, an approximation to the Hessian.

Our present approach to maximizing the log-likelihood ratio function is based on both Yung (1997, 1994) and Blåfield (1980). We employ two methods to maximize the loglikelihood function. We use the NAG FORTRAN library routine E04VDF (NAG, 1990) to maximize the loglikelihood function using exact gradients (see Appendix 1). This quasi-Newton approach is closely related to Blåfield's. In addition, we use a method closely related to Yung's two-stage procedure to obtain starting values. We first fit unconstrained multivariate normal mixtures using the EM algorithm. The estimates of mean vectors and covariance matrices of each component are subsequently treated as independent in a regular multi-group covariance structure analysis (Jöreskog, 1971). Here we obtain estimates by minimizing the multi-group likelihood ratio function, again using E04VDF with exact gradients. Estimates so obtained are identical to those produced by Yung's two-stage procedure. These estimates are then used as starting values in maximizing the log-likelihood function, (10), using E04VDF.

An advantage of the routine E04VDF is that it offers the possibility of imposing non-linear and simple bounds on the parameters in $\boldsymbol{\tau}$ and $\mathbf{p}$. In E04VDF, these constraints are accommodated using Lagrange multipliers (Gill, Murray & Wright, 1981). In addition, the extension to multi-groups (i.e., multigroup normal mixtures subject to SEM) are quite easy to program using the present approach. Further facilities available in E04VDF to impose equality constraints are not exploited. It is more efficient to accommodate equality constraints by concentrating the loglikelihood function. In the present case this means that we specify an equality constraint by inserting a given parameter in two or more positions in the parameter matrices in (5) and (6).

Standard Errors and Confidence Intervals

We calculate standard errors by approximating the Hessian using central differences and exact gradients. This approximation, which is very simple to program, is known to be good in regular multi-group structural equation modeling (Dolan & Molenaar, 1991). These standard errors are based on the observed information instead of the expected information, but these are asymptotically equivalent (Azzalini, 1996, p. 91). We use these standard errors to obtain a rough indication of the precision of the ML estimates. Subsequently we calculate likelihood-based confidence intervals (Azzalini, 1996, sec. 4.5.3; Venzon & Moolgavkar, 1988) for the subset of parameters of special interest using a method suggested by Neale and Miller (1997). Let L^* denote the maximum of the log-likelihood function, (10), for a given model and let τ^* denote a parameter of special interest. Following Neale and Miller (1997), we minimize the following functions to obtain the upper (f_u) and lower (f_l) endpoints of the confidence interval (CI) of τ^*:

$$f_l(\boldsymbol{\tau}, \mathbf{p}) = w[L^* - L(\mathbf{p}, \boldsymbol{\tau}; \mathbf{Y}, R) + .5c]^2 + \tau^*, \tag{11}$$

$$f_u(\boldsymbol{\tau}, \mathbf{p}) = w[L^* - L(\mathbf{p}, \boldsymbol{\tau}; \mathbf{Y}, R) + .5c]^2 - \tau^*. \tag{12}$$

In these equations, the parameter of interest features twice: once as a component of parameter vector $\mathbf{p}$ or $\boldsymbol{\tau}$, and one as τ^*. At the minimum of these functions, τ^* assumes the minimum (maximum) value for which minus twice the log-likelihood ($-2[L^* - L(\mathbf{p}, \boldsymbol{\tau}; \mathbf{Y}, R)]$) equals $c + 1/[w(\partial\tau^*/\partial L(\mathbf{p}; \boldsymbol{\tau}; \mathbf{Y}, R))]$. Now, the desired width of the CI can be obtained by setting c to equal $\chi^2_{\alpha,1}$, the value of the cumulative chi-square distribution, given 1 degree of freedom and the significance level, α. For instance, if α equals 0.05, c equals 3.84, that is, the values of τ^* obtained are the approximate endpoints of the 95% confidence intervals of the parameter of interest. These are not the exact endpoint, because the term $\partial\tau^*/\partial L(\mathbf{p}, \boldsymbol{\tau}; \mathbf{Y}, R)$ will not generally equal zero. As pointed out by Neale and Miller (1997), this bias can be evaluated easily and can be reduced by setting w to equal a positive number. In practice, with $w = 1$, we find that the bias is quite negligible. When we do desire an improvement, we set $w = 5$. We refer to these CI's as likelihood-based CI's. Asymptotic $100(1 - \alpha)\%$ CI's may also be calculated using the standard errors: $\tau^* \pm se(\tau^*) * \phi(\alpha/2)$, where $\phi(\alpha)$ is the value of the standard normal distribution function associated with the significance level α. We refer to these CI's as Hessian-based CI's.

Hessian-based CI's are expected to be inferior to likelihood-based CI's, because the former are based on a quadratic approximation of curvature of the loglikelihood function at its maximum. As the likelihood-based *CI*'s do not depend on such an approximation, they are more accurate. For instance, likelihood-based *CI*'s are not necessarily symmetric about the ML estimate. In addition Hessian based *CI*'s may assume nonsensical upper or lower bounds. Multidimensional *CI*'s can also be based on the loglikelihood instead of the Hessian (Venzon & Moolgavkar, 1988), but we do not consider these here.

In the presence of non-linear constraints, standard errors of the parameter estimates can be obtained by inverting the Hessian after it has augmented with the Jacobian matrix of the constraints (Aitchison & Silvey, 1958). The presence of nonlinear constraints does not complicate the calculation of likelihood-based *CI*'s (Neale & Miller, 1997).

Hypothesis Testing and Identification

We distinguish hypotheses relating to the number of components in the normal mixture and hypotheses relating to the restrictions placed upon the covariance matrices, mean vectors and proportions. As pointed out by Yung (1994, 1997), and Jedidi, Jagpal and DeSarbo (1997a, 1997b) the latter type of hypothesis can be tested using the generalized likelihood ratio test as long as competing hypotheses are nested. In covariance structure modeling, an unrestricted mean vector and covariance matrix usually features as a baseline to evaluate a model incorporating restrictions, assuming the more restrictive model is identified. The differences in the number of parameters estimated are the degrees of freedom (df) and minus twice the difference between the loglikelihoods provides a χ^2 test for the restricted model. Examples of this application of the likelihood ratio test are given below. The identification of structural equation models in regular multi-group covariance structure modeling requires consideration, but can usually be resolved easily. Many commonly employed models are known to be identified.

Identification of structural equation models within multivariate normal mixtures does not pose a problem. An obvious necessary condition is that the structural equation model is identified in a regular multi-group analysis. Normal mixtures are typically identified (Titterington, et al., 1985, p. 162). Concerning identification, Jedidi, Jagpal and DeSarbo (1997a) provide a proof that the established identification of the structural equation model and the requirement of normality within each component are sufficient and necessary conditions for identification of the model within the multivariate normal mixture.

Hypotheses relating to the number of components in a mixture cannot be tested using the likelihood ratio test, because a regularity condition for the generalized likelihood ratio test (e.g., Azzalini, 1996, p. 71) does not hold (Everitt & Hand, 1981, sec. 5.2.2). Wedel and DeSarbo (1995, 1994; see also Jedidi, et al. 1996; Jedidi, Jagpal, & DeSarbo, 1997a, 1997b) rely on information criteria based Akaike's information criterion as informal indicators of the number of components of the mixture. Although these criteria are themselves based on the loglikelihood, they appear to be useful (see Jedidi, Jagpal & DeSarbo, 1997a). Monte Carlo procedures have been suggested to determine the number components (Arminger & Stein, 1997; Feng & McCulloch, 1996; McLachlan, 1987), but these are computationally intensive. Goodness of fit procedures based on the Pearson χ^2 test and the Kolmogorov-Smirnoff test have recently been suggested (Agha & Branker, 1997), but these have yet to be applied in the context of multivariate normal mixture subject to SEM.

Separation of Components

In fitting regular multigroup structural equation models as well as in fitting normal mixtures, validity of asymptotic results, relating to standard errors and the χ^2 goodness of fit index, depends on distributional aspects of the data, sample size, and the accuracy of the model under consideration. In fitting normal mixtures, however, an additional consideration of great importance is the separation of the components. When separation is poor, the χ^2 goodness-of-fit measure and standard errors cannot be trusted (Everitt & Hand, 1981, p. 44ff.; Yung, 1994, p. 49).

Two measures relating to the degree of separation have been suggested. Wedel and DeSarbo (1994; Jedidi et al. 1996) use an entropy measure that represents the degree of fuzziness in component membership and is bounded by zero and one (perfect separation). Yung (1994) assesses the separation between each pair of components in a multicomponent mixture by calculating a multivariate version of Hosmer's measure of separation (Hosmer, 1974):

$$d_{ij} = \max_{h \in \{i,j\}} [(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_h^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)]^{1/2}. \qquad (13)$$

Because Yung's measure provides more information, especially considering his simulation results (Yung, 1994), we report these. Results of Yung's simulation study suggest that the asymptotic theory works well for values of d_{ij} of about 3.8 or over. Below we investigate the effects of separation of likelihood-based and Hessian-based confidence intervals.

Illustrations

We illustrate the fitting of multivariate normal mixture subject to constraints using 3 data sets.

We report ML estimates, standard errors, and likelihood-based 95% *CI*'s. Although they are redundant, we also report Hessian-based 95% *CI*'s to facilitate comparison with the likelihood-based *CI*'s.

Murphy-Bolling Isoniazid data: a univariate normal mixture with nonlinear constraints (Murphy & Bolling, 1967). Hamilton (1991) uses these data to illustrate his quasi-Bayesian method of fitting normal mixtures. Here we use these data to illustrate the use of non-linear constraints. A sample of 220 subjects was administered a tuberculostatic drug. After 6 hours the concentrations of the drug in the blood were measured. Individual differences in metabolic rate are hypothesized to be determined by a single locus (gene) with two alleles, *F* (fast rate) and *S* (slow rate). Under full dominance, homozygotic *FF* individuals are indistinguishable from heterozygotic *FS* individuals concerning metabolic rate. Full dominance appears to be the accepted model, as documentation of the drug speaks only of a slow and fast rate (Farmacotherapeutisch Kompas, 1994). Like Hamilton (1991), we are interested in fitting the three component model.

A problem with these data the poor separation of two components, as can be seen in Figure 1. The *SS* distribution is clearly discernible, but, assuming the absence of complete dominance, the distributions of the *FS* and *FF* cases are hard to distinguish. Although the histogram is quite compatible with a three component mixture (see Titterington, et al., 1985, Fig. 4.1.1), the degree of overlap makes the three component mixture hard to fit (Murphy & Bolling, 1967; Hamilton, 1991).

Assuming the single gene model without dominance, we can fit a mixture of three normals subject to two non-linear constraints that are derived from a simple biometric model. Let a and $1 - a$ denote the allele frequencies of S and F in the population. We assume that the sample is representative and that environmental effects do not contribute to the between-phenotype variance. The means of each phenotype are then $\mu(SS) = [\gamma - a^2\delta]$, $\mu(FS) = [\gamma]$, and $\mu(FF) = [\gamma + (1 - a)^2\delta]$. The parameter δ is the genotypic effect and the parameter γ is the so-called mid-parent value, which can be interpreted as a constant from which the effects of the alleles are expressed. Furthermore, the proportions of each component of the mixture are $\mathbf{p}^T = [a^2 \;\; 2a(1 - a) \;\; (1 - a)^2]$. The model

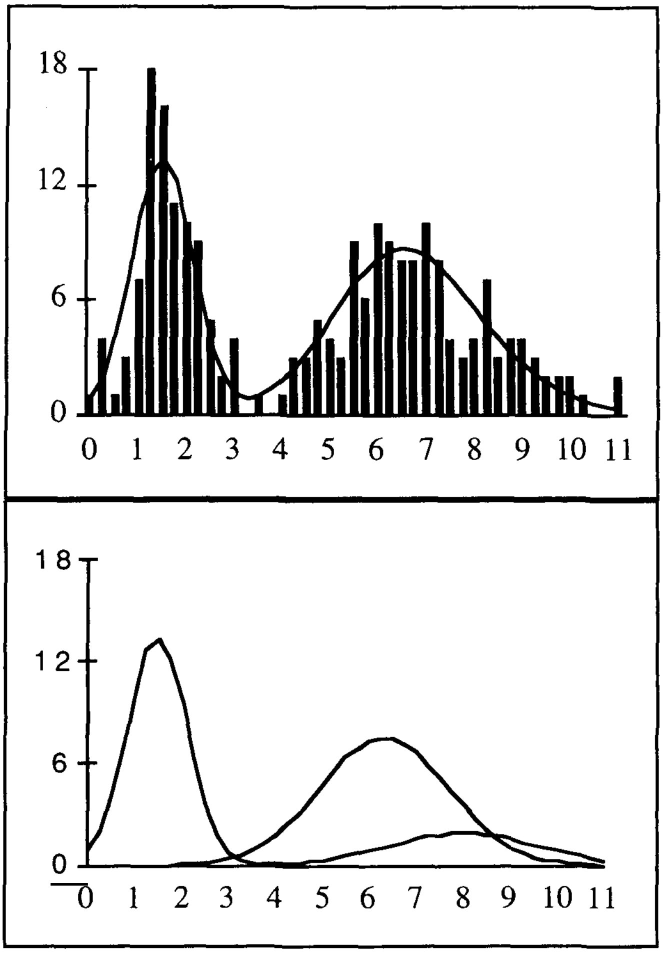

FIGURE 1.
Top: Histogram of the Murphy-Bolling Isoniazid data ($N = 220$; Adapted from Hamilton, 1991) with superimposed normal distributions of Model 2. Bottom: Components of Model 2.

implies that the proportions can be modeled using a single parameter (instead of 2), namely a, and that the means can be modeled using 2 parameters (instead of 3).

Table 1 contains the parameter estimates for the mixture of two component (Model 0) and the mixture of three normals, without (Model 1) and with (Model 2) the restrictions mentioned.[3] The estimates reported by Hamilton using his quasi-Bayesian method of estimation (Model 3) and the two component solution are included for comparison. Comparing model 1 and model 2, we find we cannot reject the nonlinear constraints: minus twice the log-likelihood equals $\chi^2(2) = 3.14$ ($p = .20$). In inspecting the estimates, the only striking differences are in the variances of the second and third component. With the nonlinear constraints, these variances are larger (1.87 vs. 1.25 and 2.39 vs. 1.05). There are considerable differences in the estimated *CI*'s in model 2. In three cases the Hessian based *CI*'s differ from the likelihood-based *CI*'s. In these three cases the latter hit a lower bound. In Model 3, except for the variance in the third component, the *CI*'s agree quite well, even

[3] The data were derived from the histogram published in Hamilton (1991). Hamilton demonstrates that the inherent loss of information does not seriously affect the results.

Table 1: Parameter estimates, standard errors, and 95% CI's (upper/lower endpoints) for the Murphy-Bolling data. Standard errors in parentheses.

parameter	Model 0	Model 1	Model 2	Hamilton
μ_1 (SS)	1.71 (.072)	1.74 (.071)	1.72 (.071)	1.75
CI (s.e)	1.57/1.85	1.60/1.88	1.58/1.86	
CI (logL)	1.57/1.86	1.60/1.88	1.58/1.86	
σ^2_1 (SS)	0.40 (.066)	.43 (.068)	.41 (.067)	.45
CI (s.e)	0.27/0.53	.29/.56	.28/.54	
CI (logL)	0.29/0.56	.31/.59	.30/.57	
p_1 (SS)	$.40^*$	.41 (.033)	.39 (.036)	.41
CI (s.e)		.35/.48	.32/.46	
CI (logL)		.35/.48	.33/.47	
μ_2 (SF)	6.92 (.15)	6.35 (.32)	6.56 (.18)	6.14
CI (s.e)	6.62/7.23	5.73/6.98	6.21/6.92	
CI (logL)	6.59/7.22	5.33/6.96	6.18/6.90	
σ^2_2 (SF)	2.68 (.40)	1.25 (.48)	1.87 (.49)	1.06
CI (s.e)	1.90/3.46	.31/2.20	.90/2.84	
CI (logL)	2.04/3.68	$.01^{\#}$/2.73	.99/3.12	
p_2 (SF)	.60 (.034)	.45 (.10)	.47 (.015)	.36
CI (s.e)	.53/.66	.25/.64	.44/.50	
CI (logL)	.53/.66	$.01^{\#}$/.60	.43/.49	
μ_3 (FF)	-	8.98 (.75)	$8.25^{\dagger}$ (.31)	8.31
CI (s.e)		7.51/10.44	7.64/8.86	
CI (logL)		6.74/11.36	7.69/8.97	
σ^2_3 (FF)	-	1.05 (.72)	2.39 (1.03)	1.74
CI (s.e)		-.36/2.47	.36/4.42	
CI (logL)		$.01^{\#}$/4.31	.87/6.12	
p_3 (FF)	-	$.14^*$	$.14^*$	.23
d_{12}	8.23	7.02	7.54	
d_{13}	-	11.07	10.17	
d_{23}	-	2.55	1.24	
logL	-477.358	-475.059	-476.628	

Note:
*An asterisk indicates that the parameter is subject to a linear constraint (p_1 in model 0, p_3 elsewhere). †A dagger indicates that the parameters is non-linearly constrained. #Indicates that the lower bound on the parameter was hit during optimization.

Model 0: two component model with standard constraint ($\sum p_i=1$). Under this model SF and FF are indistinguishable. Model 1: three component model with standard constraint ($\sum p_i=1$). Model 2: standard constraint & 2 nonlinear constraints.

though the distance between the second and third components is smaller than in model 2 (d_{23} = 2.55 vs. d_{23} = 1.24). In view of the poor separation, the *CI*'s cannot be trusted.

The estimates of Model 2 differ slightly from those reported by Hamilton (1991) in

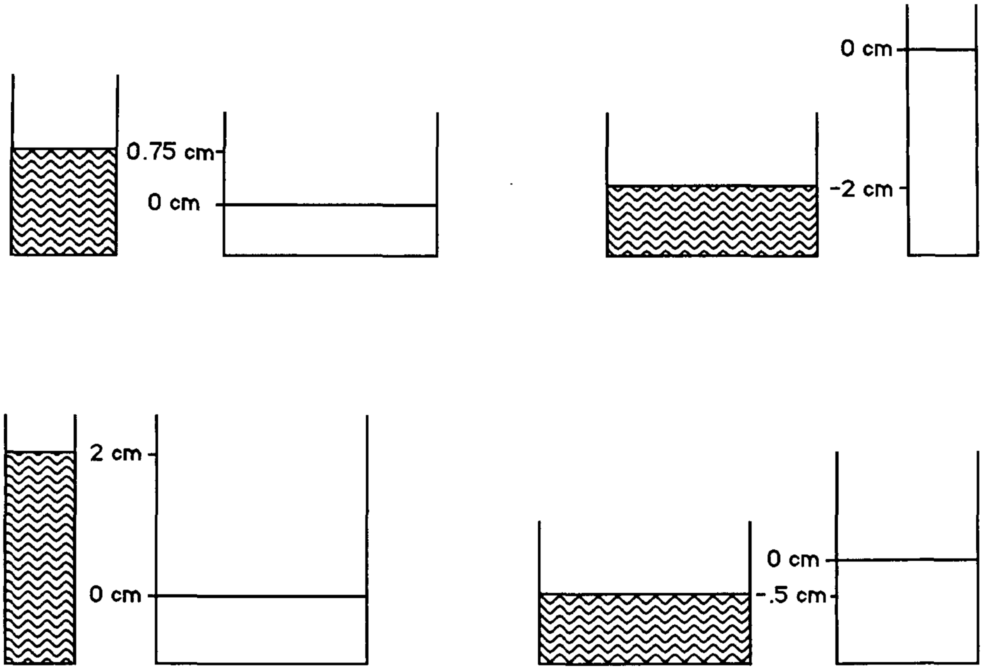

FIGURE 2.
Four liquid conservation items.

that the variances of the second and third component are larger (1.87 vs. 1.06 and 2.39 vs. 1.74). Furthermore, Hamilton assigns more individuals to the third component (.14 vs. .23). Figure 1 is a graphical representation of the fitted components.

Qualitative developmental change: Analysis of conservation data. Van der Maas (1993) designed a nonverbal computer-based test to assess the ability to conserve continuous quantity, an ability that is characteristic of the concrete operational stage in Piaget's structural theory of cognitive development (Piaget & Inhelder, 1969). This test, which is concerned with pouring equal amounts, consists of 4 items that are reproduced in Figure 2. The subjects are required to imagine that the liquid in the transparent vessel on the left is poured into the transparent vessel on the right. They are instructed to indicate their expectation concerning the level of the liquid in the vessel to the right by moving the level of the liquid up or down using designated keys on the computer keyboard. The movable level is represented by a thin black line. The correct level is indicated in Figure 2 by "0 cm" (i.e., zero centimeters). Subjects in the pre-operational stage are hypothesized to set the expected level to the level in the original vessel, that is, they simply align the levels. Subjects in the concrete operation stage are expected to realize that the dimensions of the vessel will affect the observed level of the liquid and act accordingly.

A total of 90 children ranging in age from 6.5 to 11 years completed the computer test within their school settings (see van der Maas, 1993, chap. 2). We specify a two component mixture as the subjects are expected to be in the concrete operational stage (conservers), or in the pre-operational stage (nonconservers). The mean vector of the conservers, $\boldsymbol{\mu}_C$, is expected to equal $[0\ 0\ 0\ 0]^T$, and the mean vector of the nonconservers, $\boldsymbol{\mu}_{NC}$, is expected to equal $[.75\ -2\ 0\ -.5]^T$, where the subscript NC stands for nonconserver and C for conserver. We do not initially impose any restrictions on the covariance matrices. Rather than estimating the covariance matrices, however, we do specify $\Sigma_C = \Lambda_C \Psi_C \Lambda_C^T$ and

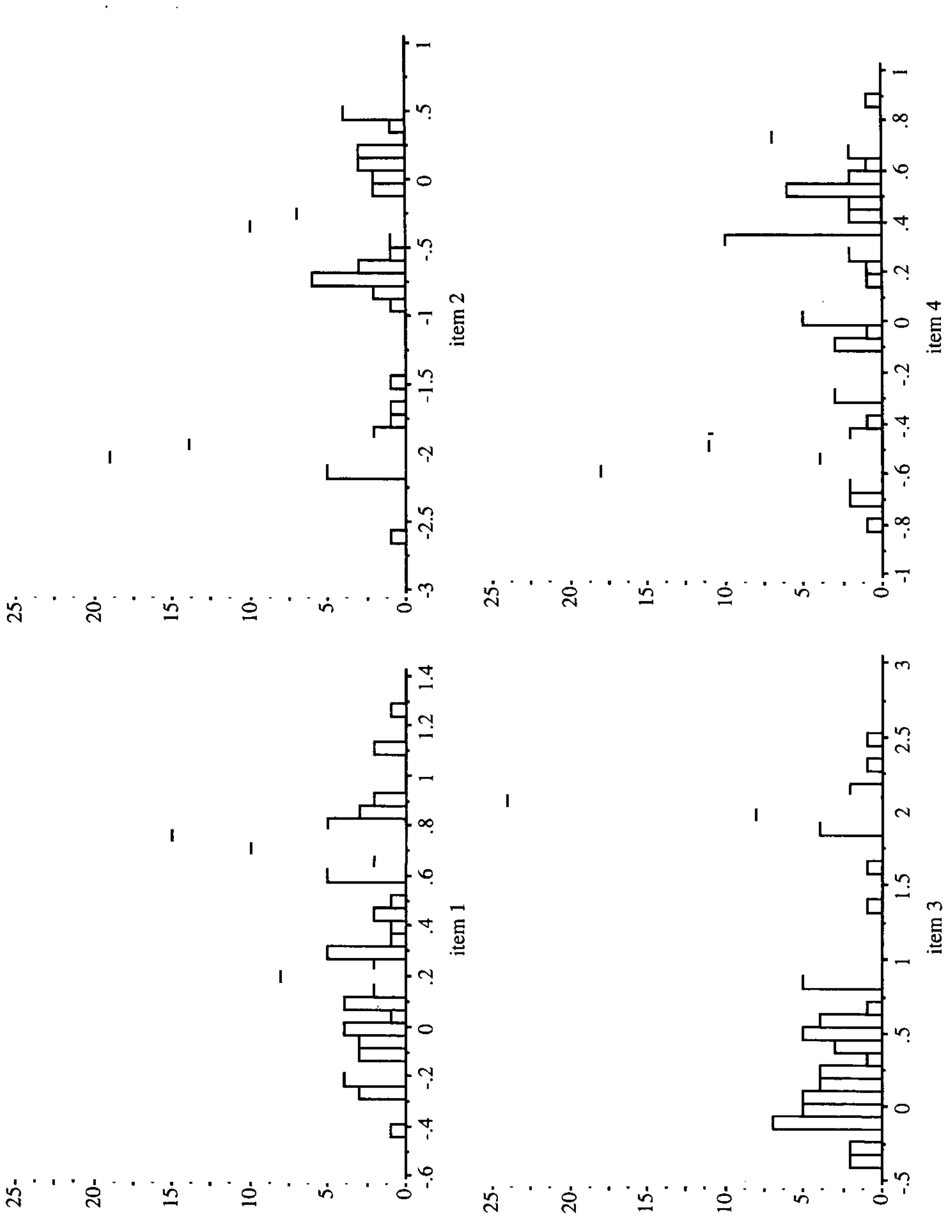

FIGURE 3.
Histograms of conservation data (N = 90).

$\boldsymbol{\Sigma}_{NC} = \boldsymbol{\Lambda}_{NC}\boldsymbol{\Psi}_{NC}\boldsymbol{\Lambda}_{NC}^T$. The diagonal 4 × 4 matrix $\boldsymbol{\Lambda}_{C(NC)}$ contains the standard deviations and the symmetric 4 × 4 matrix $\boldsymbol{\Psi}_{C(NC)}$ is a correlation matrix. This parametrization is convenient, because it yields estimates and standard errors of the correlations among the 4 items rather than the covariances. Histograms of the data are shown in Figure 3.

We start by fitting the model with and without the fixed means. The log-likelihoods equal −49.28 and −11.04, respectively so we reject the hypothesis concerning the means ($\chi^2(8) = 76.5$ ($p < 0.001$). Parameter estimates are shown in Table 2 for the model with unconstrained means. Both visual inspection of Figure 4, and Yung's separation measure ($d_{12} \approx 38$) indicate that the components are very well separated.

Analysis of the means indicates that all means in the identified C group deviate from their expected values (two-sided univariate t-tests, $\alpha \approx .006$, i.e., 0.05/8). In the NC group the means of the third and fourth items differ from their expected values. In comparing the variances, it is evident that the C group is a lot more variable than the NC group. It is likely that the items are easier for members of the NC group: A typical subject is assumed to merely align the level in the vessel to the level in the original vessel. In addition, the C group may not be quite as homogeneous as we suppose it is.

The correlations in the C group appear to be due to overestimation on Item 1 and 3 and underestimation on Item 2. Except for the mean on Item 4, which is clearly more difficult, the mean values are compatible with this idea of over- and underestimation. In the NC group, none of the correlations appears to be significantly greater than zero, judging by the standard errors. If the NC children are carrying out a simple alignment, the most obvious hypothesis is that the covariance matrix is diagonal.

We refit the model, subject to the following constraints. The covariance matrix the NC group is constrained to be diagonal. In the C group, we specify a constrained oblique two factor model. Given this hypothesis, the density is:

$$f(\mathbf{y}_i;\ \mathbf{p},\ \boldsymbol{\Sigma}\{\boldsymbol{\tau}\},\ \boldsymbol{\mu}\{\boldsymbol{\tau}\}) = p_C g_C(\mathbf{y}_i;\ \{\boldsymbol{\Lambda}_C\boldsymbol{\Psi}_C\boldsymbol{\Lambda}_C^T + \boldsymbol{\Theta}_C\},\ \boldsymbol{\nu}_C) + p_{NC} g_{NC}(\mathbf{y}_i;\ \{\boldsymbol{\Lambda}_{NC}\boldsymbol{\Lambda}_{NC}^T\},\ \boldsymbol{\nu}_{NC}),$$

where ($p_C + p_{NC} = 1$). The 4 × 4 diagonal matrix $\boldsymbol{\Lambda}_{NC}$ contains the standard deviations of the scores on the items. The model matrices in the C group are diag $[\boldsymbol{\Theta}_C] = [\sigma_{\varepsilon 1}^2\ \sigma_{\varepsilon 2}^2\ \sigma_{\varepsilon 3}^2\ \sigma_{\varepsilon 4}^2]$ and

$$\boldsymbol{\Lambda}_C = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \\ \lambda_2 & 0 \\ 0 & \lambda_1 \end{bmatrix} \qquad \boldsymbol{\Psi}_C = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

The equality constraints imposed on the factor loadings are inspired by the fact that Items 2 and 3 are essentially of the same type (see Figure 2). Item 1 does not resemble item 4 to the same degree, but these items are similar. In fitting the model, the parameter $\sigma_{\varepsilon 3}^2$ kept hitting the lower bound (.0001) and the parameter $\sigma_{\varepsilon 2}^2$ kept assuming very small values. We therefore specified diag $\{\boldsymbol{\Theta}_C\} = [\sigma_{\varepsilon 1}^2\ \mathbf{0}\ \mathbf{0}\ \sigma_{\varepsilon 4}^2]$. The log-likelihood ratio for this model equals −18.56, so we have a χ^2 of 15.0 ($-2 * [11.04 - 18.56]$) with 29 − 18 = 11 df ($p = .18$). There is no reason to reject the restrictions imposed. The estimate of $\boldsymbol{\tau}_C^T = [\lambda_1\ \lambda_2\ \rho\ \sigma_{\varepsilon 1}^2\ \sigma_{\varepsilon 4}^2]$ equals [.21(.037) .72(.059) −.68(.072) .10(.019) .13(.024)] (standard errors in parentheses). The mixing proportion equals .61 (.05). The fitted moments in the C group are shown in Table 2. Standard deviations in the NC group equal .056 (.006), .071 (.008), .068 (.009), and .072 (.009).

Besides the item scores, IQ scores (Raven Progressive Matrices) and age are available. We now investigate the relationship between these variables and the item scores in the NC group and the scores on the common factors in the C group. In the NC group both the item scores and the covariates are treated simply as observed variables. In the C group,

Table 2: Parameter estimates for analysis of conservation data (N=90), standard errors in parentheses, 95% CI's for means (upper/lower end point).

Unconstrained model (logL = -11.04).Identified concrete operational group, proportion =.59 (s.e. .052), N ≈ 53.
correlation matrix (Ψ)

	1^{*}			
	-.25 (.13)	1^{*}		
	0.45 (.11)	-.66 (.08)	1^{*}	
	-.11 (.13)	0.53 (.10)	-.31 (.12)	1^{*}
st. dev.'s	.36 (.035)	.73 (.070)	.63 (.061)	.42 (.040)
μ (means)				
expected	0.0	0.0	0.0	0.0
observed	.20† (.05)	-.53†(.10)	.42† (.09)	.25†(.057)
CI (s.e)	.11/.30	-.73/-.33	.25/.59	.14/.37
CI (logL)	.10/.30	-.73/-.33	.24/.59	.14/.37

Identified pre-operational group, proportion =.41, N ≈ 37.
correlation matrix

	1^{*}			
	-.16 (.16)	1^{*}		
	.32 (.15)	.01 (.17)	1^{*}	
	-.03 (.16)	.04 (.16)	-.30 (.15)	1^{*}
st dev.'s	.06 (.007)	.07 (.008)	.11 (.014)	.07 (.008)
μ (means)				
expected	.75	-2	2	-.5
observed	.74 (.010)	-2.01(.012)	2.05†(.019)	-.54† (.012)
CI (s.e)	.72/.76	-2.03/-1.99	2.02/2.07	-.57/-.52
CI (logL)	.71/.76	-2.03/-1.99	2.01/2.07	-.57/-.52

Fitted moments for C group. Constrained model (logL=-18.56). Proportion =.61 (s.e. .051), N ≈ 55.
Correlation matrix

	1			
	-.39	1		
	.54	-.68	1	
	-.20	.51	-.35	1
St. dev.'s	.37	.72	.72	.41
Means	.21 (.05)	-.56 (.09)	.45 (.09)	.24 (.055)

†Deviates significantly ($|t| > t_{\alpha/2;N-1}$, α=.05/8, .006) from expected values according to the two-sided univariate t-tests based on the reported estimates of means and standard deviations and expected values of the means. *Parameters fixed to values reported.

we regress the observed covariates on unobserved, latent variables. The covariates, which correlate .48 in the entire sample, were examined separately. The model for a given item i in the *NC* group is (discarding subject subscripts; the model for age is identical):

$$I_i - \mu(I_i) = \beta_i[IQ - \mu(IQ)] + \zeta(I_i), \qquad i = 1, \ldots, 4$$

Table 3: ML parameter estimates, standard errors (in parentheses), and 95% CI's (lower/upper endpoint) for the regression on IQ and AGE of factor and items scores.

covariate:	IQ		AGE	
C group	estimate	%var	estimate	%var
var (cov.)	.84(.16)		.14 (.027)	
mean (cov.)	3.18(.12)		3.03 (.051)	
ß (on η_1)	-.048(.027)	6%	-.097 (.067)	3.5%
CI (s.e)	-.10/.005		-.230/.034	
CI (logL)	-.10/.004		-.402/.031	
ß (on η_2)	.052 (.028)	7%	.114 (.068)	5.5%
CI (s.e)	-.003/.10		-.021/.246	
CI (logL)	-.001/.11		-.014/.404	
NC group	estimate	%var	estimate	%var
var(cov.)	.65 (.15)		.14 (.032)	
mean(cov.)	2.41 (.13)		2.83 (.061)	
ß (Item 1)	.007 (.012)	<1%	.002 (.027)	0%
CI (s.e)	-.017/.032		-.051/.054	
CI (logL)	-.018/.030		-.052/.057	
ß (Item 2)	-.050 (.012)	32%	-.042 (.031)	4.8%
CI (s.e)	-.074/-.027		-.103/.017	
CI (logL)	-.074/-.025		-.104/.020	
ß(Item 3)	0.013 (.023)	<1%	-.032 (.049)	1.2%
CI (s.e)	-.033/.057		-.130/.063	
CI (logL)	-.033/.059		-.132/.067	
ß (Item 4)	0.010 (.015)	1%	0.032 (.031)	2.2%
CI (s.e)	-.020/.037		-.031/.093	
CI (logL)	-.020/.039		-.032/.095	

where I_i stands for item i. In the C group, the each common factor is regressed on IQ as follows:

$$\eta = \beta[IQ - \mu(IQ)] + \zeta(\eta).$$

Table 3 contains the results for IQ and age. In each analysis all parameters were estimated simultaneously, including those that featured in the previous analysis. The covariate IQ explains about 6% and 7% of the variance in the common factors in the C group. Judging by the *CI*'s, the regression coefficients do not deviate from zero. In the NC group we find the unexpected result that IQ explains a substantial portion (32%) of the variance associated with item 2. In all other cases in the NC group, the percentage of explained variance is negligible. Refitting the model with all regression coefficient fixed to zero results in a loglikelihood of -133.36 and a $\chi^2(6)$ of 17 ($p = 0.01$). Limiting the regression to the second item in the *NC* groups, we obtain loglikelihood of -127.6 and a $\chi^2(5)$ of 5.5. A visual inspection of the relevant scatterplot in the *NC* group suggests that the significant regression in the *NC* group cannot be attributed to any outlying cases.

Results for the covariate age are quite clear. The percentage of explained variance is quite small in both components (in the *C* group, 3.5% and 5.5%; in the *NC* group, between 0% and 4.8%). The loglikelihood for this model equals -54.8. Refitting the model with all regression coefficients fixed to zero, we find a log-likelihood of -58.08 and a $\chi^2(6)$ of 6.6. Age and IQ (with the exception mentioned) appear to do little to explain the within-component variance.

Overall, the results are quite compatible with the hypothesis that the *NC* subjects are carrying out an alignment. In the *C* group, the results suggest that the subjects know the correct responses to the items, but are prone to over- or underestimation depending on the item. This interpretation is based on the provisional assumption that the *C* group is homogeneous.

An illustration based on the quasi-Markov simplex model: An analysis of simulated data. We simulate two longitudinal datasets according to a first order autoregression comprising 6 equidistant measurement occasions. In the first group, the time series is stationary throughout. In the second group, the time series is identical in means and covariance structure up to occasion 3, but thereafter the means and covariance structure change. From occasion 4 onwards, the time series in the second group is again stationary, but is characterized by a different mean and different autoregressive parameter.

In the first group, the model is as follows (Jöreskog, 1970; subject subscript discarded; subscript j denotes occasion):

$$\eta_1 - \mu(\eta_1) = \zeta_1,$$

$$\eta_j - \mu(\eta_j) = \beta_1[\eta_{j-1} - \mu(\eta_{j-1})] + \zeta_j, \qquad j = 2, \ldots, 6$$

$$y_j = \eta_j + \mu(\eta_j) + \varepsilon_j, \qquad j = 1, \ldots, 6$$

where $\beta_1 = .5$ and $\mu(\zeta_j) = 0$. The observations are characterized by variances, covariance and means which equal:

$$\sigma^2(\varepsilon_j) = 2, \qquad j = 1, \ldots, 6$$

$$\sigma^2(\zeta_j) = 7.5, \qquad j = 2, \ldots, 6$$

$$\sigma^2(\eta_1) = 10,$$

$$\sigma^2(\eta_j) = \beta_1^2\sigma^2(\eta_{j-1}) + \sigma^2(\zeta_j) = 10, \qquad j = 2, \ldots, 6$$

$$\text{cov}\,(\eta_j\eta_{j-1}) = \beta_1\sigma^2(\eta_{j-1}) = 5, \qquad j = 2, \ldots, 6$$

$$\mu(\eta_j) = 10, \qquad j = 1, \ldots, 6$$

$$\sigma^2(y_j) = \sigma^2(\eta_j) + \sigma^2(\varepsilon_j) = 12. \qquad j = 1, \ldots, 6$$

In the second group, we have the same model from occasions 1 to 3, but hereafter the autoregressive parameter, β_2, equals .70, the mean equals 14, the residual variance equals 5.1, and the error variance equals 4:

$$\sigma^2(\varepsilon_j) = 4, \qquad j = 4, \ldots, 6$$

$$\sigma^2(\zeta_j) = 5.1, \qquad j = 4, \ldots, 6$$

$$\sigma^2(\eta_j) = \beta_2^2\sigma^2(\eta_{j-1}) + \sigma^2(\zeta_j) = 10, \qquad j = 4, \ldots, 6$$

$$\text{cov}\,(\eta_j\eta_{j-1}) = \beta_2\sigma^2(\eta_{j-1}) = 7, \qquad j = 4, \ldots, 6$$

$$\mu(\eta_j) = 14, \qquad j = 4, \ldots, 6$$

$$\sigma^2(y_j) = \sigma^2(\eta_j) + \sigma^2(\varepsilon_j) = 14. \qquad j = 1, \ldots, 6$$

The density is now:

$$f(\mathbf{y}_i; \mathbf{p}, \boldsymbol{\Sigma}\{\tau\}, \boldsymbol{\mu}\{\boldsymbol{\tau}\}) = p_1g_1(\mathbf{y}_i; \{(\mathbf{I} - \mathbf{B}_1^{-1})\boldsymbol{\Psi}_1(\mathbf{I} - \mathbf{B}_1^{-1})^T + \boldsymbol{\Theta}_1\}, \boldsymbol{\nu}_1)$$
$$+ p_2g_2(\mathbf{y}_i; \{(\mathbf{I} - \mathbf{B}_2^{-1})\boldsymbol{\Psi}_2(\mathbf{I} - \mathbf{B}_2^{-1})^T + \boldsymbol{\Theta}_2\}, \boldsymbol{\nu}_2)$$

where diag $[\boldsymbol{\Psi}_1]$ = [10, 7.5, 7.5, 7.5, 7.5, 7.5], diag $[\boldsymbol{\Psi}_2]$ = [10, 7.5, 7.5, 5.1, 5.1, 5.1], $\boldsymbol{\Theta}_1$ = 2I, and diag $[\boldsymbol{\Theta}_2]$ = [2, 2, 2, 4, 4, 4]. The first lower subdiagonal[4] of the otherwise zero matrix $\mathbf{B}_1$ equals [.5, .5, .5, .5, .5] and the first lower sub-diagonal of $\mathbf{B}_2$ equals [.5, .5, .7, .7, .7]. Finally $\boldsymbol{\nu}_1^T$ = [10, 10, 10, 10, 10, 10] and $\boldsymbol{\nu}_2^T$ = [10, 10, 10, 14, 14, 14].

For each component, 150 cases were created, a total of 300 cases ($p_1 = p_2 = .5$). By fitting the unconstrained mixture (55 parameters), we obtain a log likelihood of −4724.6. The estimate of the proportion equals .44. Fitting the constrained true model, we find a loglikelihood of −4749.9. The goodness of fit for the constrained model is $\chi^2(45) = 50.6$ ($p = .36$). The mixing proportion equals .59. Table 4 contains the parameter estimates obtained by fitting the constrained mixture and estimates obtained using Yung's two stage procedure. The results of these two analyses are quite similar both in terms of χ^2 (50.6 vs. 58.9) and of the parameter estimates. The separation of the components equals 1.98. Recalculating this measure for the last 3 occasions, we find a value of 2.84. The *CI*'s based on the standard error and the loglikelihood are quite similar. In the cases of $\sigma^2(\zeta_j)$ (true value: 5.1) and $\sigma^2(\varepsilon_j)$ (2), the *CI*'s diverge somewhat. In the latter case, the loglikelihood *CI* hits a lower bound. As in Illustration 1, the relatively poor separation between the components renders the *CI*'s unreliable, regardless of how they are calculated.

This illustration is indicative of the type of analysis that we would like to carry out in studying qualitative development using the sort of data that featured in the previous illustration. Given J measurement occasions, the model could be extended to a maximum of $J + 1$ components. The assumption that the process before and after the stage transition is stable, gives rise to a highly constrained model that renders the specification of $J + 1$ components feasible.

Confidence Intervals: A Simulation Study

In the illustrations, we generally observe quite good agreement between the likelihood-based *CI*'s and the Hessian-based *CI*'s. A simulation study was carried out to arrive at a more systematic comparison of the *CI*'s and their accuracy. We simulated a 5-variate normal mixture consisting of two components of equal size ($p_1 = .5$). Two factors were varied: sample size with three levels ($N = 100, N = 200, N = 400$), and separation of the components with two levels ($d_{12} = 3.69$ and $d_{12} = 2.076$). Within each condition, 250 replications were carried out (a total of 1500 data sets). The 5 variables were simulated according to a single common factor model that was identical in the two components. The factor loadings equaled $\boldsymbol{\Lambda}^T$ = [1 1.3 .9 1.2 .8] and the error variances equaled diag $(\boldsymbol{\Theta})$ = [1 1.3 .7 1.2 .6]. The variance of the common factor equaled 1 ($\boldsymbol{\Psi}$ = [1]). The variances of the indicators attributable to the common factor ranged between 50% and 56%. The means in the first component equaled $\boldsymbol{\nu}_1^T$ = [2 3 4 3 2]. In the $d_{12} = 3.69$ condition, the

[4] That is, the elements in row i and column $i - 1$, where $i = 2, \ldots, 6$.

Table 4: ML parameter estimates, standard errors in parentheses, 95% CI's (lower/upper endpoint) for Illustration 3.

Component 1	ML (mixture)[1]	2 stage GLS[2]
$\sigma^2(\varepsilon_j) = 2$ (j=1,6)	1.11 (1.09)	1.04
CI (s.e)	-1.04/3.27	
CI (logL)	.01#/3.07	
$\sigma^2(\zeta_j) = 7.5$ (j=2,6)	8.79 (1.40)	9.08
CI (s.e)	6.05/11.53	
CI (logL)	6.31/10.87	
$\sigma^2(\eta_1) = 10$	11.73 (1.42)	11.80
CI (s.e)	8.95/14.52	
CI (logL)	9.14/14.67	
$\mu(\eta_j) = 10$ (j=1,6)	10.06 (.15)	9.96
CI (s.e)	9.78/10.35	
CI (logL)	9.77/10.35	
$\beta_{j,j-1} = .5$ (j=2,6)	.48 (.05)	.43
CI (s.e)	.38/.58	
CI (logL)	.38/.58	
Component 2		
$\sigma^2(\varepsilon_j) = 4$ (j=4,6)	5.32	3.68
CI (s.e)	3.04/7.62	
CI (logL)	2.88/7.75	
$\sigma^2(\zeta_j) = 5.1$ (j=4,6)	3.09 (1.44)	5.18
CI (s.e)	.26/5.92	
CI (logL)	.81/6.75	
$\mu(\eta_j) = 14$ (j=4,6)	14.57 (.47)	13.52
CI (s.e)	13.64/15.48	
CI (logL)	13.60/15.43	
$\beta_{j,j-1} = .7$ (j=4,6)	.74 (.06)	.73
CI (s.e)	.61/.87	
CI (logL)	.60/.86	
proportion (.5)	.59 (.06)	-
CI (s.e)	.47/.70	
CI (logL)	.47/.72	
d_{12}(j=1,6)	1.98	
d_{12}(j=4,6)	2.84	
logL	-4749.9	-4754.1
χ^2(df)	50.6(45)	58.9(45)

Note

1 Parameter estimates obtained by fitting the mixture. 2 Parameters estimated using Yung's (1994) two stage GLS. Unstructured means and covariance matrix for 2 stage GLS were calculated using EM algorithm. Subsequent parameter estimation based carried out by minimizing the loglikelihood ratio. #Bound on parameter was hit during optimization.

means in the second component equal $\nu_2^T = [6\ 8.2\ 7.6\ 7.8\ 5.2]$, and in the $d_{12} = 2.067$ condition, $\nu_2^T = [4.25\ 5.92\ 6.02\ 5.7\ 3.8]$.

Data simulation and analysis were carried out using FORTRAN programs on a Pentium 75 personal computer. For each replication, we first fitted the unconstrained mixture using the EM algorithm to obtain a baseline log likelihood. The log likelihood for the constrained mixture was maximized using E04VDF with the parameters in Λ and Θ constrained to be equal over the components. The variance of the common factor was standardized in both components. A χ^2 goodness-of-fit index for the factor model was calculated as minus twice the difference between the log likelihood of the unconstrained model and that of the constrained model. The number of parameters estimated equaled 41 in the unconstrained analysis, and 21 in the constrained analysis. There are 20 (41 − 21) degrees of freedom for the common factor model. Following each analysis, checks were carried out to ascertain that no parameter had hit a bound and that the program had converged properly. Subsequently the Hessian-based standard errors were calculated. Finally, the function in Eqs. 11 and 12 were minimized using E04VDF to obtain the log-likelihood based *CI*'s. A check was carried out to ascertain that a minimum had been reached within the permitted number of iterations. Whether a bound was hit during the calculation of the log-likelihood *CI*'s, was not considered. This implies that some *CI*'s were found to equal the stated parameter bounds. The bounds were set very wide. For the components of Λ, $\boldsymbol{\nu}_1$ and $\boldsymbol{\nu}_2$, the bounds were set to equal ±200. Bounds on the diagonals of $\boldsymbol{\Theta}$ equaled .00001 and 200. The bounds on the mixing proportion, p_1, equaled .001 and .999. We limit our attention to the following parameters: the second factor loading, the second error variance, the first mean in the first component, and mixing proportion. These are denoted λ_2 (true value 1.3), $\sigma^2\varepsilon_2$ (1.3), ν_{11} (2), and p_1 (.5), respectively.

Starting values were set to equal the true values in all analyses. In the cells $N = 200$ & $d_{12} = 2.076$ and $N = 100$ & $d_{12} = 2.076$, 3.9% and 7.6%, respectively, of the analyses failed due to divergence during maximization of the log-likelihood for the constrained model. Failed analyses were repeated with new data to make up the total of 250 replications.

The mean and standard deviation of χ^2 goodness of fit indices are reported in Table 5 and 6 for each cell in the 2 × 3 design. In the $d_{12} = 3.69$ conditions, the χ^2 follows its expected distribution when $N = 200$, or $N = 400$. When $N = 100$, the χ^2 does not approach its expected distribution very well and the factor model is rejected at a significance level of 0.05, in 14% (35/250) of the replications ($\chi^2(20) > 31.41$; expected number of rejections 12.5). In the $d_{12} = 2.076$ conditions, the χ^2 statistic cannot be trusted regardless of the considered sample sizes, although the number of rejections increase as the sample size is smaller. The results relating to the χ^2 statistic agree with those reported by Yung (1994; Table 14).

We first discuss the results relating to the parameter estimates in the $d_{12} = 3.69$ cells. The means of the estimates are generally close to their true values. In the $N = 100$ cell the mean of the estimate of the error variance is underestimated (1.228 vs. 1.3). We observe a good agreement between the standard deviation of the estimates of the factor model and the associated mean Hessian-based standard errors. The mean *CI*'s are likewise very close in value. Besides the likelihood-based and Hessian-based intervals, we calculate so-called empirical *CI*'s based on the mean and standard deviation of the parameter estimates (see footnote Table 5). The mean standard errors of the proportions are systematically larger than the standard deviation of the estimate (.011 vs. .027 [$N = 400$], .015 vs. .039 [$N = 200$], and .028 vs. .055 [$N = 100$]). This finding is surprising, because Yung (1994, Table 15) finds that the mean standard errors are consistently smaller than the standard deviation of the estimate of the proportion, even in conditions characterized by separation better than our 3.69. The likelihood-based *CI*'s and Hessian-based *CI*'s are quite similar, but are slightly wider than the empirical *CI*'s.

In the $d_{12} = 2.076$ condition, we find that the true variability of the estimates is systematically underestimated by the Hessian-based standard errors. This is now also the case with the variability of the estimates of the proportion. The smaller the sample size, the greater the underestimation. In the $N = 100$ cell, the mean standard errors are too small by about a factor 1.5. The upper endpoints of the empirical and likelihood-based *CI*'s of the parameters of the factor model agree quite well, even in the $N = 100$ cell. The lower endpoints agree well only in the $N = 400$ conditions. Except for results relating to the factor loading (λ_2), the likelihood-based lower endpoints of the *CI*'s are closer in value to the empirical lower endpoints than the Hessian-based endpoints. The lower and upper endpoints of the likelihood-based *CI*'s of the proportion in the $N = 200$ and $N = 100$

Table 5: Summary statistics calculated in the d_{12}=3.69 conditions. 250 replications within each condition.

Condition N=400	($\mu(\chi^2)$=20.23[3],	$\sigma(\chi^2)$=6.41,	rejected=13[2],	5.2%)
parameter	λ_2=1.3	ν_{11}=2	$\sigma^2\varepsilon_2$=1.3	p_1=.5
μ(est.)	1.294	1.982	1.299	.500
σ(est.)	.091	.113	.119	.011
μ(st.err.)	.094	.111	.120	.027
σ(st.err.)	.007	.007	.009	.001
upper CI(emp.)[1]	1.472	2.203	1.532	.521
upper μ(loglCI)	1.491/1.01[4]	2.202/1.00	1.556/1.02	.553/1.06
upper μ(hessCI)	1.489/1.01	2.209/1.00	1.544/1.01	.563/1.08
lower CI(emp.)	1.115	1.760	1.065	.478
lower μ(loglCI)	1.118/1.00	1.765/1.00	1.080/1.01	.448/.94
lower μ(hessCI)	1.119/1.00	1.775/1.00	1.074/1.01	.457/.96
Condition N=200	($\mu(\chi^2)$=21.28,	$\sigma(\chi^2)$=7.05,	rejected=20,	8.0%)
μ(est.)	1.316	1.939	1.279	.499
σ(est.)	.146	.145	.172	.015
μ(st.err.)	.135	.158	.170	.039
σ(st.err.)	.020	.015	.019	.002
upper CI(emp.)	1.602	2.214	1.616	.528
upper μ(loglCI)	1.606/1.00	2.256/1.02	1.657/1.03	.574/1.09
upper μ(hessCI)	1.590/.99	2.259/1.02	1.622/1.00	.584/1.11
lower CI (emp.)	1.029	1.645	.942	.469
lower μ(loglCI)	1.068/1.04	1.628/.99	.975/1.04	.424/.90
lower μ(hessCI)	1.062/1.03	1.640/1.00	.955/1.01	.433/.92
Condition N=100	($\mu(\chi^2)$=23.38,	$\sigma(\chi^2)$=8.37,	rejected=35,	14%)
μ(est)	1.309	1.981	1.228	.501
σ(est.)	.222	.210	.259	.028
μ(st.err.)	.184	.223	.234	.055
σ(st.err.)	.028	.032	.034	.005
upper CI(emp.)	1.744	2.393	1.735	.556
upper μ(loglCI)	1.740/1.00	2.447/1.02	1.780/1.03	.609/1.10
upper μ(hessCI)	1.679/.96	2.428/1.01	1.697/.98	.618/1.11
lower CI(emp.)	.874	1.569	.720	.446
lower μ(loglCI)	.974/1.11	1.533/.98	.821/1.14	.394/.88
lower μ(hessCI)	.960/1.10	1.553/.99	.778/1.08	.404/.91

[1] "empirical" CI's are calculated as μ(est.) ± σ(est.)*1.96.

[2] rejected means that the $\chi^2(20)$ for the common factor model exceeded 31.41 (α=0.05).

[3] expected mean and standard deviation equal df and $\sqrt{(2df)}$, i.e. 20 and 6.324.

[4] ratio of μ(loglCI) to CI(emp.).

conditions hit the lower (.001) and upper bound (.999) in a number of cases (see footnote Table 6). The same applied to the lower endpoint of the likelihood-based *CI* of the error variance in the N = 100 condition.

Table 6[1]: Summary statistics relating to 4 parameters calculated in the d_{12}=2.076 conditions. 250 replications within each condition.

Condition N=400	($\mu(\chi^2)$=26.04,	$\sigma(\chi^2)$=9.56,	rejected 59,	23.6%)
parameter	λ_2=1.3	v_1=2	$\sigma^2\varepsilon_2$=1.3	p_1=.5
μ(est.)	1.328	1.999	1.236	.496
σ(est.)	.183	.225	.152	.069
μ(st.err.)	.145	.174	.132	.052
σ(st.err.)	.054	.048	.025	.014
upper CI(emp.)	1.687	2.440	1.534	.631
upper μ(loglCI)	1.689/1.00	2.485/1.02	1.514/.99	.613/.97
upper μ(hessCI)	1.622/.96	2.351/.96	1.505/.98	.607/.96
lower CI(emp.)	.969	1.558	.938	.361
lower μ(loglCI)	1.068/1.10	1.626/1.02	.955/1.02	.369/1.02
lower μ(hessCI)	1.053/1.09	1.667/.96	.988/1.05	.405/1.12
Condition N=200	($\mu(\chi^2)$=29.13,	$\sigma(\chi^2)$=9.36,	rejected=94,	37.6%)
μ(est.)	1.365	2.040	1.225	.517
σ(est.)	.254	.377	.203	.100
μ(st.err.)	.178	.214	.186	.064
σ(st.err.)	.047	.056	.034	.017
upper CI(emp.)	1.863	2.778	1.623	.713
upper μ(loglCI)	1.867/1.00	2.929/1.05	1.642/1.01	.693[4]/.97
upper μ(hessCI)	1.724/.93	2.470/.89	1.600/.99	.651/.91
lower CI(emp.)	.867	1.301	.827	.321
lower μ(loglCI)	1.030/1.19	1.448/1.11	.820/.99	.328[3]/1.02
lower μ(hessCI)	1.026/1.18	1.630/1.25	.871/1.05	.402/1.25
Condition N=100	($\mu(\chi^2)$=28.37,	$\sigma(\chi^2)$=9.54,	rejected=92,	37%)
μ(est.)	1.290	1.957	1.139	.505
σ(est.)	.305	.406	.302	.123
μ(st.err.)	.203	.263	.235	.071
σ(st.err.)	.041	.064	.049	.014
upper CI(emp.)	1.889	2.753	1.731	.746
upper μ(loglCI)	1.888/1.00	2.912/1.05	1.725/1.00	.690[6]/.92
upper μ(hessCI)	1.698/.90	2.481/.89	1.610/.93	.655/.88
lower CI(emp.)	.692	1.161	.547	.264
lower μ(loglCI)	.908/1.31	1.226/1.11	.670[2]/1.22	.307[5]/1.16
lower μ(hessCI)	.901/1.30	1.452/1.25	.689/1.26	.374/1.42

[1] see footnotes Table 5

[2] In 10/250% of the replications, the endpoint hit the bound (.00001).

[3] In 27/250% of the replications, the endpoint hit the bound (.001).

[4] In 25/250% of the replications, the endpoint hit the bound (.999).

[5] In 22/250% of the replications, the endpoint hit the bound (.001).

[6] In 23/250% of the replications, the endpoint hit the bound (.001).

Table 7, finally, contains the correlation (Spearman's ρ) between the standard errors and the observed measure of separation calculated within each cell of the design. A

Table 7: Spearman's ρ between observed distance measures, d_{12}, and standard errors, and χ^2 within each cell.

		standard error of				
cell		λ_2	ν_{11}	$\sigma^2\varepsilon_2$	p_1	χ^2
d_{12}=3.69	N=400	-.54	-.72	-.06	-.94	.00
d_{12}=3.69	N=200	-.53	-.71	.00	-.91	-.08
d_{12}=3.69	N=100	-.59	-.65	.00	-.88	-.15
d_{12}=2.076	N=400	-.48	-.72	-.11	-.88	-.25
d_{12}=2.076	N=200	-.37	-.60	-.16	-.90	-.27
d_{12}=2.076	N=100	-.38	-.50	.02	-.83	-.28

striking result, that Yung (1994, p. 49) also reports, is that the standard error of the error variance is hardly affected by the separation between the components. Concerning the other parameters, it is clear that the standard error of the proportion is highly dependent on the degree of separation. This finding is consistent with the rather large discrepancies between the standard deviations of the estimates and their mean standard errors. The standard errors of the factor loading and the mean are also quite sensitive to the degree of separation, although the latter is more sensitive than the former. Table 7 also contains the correlations between the χ^2's and the degree of separation. Within the d_{12} = 3.69 cells, the correlations are small, but there is a clear relationship with sample size. The greater the sample size, the less the χ^2 is influenced by the degree of separation. The correlation ranges from .0 (N = 400) to −.15 (N = 100), In the d_{12} = 2.067 cells, the correlations are about −.27, and much less affected by sample size.

The results of the present simulation study indicate that the loglikelihood-based *CI*'s and Hessian-based intervals are comparable and quite accurate when the separation is good (d_{12} = 3.69). When the separation of the components is poorer (d_{12} = 2.076), the likelihood-based *CI*'s are generally slightly more accurate. The *CI*'s of the factor loading are an exception. Here the likelihood-based and Hessian-based *CI*'s are similar and differ equally from the empirical *CI*'s.

Most results are compatible with those presented by Yung (1994). An important difference relates to the variability of the estimates of the proportions in the d_{12} = 3.69 condition. As mentioned, we find that the standard deviations of the estimates are smaller than the means of the standard error. Yung observed the exact reverse. We suspect that this difference is due to the difference in the model that featured in the simulations. Yung (1994) based his simulations on an oblique two common factor model, where the differences between the components were due to latent factor means and latent factor (co-)variances (see Sörbom, 1974). The present model is a lot simpler, and, especially within the context of normal mixtures, a lot easier to fit. It is our experience that constrained normal mixtures are difficult to fit when the mixture is specified at the level of the latent variables, as is the case in Yung's (1994) simulation.

Discussion

We have found that fitting multivariate normal mixtures subject to SEM does not pose any additional computational problem to those that are generally recognized. Rather, the introduction of constraints reduces the number of parameters and so alleviates the computational load. Although individual bounds may keep parameter estimates within the admissible parameter space, such constraints do not necessarily guarantee that a covariance matrix will remain positive definite during optimization. In optimizing the loglikelihood function using the quasi-Newton method, we used a penalty to ensure that covariance matrices remained positive definite. In maximizing the log-likelihood function using the EM algorithm in contrast, we found that a covariance matrix that was positive definite thanks to sensible starting values, usually stayed so during optimization. Regardless of the number, or type of constraints, and indeed of the method of estimation, the danger of local maxima in fitting normal mixtures is ever present and the only way to gain confidence in a solution is to vary starting values. We have often found it necessary to go through the, at times, tedious process of finding suitable initial values and checking solutions.

A number of possibilities in estimating parameters remain to be considered. One possibility is to implement Hamilton's quasi-Bayesian approach to parameter estimation in normal mixtures (Hamilton, 1991). Judging by the results of Hamilton's Monte Carlo study, this appears to be a useful option. A second possibility is the use of a genetic algorithm to optimize the loglikelihood (Goldberg, 1989). Genetic algorithms are computationally intensive, but are quite insensitive to local maxima and to the choice of initial values, i.e., vexing problems in fitting multivariate normal mixtures. Van der Maas and Raijmakers (1997) have demonstrated the effectiveness of this method of optimization in exploratory latent class analysis.

We have found the quasi-Newton based approach to optimization using exact gradients to be feasible. The combined use of the EM algorithm and the Quasi-Newton algorithm in Yung's two stage GLS procedure (Yung, 1994) is helpful in finding good starting values. Solutions still have to be checked, however, because the EM algorithm may converge to a local maximum.

We have focused solely on mixtures of multivariate normals. The assessment of multivariate normality is problematic, because the histograms of a mixture of univariate normals may assume a variety of shapes (Everitt & Hand, 1981, p. 27 ff.; Titterington, et al., 1985, p. 49). To assess normality, one may first to fit the mixture and obtain assignments of the cases to the components based on the posterior probability, (9). Given these, the assumption of normality can be assessed using goodness-of-fit procedures.

We have further limited our attention to fitting mixtures in the absence of any information concerning the component membership of the cases. Titterington, et al. (1985, p. 3) discusses fitting mixtures given an addition sample of fully categorized cases (e.g., cases whose component membership is known a priori). Yung (1994, 1997) presents versions of his estimation procedures that can accommodate such complex sampling schemes. The ability to include categorized data in fitting a mixture is useful, because the addition of such information is known to be very beneficial from a computational point of view (Hamilton, 1991; Titterington, et al., 1985; Yung, 1994). The quasi-Newton approach to estimation, that we have adopted here, has the advantage that it is easy to generalize to include multigroups (e.g., males and females). Because the loglikelihood function and gradients need only be weighed by the relative group size prior to summation over the groups, additional programming requirement is quite small. Fitting multigroup mixtures, where the data in each group may or may not be a mixture, provides an alternative to Arminger and Stein's approach of incorporating group membership as a fixed observed covariate.

On the basis of the results of our simulation study, we have found that likelihood-based *CI*'s are only slightly more accurate than the Hessian-based *CI*'s. Like Yung (1994), we find that asymptotic results relating to standard errors, *CI*'s, and the likelihood ratio cannot be trusted when the separation between the components is insufficient. It is therefore advisable to consider the separation between the components, before setting too great a store by such results.

Appendix 1: Derivatives of the Log-Likelihood Function

This appendix contains the derivatives of the loglikelihood function in (10). The multivariate normal mixture density is given in (7). Rather than ensuring that the mixing proportions sum to unity by introducing a Lagrange multiplier, we specify the bounds $l_k \le p_k \le u_k$, where $0 < l_k < u_k < 1$ $(k = 1, \ldots, R-1)$, and concentrate the loglikelihood function by solving for p_R: $p_R = 1 - p_1 - p_2 - \ldots - p_{R-1}$. Depending in the value of R and the values of the bounds, l_k and u_k, it is possible that p_R assumes a negative value. As a precaution, we check whether $p_R > 0$, and introduce a penalty if this is not so (a rare occurrence). Maximum likelihood estimates of $\mathbf{p}$ and $\boldsymbol{\tau}$ are obtained by maximizing the loglikelihood function, $L(\mathbf{p}, \boldsymbol{\tau}; \mathbf{Y}, R)$, (10). We first require the derivatives of $L(\mathbf{p}, \boldsymbol{\tau}; \mathbf{Y}, R)$ with respect to $\boldsymbol{\mu}_k$, $\mathbf{p}$ and $\boldsymbol{\Sigma}_k$. To ease presentation, let f and g_k represent $f(\mathbf{y}_i, \mathbf{p}, \boldsymbol{\Sigma}\{\boldsymbol{\tau}\}, \boldsymbol{\mu}\{\boldsymbol{\tau}\})$ and $g_k(\mathbf{y}_i, \boldsymbol{\Sigma}_k\{\boldsymbol{\tau}_k\}, \boldsymbol{\mu}_k\{\boldsymbol{\tau}_k\})$, respectively, and let L stand for $L(\mathbf{p}, \boldsymbol{\tau}; \mathbf{Y}, R)$.

Bearing in mind that p_R depends on the other components of the vector $\mathbf{p}$, the derivative with respect to the components of $\mathbf{p}$ is (e.g., Everitt & Hand, 1981, Eq. 2.16):

$$\partial L/\partial p_k = \sum_{i=1}^{N} f^{-1}[g_k - g_R]. \qquad (k = 1, \ldots, R-1) \tag{1A}$$

The derivative with respect to $\boldsymbol{\mu}_k$ is (e.g., Everitt & Hand, 1981, Eq. 2.17):

$$\partial L/\partial \boldsymbol{\mu}_k = \sum_{i=1}^{N} \omega_{ki}[\boldsymbol{\Sigma}_k^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_k)], \qquad (k = 1, \ldots, R) \tag{2A}$$

where ω_{ki} is posterior probability defined as $p_k g_k f^{-1}$. Everitt and Hand (1981, Eq. 2.8) provide the derivative with respect $\boldsymbol{\Sigma}_k^{-1}$. We require the derivative with respect to $\boldsymbol{\Sigma}_k$. Using results published in Graybill (1983, p. 356–359), we have:

$$\partial L/\partial \boldsymbol{\Sigma}_k = \sum_{i=1}^{N} \frac{1}{2}\, \omega_{ki}\{[\boldsymbol{\Sigma}_k^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_k)(\mathbf{y}_i - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}] - \boldsymbol{\Sigma}_k^{-1}\}. \qquad (k = 1, \ldots, R) \tag{3A}$$

Once we have calculated $[\partial L/\partial \boldsymbol{\Sigma}_k]$ and $[\partial L/\partial \boldsymbol{\mu}_k]$, we use the chain rule to obtain the derivatives of the loglikelihood with respect to the model matrices $\boldsymbol{\Lambda}_k$, $\mathbf{B}_k$, $\boldsymbol{\Psi}_k$, and $\boldsymbol{\Theta}_k$, and vectors $\boldsymbol{\nu}_k$ and $\boldsymbol{\alpha}_k$ (Equations (4A) to (7A) can be found in Jöreskog, 1977):

$$\partial L/\partial \boldsymbol{\Sigma}_k \partial \boldsymbol{\Sigma}_k/\partial \boldsymbol{\Lambda}_k = 2[\partial L/\partial \boldsymbol{\Sigma}_k]\boldsymbol{\Lambda}_k(\mathbf{I} - \mathbf{B}_k)^{-1}\boldsymbol{\Psi}_k(\mathbf{I} - \mathbf{B}_k)^{-1T} \tag{4A}$$

$$\partial L/\partial \boldsymbol{\Sigma}_k \partial \boldsymbol{\Sigma}_k/\partial(\mathbf{I} - \mathbf{B}_k) = -2(\mathbf{I} - \mathbf{B}_k)^{-1T}\boldsymbol{\Lambda}_k^T[\partial L/\partial \boldsymbol{\Sigma}_k]\boldsymbol{\Lambda}_k(\mathbf{I} - \mathbf{B}_k)^{-1}\boldsymbol{\Psi}_k(\mathbf{I} - \mathbf{B}_k)^{-1T} \tag{5A}$$

$$\partial L/\partial \boldsymbol{\Sigma}_k \partial \boldsymbol{\Sigma}_k/\partial \boldsymbol{\Psi}_k = (\mathbf{I} - \mathbf{B}_k)^{-1T}\boldsymbol{\Lambda}_k^T[\partial L/\partial \boldsymbol{\Sigma}_k]\boldsymbol{\Lambda}_k(\mathbf{I} - \mathbf{B}_k)^{-1} \tag{6A}$$

$$\partial L/\partial \boldsymbol{\Sigma}_k \partial \boldsymbol{\Sigma}_k/\partial \boldsymbol{\Theta}_k = [\partial L/\partial \boldsymbol{\Sigma}_k] \tag{7A}$$

$$\partial L/\partial \boldsymbol{\mu}_k \partial \boldsymbol{\mu}_k/\partial \boldsymbol{\Lambda}_k = [\partial L/\partial \boldsymbol{\mu}_k](\mathbf{I} - \mathbf{B}_k)^{-1}\boldsymbol{\alpha}_k \tag{8A}$$

$$\partial L/\partial \boldsymbol{\mu}_k \partial \boldsymbol{\mu}_k/\partial(\mathbf{I} - \mathbf{B}_k) = -(\mathbf{I} - \mathbf{B}_k)^{-1T}\boldsymbol{\Lambda}_k^T[\partial L/\partial \boldsymbol{\mu}_k]\boldsymbol{\alpha}_k^T \tag{9A}$$

$$\partial L/\partial \boldsymbol{\mu}_k \partial \boldsymbol{\mu}_k/\partial \boldsymbol{\alpha}_k = (\mathbf{I} - \mathbf{B}_k)^{-1T}\Lambda_k^T[\partial L/\delta \boldsymbol{\mu}_k] \tag{10A}$$

$$\partial L/\partial \boldsymbol{\mu}_k \partial \boldsymbol{\mu}_k/\partial \nu_k = [\partial L/\partial \boldsymbol{\mu}_k] \tag{11A}$$

Finally, letting $\mathbf{M}_k \in \{\boldsymbol{\Lambda}_k, (\mathbf{I} - \mathbf{B}_k), \boldsymbol{\Psi}_k, \boldsymbol{\Theta}_k, \boldsymbol{\nu}_k, \boldsymbol{\alpha}_k\}$, we have:

$$\partial L/\partial \mathbf{M}_k = \partial L/\partial \boldsymbol{\Sigma}_k \partial \boldsymbol{\Sigma}_k/\partial \mathbf{M}_k + \partial L/\partial \boldsymbol{\mu}_k \partial \boldsymbol{\mu}_k/\partial \mathbf{M}_k \qquad (k = 1, \ldots, R).$$

References

Agha, M. & Branker, D. S. (1997). Maximum Likelihood estimations and goodness of fit tests for mixtures of distributions (AS 317). *Applied Statistics, 46,* 399–407.

Aitchison, J., & Silvey, S. D. (1958). Maximum-likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics, 29,* 813–828.

Arminger, G., & Stein, P. (1997). Finite mixtures of covariance structure models with regressions: likelihood, function, minimum distance estimation, fit indices, and a complex example. *Submitted.*

Azzalini, A. (1996). *Statistical inference based on the likelihood.* London: Chapman and Hall.

Blåfield, E. (1980). *Clustering of observations from finite mixtures with structural information* (Jyväskylä studies in computer science, economics and statistics, No. 2). Jyväskylä, Finland: Jyväskylä University.

Dolan, C. V., & Molenaar, P. C. M. (1991). A comparison of 4 methods of calculating standard errors of maximum likelihood estimates in the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 44,* 359–368.

Everitt, B. S., & Hand, D. J. (1981). *Finite mixture distributions.* London: Chapman and Hall.

Farmacotherapeutisch Kompas (1994). *Farmacotherapeutisch kompas: medisch farmaceutisch voorlichting* [Pharmacotherapeutic guide: medical pharmaceutical information]. Amstelveen: Ziekenfondsraad.

Feng, Z. D. & McCulloch, C. E. (1996). Using bootstrap likelihood ratios in finite mixture models. *Journal of the Royal Statistical Society, Series B, 58,* 609–617.

Gill, P. E., Murray, W., & Wright, M. H. (1981). *Practical Optimization.* London: Academic Press.

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization and machine learning.* Reading: Addison-Wesley.

Graybill, F. A. (1983). *Matrices with applications in statistics* (2nd ed.). Belmont, CA: Wadsworth.

Hamilton, J. D. (1990). Analysis of time series subject to changes in regime. *Journal of Econometrics, 45,* 39–70.

Hamilton, J. D. (1991). A quasi-Bayesian approach to estimating parameters for mixtures of normal distributions. *Journal of Business and Economic Statistics, 9,* 27–39.

Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics, 13,* 795–800.

Hosmer, D. W. (1974). Maximum Likelihood estimates of parameters of a mixture of two regression lines. *Communication in Statistics. Theory and Methods, 3,* 995–1006.

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

Jedidi, K., Jagpal, H. S., & DeSarbo, W. S. (1997b). STEMM: A general finite mixture structural equation model. *Journal of Classification, 14,* 23–50.

Jöreskog, K. G. (1970). Estimation and fitting of simplex models. *British Journal of Mathematical and Statistical Psychology, 23,* 121–145.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 57,* 409–426.

Jöreskog, K. G. (1977). Structural equation models in the social sciences: Specification, estimation and testing. In P. R. Krishnaiah (Ed.), *Applications of Statistics.* Amsterdam: North-Holland.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8 user's reference guide.* Chicago: Scientific Software International.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics, 36,* 318–324.

Murphy, A. E., & Bolling, D. R. (1967). Testing of a single locus hypothesis where there is incomplete separation of the phenotypes. *American Journal of Human Genetics, 19,* 322–334.

Numerical Algorithms Group. (1990). *The NAG Fortran Library Manual, Mark 14.* Oxford: Author.

Neale, M. C. (1995). *Mx: Statistical Modeling* (3rd ed.). Richmond, VA: Medical College of Virginia.

Neale, M. C., & Miller, M. (1997). The use of likelihood-based confidence intervals in genetic models. *Behavior Genetics, 27,* 113–120.

Piaget, J., & Inhelder, B. (1969). *The psychology of the child.* New York: Basic Books.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structures between groups. *British Journal of Mathematical and Statistical Psychology, 27,* 229–239.

Stein, P. (1997). *Mischungen von konditionalen Mittlewertund Kovarianzstrukturmodellen mit Anwendungen auf die analyse von Lebensstilen*. Unpublished doctoral dissertation, Department of Social Sciences, Gerhard Mercator University, Duisburg, Germany.

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Chicester: John Wiley & Sons.

Van der Maas, H. J. L. (1993). *Catastrophe analysis of stagewise cognitive development: model, method and applications* (Dissertatie reeks 1993–2). Amsterdam: University of Amsterdam, Psychology Faculty.

van der Maas, H. J. L. & Raijmakers, M. E. J. (1997). *Optimizing latent class models by genetic algorithms* (Internal Report). Amsterdam: University of Amsterdam, Developmental Psychology, Psychology Faculty.

Venzon, D. J., & Moolgavkar, S. H. (1988). A method for computing profile-likelihood-based confidence intervals. *Applied Statistics*, *37*, 87–94.

Wedel, M. (1995). GLIMMIX, A program for mixtures of generalized linear regression models, and its applications in marketing. *Kwantitatieve Methoden*, *50*, 55–70.

Wedel, M., & DeSarbo, W. S. (1994). A review of recent developments in latent class regression models. In R. P. Bagozzi (Ed.), *Advanced methods of marketing research* (pp. 352–388) Cambridge, MA:Blackwell.

Wedel, M., & DeSarbo, W. S. (1995). A mixture likelihood approach for generalized linear models. *Journal of Classification*, *12*, 21–55.

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, *5*, 329–350.

Yung, Y. F. (1994). *Finite mixtures in confirmatory factor-analytic models*. Unpublished doctoral dissertation, University of California, Los Angeles. (Also available as Yung, Y. F., 1995, *Finite Mixtures in Confirmatory Factor-Analytic Models* (microfilm). Ann Arbor, MI: University Microfilms.

Yung, Y. F. (1997). Finite Mixtures in Confirmatory Factor-Analysis Models. *Psychometrika*, *62*, 297–330.

Manuscript received 5/24/96
Final version received 10/7/97