FACTOR ANALYSIS FOR CLUSTERED OBSERVATIONS

N. T. LONGFORD

EDUCATIONAL TESTING SERVICE

B. O. MUTHÉN

UNIVERSITY OF CALIFORNIA, LOS ANGELES

Classical factor analysis assumes a random sample of vectors of observations. For clustered vectors of observations, such as data for students from colleges, or individuals within households, it may be necessary to consider different within-group and between-group factor structures. Such a two-level model for factor analysis is defined, and formulas for a scoring algorithm for estimation with this model are derived. A simple noniterative method based on a decomposition of the total sums of squares and crossproducts is discussed. This method provides a suitable starting solution for the iterative algorithm, but it is also a very good approximation to the maximum likelihood solution. Extensions for higher levels of nesting are indicated. With judicious application of quasi-Newton methods, the amount of computation involved in the scoring algorithm is moderate even for complex problems; in particular, no inversion of matrices with large dimensions is involved. The methods are illustrated on two examples.

Key words: factor analysis, Fisher scoring algorithm, maximum likelihood, multilevel analysis.

1. Introduction and Motivation

Classical factor analysis (see, e.g., Lawley & Maxwell, 1971) is a commonly applied multivariate statistical technique. The estimation theory is well-developed for both exploratory and confirmatory modes (see, e.g., Jöreskog, 1977, and Jöreskog & Sörbom, 1979). Factor analysis is, however, frequently applied to observational data for which the standard assumption of independence of the vectors of observations, or that of simple random sampling, is not appropriate. For example, students are usually observed within classrooms and schools, and individuals are observed within households. It is often reasonable to assume that the observations within a group are more similar, because the subjects share common environment, experiences, and interactions. This within-group homogeneity, or between-group variation, can be modeled by a *group-level* correlation structure; at the same time an *individual-level* correlation structure is considered. Such a development runs parallel with the extension of the ordinary regression to random coefficient (mixed) models for clustered observations (see, e.g., Aitkin & Longford, 1986; de Leeuw & Kreft, 1986; Goldstein, 1986, 1987; Jennrich & Schluchter, 1986; Longford, 1987; Mason, Wong, & Entwisle, 1984; Raudenbush & Bryk, 1985), since factor analysis models can be formally regarded as ordinary regression models with unknown regressors.

Recent work related to extensions of factor analysis and of structural equations for correlated vectors of observations is that of Goldstein and McDonald (1988), McDonald

Suggestions and corrections of three anonymous referees and of an Associate Editor are acknowledged. Discussions with Bob Jennrich on computational aspects were very helpful. Most of research leading to this paper was carried out while the first author was a visiting associate professor at the University of California, Los Angeles.

Requests for reprints should be sent to N. T. Longford, 21-T, Educational Testing Service, Rosedale Rd., Princeton, NJ 08541.

0033-3123/92/1200-90280$00.75/0

#41

and Goldstein (1989), Lee (1990), Muthén and Satorra (1989), and Muthén (1989, 1992). The former three papers outline maximum likelihood estimation in a general two-level structural model with the aim of developing specially designed software to carry out the complex computational tasks. Muthén and Satorra (1989) and Muthén (1989) discuss model specification and estimation for two-level structural equation models with balanced data, and Muthén (in press) describes an implementation in structural equation modeling software for the unbalanced case.

The present paper focuses on efficient computation for maximum likelihood estimation in the factor analysis model. In particular, the Fisher scoring algorithm for a two-level analysis is described. The algorithm relies on formulas similar to those derived by McDonald and Goldstein (1989) for a more general class of models with balanced design, whereas here a general unbalanced design is assumed. The framework of models is described in section 2, and section 3 presents the scoring algorithm. In section 4 a noniterative method for fitting the "unrestricted" model that imposes no constraints on the factor structure at either level is discussed. This method, in conjunction with the Fisher-scoring method, provides a computationally economic way of calculating the likelihood ratio χ^2 test statistic for a model fit.

The methods presented in this paper are easy to implement using any software with standard matrix algebra tools. They do not require balanced data and in principle they can be extended to data structures with further layers of nesting, such as subjects within groups within areas.

The focus of a substantive analysis may be on the within-group, the between-group factor structure, on the structure at both levels, or on the comparison of the factor structures. Although McDonald and Goldstein (1989) suggest that the within- and between-group factor structures may be equal, or parallel, counterexamples are bound to arise with more frequent application of these methods. For example, in studies of academic performance of students (with students within schools), the within-school factor structure may be interpreted as a description for natural variation in the student population, while the between-school factor structure would describe a composite of the school "effect" and of the selection procedures that lead to a specific set of students attending the school. Obviously, the latter processes may be unrelated to the within-school factor structure.

Longford (1990) reported an analysis of the covariance structures underlying the subscores of an educational test. The test was administered to students, but the assessment was focused on colleges, and so the aggregate (within-college average) subscores corresponding to certain academic skills were of principal interest. The purpose of the analysis was to establish usefulness of these aggregate subscores. On the one hand, if the subscores represent the same underlying trait, then they are only less reliable versions of a linear combination of the subscores (such as their total), and the report would be more useful if it contained this single score. On the other hand, each subscore may contain some unique information, in which case reporting of the subscores would be justified. The test administrators have to decide which (linear) combinations of the subscores to report. This problem can be formulated in a factor analytic framework, with college- and student-level factor structures.

The analysis reported in Longford (1990) proceeded as follows: A multivariate variance component model was fitted for the data, decomposing the fitted variance of a set of subscores into within- and between-college components, $\hat{\Sigma} = \hat{\Sigma}_W + \hat{\Sigma}_B$, and then these two estimated variance matrices were subjected to (informal) factor analyses. The decomposition of the total variance was carried out using the software VARCL (Longford, 1988). In section 4 a computationally more efficient method for this decomposition is described. Section 5 deals with the generalization for three levels of

nesting. An example with real data and one with artificially generated data, both with two levels, are discussed in section 6. See Muthén (1991) for a detailed background to the former example.

2. Models for Multilevel Factor Analysis

Suppose there are M groups indexed $j = 1, 2, \ldots, M$, and within each group j there are n_j p-variate normally distributed random vectors of observations, $\mathbf{y}_{ij}$, $i = 1, 2, \ldots, n_j$; the total number of observations is $N (= \Sigma_j n_j)$. Assume that, conditionally on the group mean $\mathbf{m}_j$, observations within each group have a common factor structure:

$$\mathbf{y}_{ij}|\mathbf{m}_j \sim \mathbf{N}_p(\mathbf{m}_j, \mathbf{V}_1), \qquad \text{(iid)} \tag{1}$$

$$\mathbf{V}_1 = \boldsymbol{\Lambda}_1 \boldsymbol{\Psi}_1 \boldsymbol{\Lambda}_1^\top + \boldsymbol{\Theta}_1,$$

where $\boldsymbol{\Lambda}_1$ is a $p \times r_1$ matrix of constants ($r_1 \leq p$), $\boldsymbol{\Psi}_1$ is an $r_1 \times r_1$ correlation matrix, and $\boldsymbol{\Theta}_1$ is a diagonal covariance matrix. Further, assume that the mean-vectors $\{\mathbf{m}_j\}$ also have an underlying factor structure; $\mathbf{m}_j \sim \mathbf{N}_p(\boldsymbol{\mu}, \mathbf{V}_2)$, iid, with

$$\mathbf{V}_2 = \boldsymbol{\Lambda}_2 \boldsymbol{\Psi}_2 \boldsymbol{\Lambda}_2^\top + \boldsymbol{\Theta}_2, \tag{2}$$

where $\boldsymbol{\Lambda}_2$ is a $p \times r_2$ matrix of constants ($r_2 \leq p$), $\boldsymbol{\Psi}_2$ an $r_2 \times r_2$ correlation matrix, and $\boldsymbol{\Theta}_2$ a diagonal covariance matrix. The marginal covariance matrix for a vector of observations is

$$(\mathbf{W}_j =)\ \operatorname{var}(\mathbf{y}_j) = \mathbf{V}_2 \otimes \mathbf{J}_{n_j} + \mathbf{V}_1 \otimes \mathbf{I}_{n_j}, \tag{3}$$

where $\otimes$ stands for the direct product, $\mathbf{y}_j$ is the vector of length pn_j composed of the n_j vectors $\mathbf{y}_{ij}$, $\mathbf{J}_n = \mathbf{1}_n \mathbf{1}_n^\top$ is the $n \times n$ matrix of ones ($\mathbf{1}_n$ is a column vector of ones of length n), and $\mathbf{I}_n$ is the $n \times n$ identity matrix. We also use the notation $\mathbf{0}_n$ for the (column) n-vector of zeros.

An alternative model description for (1) and (2), involving random terms, is

$$\mathbf{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\Lambda}_2 \boldsymbol{\delta}_{2,j} + \boldsymbol{\Lambda}_1 \boldsymbol{\delta}_{1,ij} + \boldsymbol{\varepsilon}_{2,j} + \boldsymbol{\varepsilon}_{1,ij}, \tag{4}$$

where $\boldsymbol{\delta}_{2j} \sim \mathbf{N}(\mathbf{0}_{r_2}, \boldsymbol{\Psi}_2)$, $\boldsymbol{\varepsilon}_{2j} \sim \mathbf{N}(\mathbf{0}_p, \boldsymbol{\Theta}_2)$, $\boldsymbol{\delta}_{1,ij} \sim \mathbf{N}(\mathbf{0}_{r_1}, \boldsymbol{\Psi}_1)$, and $\boldsymbol{\varepsilon}_{1,ij} \sim \mathbf{N}(\mathbf{0}_p, \boldsymbol{\Theta}_1)$ are mutually independent normal random vectors. In a typical application to assess differences among educational units (e.g., colleges or classrooms), the term $\boldsymbol{\delta}_{1,ij}$ represents the unexplained variation among the individuals, and the between-group term $\boldsymbol{\delta}_{2j}$ is often interpreted as the "effect" of the unit, the selection "effect," or as their aggregate. Linear regression can be easily accommodated in (4) by replacing the vector of means $\boldsymbol{\mu}$ with a general linear predictor term $\mathbf{x}_{ij}^\top \boldsymbol{\beta}$, in which $\mathbf{x}_{ij}$ is the matrix of the regression variables augmented by the dummy variables associated with the components of the random vector $\mathbf{y}_{ij}$. The matrix of regressors $\mathbf{x}_{ij}$ may also contain interactions (products) of these dummy variables with the explanatory variables. Denote the associated design matrix, formed by vertical stacking of the matrices $\mathbf{x}_{ij}$, by $\mathbf{X}$. The model (4) is a special case of the class of two-level structural relations models of McDonald and Goldstein (1989).

In exploratory factor analysis there are r_1^2 and r_2^2 indeterminacies in $\mathbf{V}_1$ and $\mathbf{V}_2$, respectively. Therefore, it is assumed that $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$ are identity matrices and impose $r_1(r_1 - 1)/2$ and $r_2(r_2 - 1)/2$ independent restrictions on $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2$, respectively. In confirmatory factor analysis it is assumed that $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$ are arbitrary correlation matrices. No loss of generality is involved by assuming unit variance of each component of $\boldsymbol{\delta}_{1,ij}$ and $\boldsymbol{\delta}_{2j}$ because scalar multiples can be accommodated in the matrices of

loadings Λ_1 and Λ_2. However, when constraints involving parameters for both $\mathbf{V}_1$ and $\mathbf{V}_2$ (cross-level constraints, such as $\Lambda_1 = \Lambda_2$) are imposed, only one of the matrices Ψ_1 and Ψ_2 is assumed to be a correlation or the identity matrix.

3. Maximum Likelihood Estimation

The log likelihood associated with the observations $\mathbf{y}$, $\mathbf{y}^\top = (\mathbf{y}_1^\top, \mathbf{y}_2^\top, \ldots, \mathbf{y}_M^\top)$, is

$$L = -\frac{1}{2}\{Np \log (2\pi) + \log (\det \mathbf{W}) + \operatorname{tr} (\mathbf{W}^{-1}\mathbf{e}\mathbf{e}^\top)\}, \tag{5}$$

where $\mathbf{e} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ is the Np-vector of residuals, and $\mathbf{W}$ is the covariance matrix for the observations $\mathbf{y}$, $\mathbf{W} = \operatorname{diag}(\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_M)$ with the block diagonals given by (3) corresponding to the groups. The block-pattern of the covariance matrices $\mathbf{W}_j$ enables expressions for their inverses and the determinants in terms of the inverses and determinants of the $(p \times p)$ matrices $\mathbf{H}_{2j} = \mathbf{V}_1 + n_j\mathbf{V}_2$:

$$\mathbf{W}_j^{-1} = \mathbf{V}_1^{-1} \otimes \mathbf{I}_{n_j} - (\mathbf{H}_{2j}^{-1}\mathbf{V}_2\mathbf{V}_1^{-1}) \otimes \mathbf{J}_{n_j}, \tag{6}$$

$$\det \mathbf{W}_j = (\det \mathbf{V}_1)^{n_j - 1} \det \mathbf{H}_{2j}. \tag{7}$$

Furthermore, if r_1 and r_2 are much smaller than p, then it is advantageous to use the identities

$$\mathbf{H}_{2j}^{-1} = \mathbf{H}_{1j}^{-1} - n_j\mathbf{H}_{1j}^{-1}\Lambda_2\Psi_2\mathbf{G}_{2j}^{-1}\Lambda_2^\top\mathbf{H}_{1j}^{-1}, \tag{8}$$

$$\mathbf{H}_{1j}^{-1} = \Theta^{-1} - \Theta^{-1}\Lambda_1\Psi_1\mathbf{G}_1^{-1}\Lambda_1^\top\Theta^{-1}, \tag{9}$$

$$\det \mathbf{H}_{2j} = \{\det (\mathbf{V}_1 + n_j\Theta_2)\} \det \mathbf{H}_1, \tag{10}$$

$$\det \mathbf{H}_1 = \det \Theta \det \mathbf{G}_1, \tag{11}$$

where $\mathbf{G}_{2j} = \mathbf{I}_{r_2} + n_j\Lambda_2^\top\mathbf{H}_{1j}^{-1}\Lambda_2\Psi_2$, $\mathbf{G}_1 = \mathbf{I}_{r_1} + \Lambda_1^\top\Theta^{-1}\Lambda_1\Psi_1$, $\mathbf{H}_{1j} = \mathbf{V}_1 + n_j\Theta_2$, and $\Theta = \Theta_1 + n_j\Theta_2$, so that it is necessary to invert or evaluate determinants of matrices of sizes r_1 and r_2 only. The familiar decomposition for the sample total sum of squares and crossproducts into its within- and between-group components is used:

$$\Sigma_j\Sigma_i\mathbf{e}_{ij}\mathbf{e}_{ij}^\top = \mathbf{T}_1 + \Sigma_j\mathbf{D}_j,$$

where $\mathbf{T}_1 = \Sigma_j\Sigma_i(\mathbf{e}_{ij} - \mathbf{e}_{\cdot j})(\mathbf{e}_{ij} - \mathbf{e}_{\cdot j})^\top$, $\mathbf{e}_{\cdot j}$ is the within-group residual, $\mathbf{e}_{\cdot j} = n_j^{-1}\Sigma_i\mathbf{e}_{ij} - \mathbf{e}_{\cdot\cdot}$, $\mathbf{e}_{\cdot\cdot} = N^{-1}\Sigma_j\Sigma_i\mathbf{e}_{ij}$, and $\mathbf{D}_j = \mathbf{n}_j(\mathbf{e}_{\cdot j}\mathbf{e}_{\cdot j}^\top + \mathbf{e}_{\cdot\cdot}\mathbf{e}_{\cdot\cdot}^\top)$. It can be shown by elementary operations that $\mathbf{V}_1^{-1} - \mathbf{H}_{2j}^{-1} = n_j\mathbf{H}_{2j}^{-1}\mathbf{V}_2\mathbf{V}_1^{-1}$, and by using (6) and (7) that

$$L = -\frac{1}{2}\{Np \log (2\pi) + (N - M) \log (\det \mathbf{V}_1) + \Sigma_j \log (\det \mathbf{H}_{2j}) + \operatorname{tr} (\mathbf{V}_1^{-1}\mathbf{T}_1) + \Sigma_j \operatorname{tr} (\mathbf{H}_{2j}^{-1}\mathbf{D}_j)\}. \tag{12}$$

Thus the matrix of raw sample crossproducts, $\mathbf{Y}^\top\mathbf{Y}$, and the vectors of within-group totals, $\{\Sigma_i\mathbf{y}_{ij}\}$, are a set of sufficient statistics for the model in (4). In the balanced case ($n_j \equiv n$), $\mathbf{H}_{2j}$ is constant across the groups ($\mathbf{H}_{2j} \equiv \mathbf{H}_2$), and (12) simplifies to

$$L = -\frac{1}{2}\{Mnp \log (2\pi) + M(n-1) \log (\det \mathbf{V}_1) + M \log (\det \mathbf{H}_2) + \operatorname{tr} (\mathbf{V}_1^{-1}\mathbf{T}_1) + \operatorname{tr} (\mathbf{H}_2^{-1}\mathbf{T}_2)\}, \quad (13)$$

where $\mathbf{T}_2 = \Sigma_j \mathbf{D}_j$. The log likelihood (13) depends on the data only by means of the within- and between-group sums of squares and crossproducts, $\mathbf{T}_1$ and $\mathbf{T}_2$, respectively; $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$ and $\Sigma_j(\mathbf{Y}_j^{\mathrm{T}}\mathbf{1})(\mathbf{Y}_j^{\mathrm{T}}\mathbf{1})^{\mathrm{T}}$ are a set of minimal sufficient statistics.

In the balanced case the cluster-vectors $\mathbf{Y}_j$ are iid, and so, subject to regularity conditions, standard asymptotic results apply: For large number of clusters ($M \to \infty$), the distribution of the maximum likelihood estimator of the model parameters is normal with the mean equal to the vector of parameters, and the variance equal to the inverse of the information matrix. Lee (1990) extends this result to the unbalanced case, assuming that the group-sizes n_j are close to the average group size $\bar{n}_M = (n_1 + \cdots + n_M)/M$. We conjecture that standard asymptotic results apply whenever all the limiting points of the sequence of average sizes, $\bar{n}_M$, are greater than 1.

The first- and second-order partial derivatives of (5) with respect to the vector of regression parameters $\boldsymbol{\beta}$ are

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}}\mathbf{W}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

$$\frac{\partial^2 L}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\mathrm{T}}} = -\mathbf{X}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{X},$$

respectively, so that the estimate of $\boldsymbol{\beta}$ is updated as

$$\hat{\boldsymbol{\beta}}_{\text{new}} = \hat{\boldsymbol{\beta}}_{\text{old}} + (\mathbf{X}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{old}}).$$

This implies the generalized least squares formula

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{y}. \quad (14)$$

Note that for the model in (4) (no explanatory variables), $\mathbf{X} = \mathbf{I}_p \otimes \mathbf{1}_N$, and $\mathbf{X}^{\mathrm{T}}\mathbf{W}^{-1} = (\mathbf{H}_{21}^{-1}, \mathbf{H}_{22}^{-1}, \ldots, \mathbf{H}_M^{-1})$. If in addition the clustering design is balanced, $n_j \equiv n$, then $\partial L/\partial \boldsymbol{\mu} = \mathbf{H}_2^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu})$, and so $\hat{\boldsymbol{\mu}} = \bar{\mathbf{y}}$.

For the derivative with respect to a general covariance structure parameter ϕ (involved in $\mathbf{V}_1$, $\mathbf{V}_2$, or both), we have

$$\frac{\partial L}{\partial \phi} = -\frac{1}{2}\left\{(N-M) \operatorname{tr}\left(\mathbf{V}_1^{-1}\frac{\partial \mathbf{V}_1}{\partial \phi}\right) + \sum_j \operatorname{tr}\left(\mathbf{H}_2^{-1}\frac{\partial \mathbf{H}_{2j}}{\partial \phi}\right) - \operatorname{tr}\left(\mathbf{V}_1^{-1}\frac{\partial \mathbf{V}_1}{\partial \phi}\mathbf{V}_1^{-1}\mathbf{T}_1\right) - \sum_j \operatorname{tr}\left(\mathbf{H}_{2j}^{-1}\frac{\partial \mathbf{H}_{2j}}{\partial \phi}\mathbf{H}_{2j}^{-1}\mathbf{D}_j\right)\right\}; \quad (15)$$

the derivative matrices $\partial \mathbf{V}_1/\partial \phi$ and $\partial \mathbf{H}_{2j}/\partial \phi$ have the general form

$$\frac{\partial \mathbf{U}}{\partial \phi} = (\mathbf{A}\boldsymbol{\Delta}_1\boldsymbol{\Delta}_2^{\mathrm{T}}\mathbf{B} + \mathbf{B}^{\mathrm{T}}\boldsymbol{\Delta}_2\boldsymbol{\Delta}_1^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}),$$

where $\mathbf{A}$ and $\mathbf{B}$ are matrices and $\boldsymbol{\Delta}_1$ and $\boldsymbol{\Delta}_2$ are indicator vectors. For example, for the (k, h) element of the matrix Λ_2 of factor loadings at level 2, we have

$$\frac{\partial \mathbf{H}_{2j}}{\partial \lambda_{2,kh}} = n_j(\mathbf{M}_{2,kh} + \mathbf{M}_{2,kh}^{\mathrm{T}}), \tag{16}$$

where $\mathbf{M}_{2,kh} = \mathbf{\Lambda}_2\mathbf{\Delta}_{k,r_2}\mathbf{\Delta}_{h,p}^{\mathrm{T}}$, and $\mathbf{\Delta}_{i,p}$ is the $p \times 1$ (indicator) vector of length p containing a 1 in its i-th position and zeros elsewhere. The formulas for the first-order partial derivatives are obtained by substitution of (16) and of the inversion formulas (6), (8), and (9) into (15).

The terms required for an element of the scoring vector involving a parameter represented in $\mathbf{V}_1$ or $\mathbf{V}_2$ are:

$$\mathrm{tr}\left(\mathbf{H}_{2j}^{-1}\frac{\partial \mathbf{H}_{2j}}{\partial \phi}\right) = 2n_j\mathbf{\Delta}_2^{\mathrm{T}}\mathbf{B}\mathbf{H}_{2j}^{-1}\mathbf{A}\mathbf{\Delta}_1,$$

which is a weighted sum of an element of the matrices $\mathbf{B}\mathbf{H}_{2j}^{-1}\mathbf{A}$,

$$\mathrm{tr}\left(\mathbf{H}_{2j}^{-1}\frac{\partial \mathbf{H}_{2j}}{\partial \phi}\mathbf{H}_{2j}^{-1}\mathbf{D}_j\right) = 2n_j^2\mathbf{e}_j^{\mathrm{T}}\mathbf{H}_{2j}^{-1}\mathbf{A}\mathbf{\Delta}_1\mathbf{e}_j^{\mathrm{T}}\mathbf{H}_{2j}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{\Delta}_2,$$

and additionally, for an element of the scoring vector for a parameter in $\mathbf{V}_1$,

$$\mathrm{tr}\left(\mathbf{V}_1^{-1}\frac{\partial \mathbf{V}_1}{\partial \phi}\right) = 2\mathbf{\Delta}_2^{\mathrm{T}}\mathbf{B}\mathbf{V}_1^{-1}\mathbf{A}\mathbf{\Delta}_1,$$

and

$$\mathrm{tr}\left(\mathbf{V}_1^{-1}\frac{\partial \mathbf{V}_1}{\mathbf{V}_1^{-1}}\mathbf{V}_1^{-1}\mathbf{T}_1\right) = 2\mathbf{\Delta}_2^{\mathrm{T}}\mathbf{B}\mathbf{V}_1^{-1}\mathbf{T}_1\mathbf{V}_1^{-1}\mathbf{A}\mathbf{\Delta}_1.$$

Second-order Partial Derivatives

A general second-order partial derivative of the log-likelihood (5) with respect to a pair of elements involved in the covariance matrix $\mathbf{W}$ is equal to

$$\frac{\partial^2 L}{\partial \phi_1 \partial \phi_2} = \frac{1}{2}\mathrm{tr}\left(\mathbf{W}^{-1}\frac{\partial \mathbf{W}}{\partial \phi_1}\mathbf{W}^{-1}\frac{\partial \mathbf{W}}{\partial \phi_2}\right) - \frac{1}{2}\mathrm{tr}\left\{\mathbf{W}^{-1}\frac{\partial^2 \mathbf{W}}{\partial \phi_1 \partial \phi_2}\right\}$$
$$- \mathbf{e}^{\mathrm{T}}\mathbf{W}^{-1}\frac{\partial \mathbf{W}}{\partial \phi_1}\mathbf{W}^{-1}\frac{\partial \mathbf{W}}{\partial \phi_2}\mathbf{W}^{-1}\mathbf{e} + \frac{1}{2}\mathbf{e}^{\mathrm{T}}\mathbf{W}^{-1}\frac{\partial^2 \mathbf{W}}{\partial \phi_1 \partial \phi_2}\mathbf{W}^{-1}\mathbf{e}, \tag{17}$$

and the negative of its expectation, the (ϕ_1, ϕ_2) element of the information matrix, is equal to

$$\{U(\phi_1, \phi_2) = \} \quad \frac{1}{2}\mathrm{tr}\left(\mathbf{W}^{-1}\frac{\partial \mathbf{W}}{\partial \phi_1}\mathbf{W}^{-1}\frac{\partial \mathbf{W}}{\partial \phi_2}\right). \tag{18}$$

The form of the expectation in (18) is substantially simpler than that of the derivatives in (17), and so the Fisher-scoring algorithm is preferred to the Newton-Raphson one. In fact, the "building blocks" for (18), $\mathbf{W}^{-1}\partial\mathbf{W}/\partial\phi$, are also required for the first-order partial derivatives. For a pair of parameters ϕ_1 and ϕ_2 involved in $\mathbf{V}_2$ we have

$$U(\phi_1, \phi_2) = \Sigma_j n_j^2(\mathbf{\Delta}_{B2}^{\mathrm{T}}\mathbf{B}_2\mathbf{H}_{2j}^{-1}\mathbf{A}_1\mathbf{\Delta}_{A1}\mathbf{\Delta}_{B1}^{\mathrm{T}}\mathbf{B}_1\mathbf{H}_{2j}^{-1}\mathbf{A}_2\mathbf{\Delta}_{A2}$$
$$+ \mathbf{\Delta}_{A2}\mathbf{A}_2\mathbf{H}_{2j}^{-1}\mathbf{A}_1\mathbf{\Delta}_{A1}\mathbf{\Delta}_{B1}^{\mathrm{T}}\mathbf{B}_1\mathbf{H}_{2j}^{-1}\mathbf{B}_2^{\mathrm{T}}\mathbf{\Delta}_{B2}),$$

for the appropriate matrices $\mathbf{A}_1$, $\mathbf{B}_1$, $\mathbf{A}_2$, and $\mathbf{B}_2$ and indicator vectors Δ_{A1}, Δ_{A2}, Δ_{B1}, and Δ_{B2}. The element of the information matrix corresponding to a pair of parameters for $\mathbf{V}_1$ is equal to

$$\begin{aligned} U(\phi_1, \phi_2) = (N - M)(&\Delta_{B2}^{\top}\mathbf{B}_2\mathbf{V}_1^{-1}\mathbf{A}_1\Delta_{A1}\Delta_{B1}^{\top}\mathbf{B}_1\mathbf{V}_1^{-1}\mathbf{A}_2\Delta_{A2} \\ &+ \Delta_{A2}\mathbf{A}_2\mathbf{V}_1^{-1}\mathbf{A}_1\Delta_{A1}\Delta_{B1}^{\top}\mathbf{B}_1\mathbf{V}_1^{-1}\mathbf{B}_2^{\top}\Delta_{B2}) \\ &+ \Sigma_j(\Delta_{B2}^{\top}\mathbf{B}_2\mathbf{H}_{2j}^{-1}\mathbf{A}_1\Delta_{A1}\Delta_{B1}^{\top}\mathbf{B}_1\mathbf{H}_{2j}^{-1}\mathbf{A}_2\Delta_{A2} \\ &+ \Delta_{A2}\mathbf{A}_2\mathbf{H}_{2j}^{-1}\mathbf{A}_1\Delta_{A1}\Delta_{B1}^{\top}\mathbf{B}_1\mathbf{H}_{2j}^{-1}\mathbf{B}_2^{\top}\Delta_{B2}), \end{aligned}$$

and the element corresponding to a parameter ϕ_1 for $\mathbf{V}_1$ and a parameter ϕ_2 for $\mathbf{V}_2$ is

$$\begin{aligned} U(\phi_1, \phi_2) = \Sigma_j n_j(&\Delta_{B2}^{\top}\mathbf{B}_2\mathbf{H}_{2j}^{-1}\mathbf{A}_1\Delta_{A1}\Delta_{B1}^{\top}\mathbf{B}_1\mathbf{H}_{2j}^{-1}\mathbf{A}_2\Delta_{A2} \\ &+ \Delta_{A2}\mathbf{A}_2\mathbf{V}_1^{-1}\mathbf{A}_1\Delta_{A1}\Delta_{B1}^{\top}\mathbf{B}_1\mathbf{V}_1^{-1}\mathbf{B}_2^{\top}\Delta_{B2}). \end{aligned}$$

Elements of the information matrix for parameters involved in both $\mathbf{V}_1$ and $\mathbf{V}_2$ are linear combinations of these expressions.

If $\mathbf{W}$ is known estimation of $\boldsymbol{\beta}$ requires a single iteration of (14). Otherwise $\mathbf{W}$ in (14) is replaced by its current estimate, $\hat{\mathbf{W}}$, and since in the iterations of the Fisher scoring algorithm the estimate of $\mathbf{W}$ is updated, $\hat{\boldsymbol{\beta}}$ also has to be updated.

In summary, the elements of the information matrix for the covariance structure parameters are multiples of (sums of) products of elements of the quadratic forms $\mathbf{A}^{\top}\mathbf{F}\mathbf{B}$, where $\mathbf{F}$ is either $\mathbf{V}_1^{-1}$ or $\mathbf{H}_{2j}^{-1}$, and $\mathbf{A}$ and $\mathbf{B}$ are $\mathbf{I}$, Λ_h, or $\Lambda_h\Psi_h$, $(h = 1, 2)$. These formulas have a similar form to those for the two-level regression model (see Longford, 1987; or Jennrich & Schluchter, 1986), because the covariance structures corresponding to these models are similar.

The iterations of the Fisher-scoring algorithm involve calculation of the vector of corrections for the estimated parameters based on the vectors of first-order partial derivatives and the expected information matrix, both evaluated for the current values of the parameters. A choice for the starting solution is discussed in section 4. At each iteration an updated solution is obtained, and iterations are stopped when the corrections to the current solution are small (have a small norm), the value of the log likelihood L changes by less than a prescribed tolerance and/or the norm of the scoring vector is sufficiently small. The algorithm can be easily adapted for various constraints on the parameters by application of the chain-rule (e.g., constraining a set of parameters to be equal), or of the method of Lagrange multipliers (e.g., for orthogonality of the factors), although for the latter the convergence properties usually get worse. Similar problems are often encountered in classical confirmatory analysis. Standard adaptations of the algorithm, such as step-halving, can be employed, although they appeared not necessary in the examples discussed in section 6.

Hypothesis testing can be based on the value of the deviance (-2 log likelihood), the calculation of which is a minor component of the scoring algorithm. Note that the likelihood ratio test statistics have the χ^2 distribution with the usual number of degrees of freedom only when the vector of true parameters is in the interior of the parameter space. In particular, when one of the correlation matrices Ψ_h is singular, the difference of the deviances of two nested models does not have a χ^2 distribution. The familiar problem of Heywood cases is equally applicable to two-level factor analysis.

For problems with a large number of parameters, quasi-Newton methods (see, e.g., Luenberger, 1984) can be used with advantage. The models presented here share the problem of large numbers of parameters with the confirmatory mode of the classical

factor analysis, and in the two-level factor analysis this issue is even more acute. Quasi-Newton methods require evaluation of the scoring function, but avoid inversion of the information matrix (Hessian) by approximating its inverse through the iterations. A suitable initial approximation for the inverse of the information matrix is obtained by inverting the block-diagonal matrix, with the blocks corresponding to the Θ- and to the columns of the Λ-parameters, formed from the information matrix by deleting the elements outside the diagonal blocks. The LISREL software for structural equation modeling (Jöreskog & Sörbom, 1979) also employs a quasi-Newton method. Exploring estimation procedures for classical factor analysis, Jamshidian and Jennrich (1988) found that simple quasi-Newton and conjugate gradient methods have reasonable convergence properties, and are computationally more economic than the Newton-Raphson algorithm, principally because they avoid inversion of very large Hessian matrices. Asymptotic standard errors for the estimated parameters are obtained from the inverse of the fitted information matrix. See Lee and Jennrich (1979) for derivation of standard errors when a quasi-Newton method is used.

4. The Saturated Factor Analysis Model

In many settings it is of interest to compare the adopted parsimonious factor analysis model with the saturated factor analysis model that contains p factors at either level. The saturated model contains $p \times (p + 1)$ covariance structure parameters, and so each iteration of the algorithm would require solving a system of $p \times (p + 1)$ linear equations. This may be impractical even for moderately large p (say, $p > 12$).

The saturated two-level factor analysis model can be fitted by a noniterative procedure based on the method of moments. The likelihood in (5), as a function of the parameters $(\boldsymbol{\mu}, \boldsymbol{\Omega}) = (\boldsymbol{\mu}, \boldsymbol{\Theta}_1, \boldsymbol{\Lambda}_1, \boldsymbol{\Psi}_1, \boldsymbol{\Theta}_2, \boldsymbol{\Lambda}_2, \boldsymbol{\Psi}_2)$, belongs to the exponential family of distributions; it has the general form

$$\exp\{a_f(\boldsymbol{\mu}) + a_c(\boldsymbol{\Omega}) + \mathbf{u}(\boldsymbol{\mu}, \boldsymbol{\Omega})^{\top}\mathbf{b}(\mathbf{y})\},$$

where a_f and a_c are real functions, and $\mathbf{u}$ and $\mathbf{b}$ are vector functions of the parameters and the data, respectively. Note that $\mathbf{b}$ is a linear function of the sufficient statistics $\mathbf{T}_1$ and $\{\mathbf{D}_j\}$. The scoring vector for $\boldsymbol{\Omega}$ has the form

$$\frac{\partial a_c(\boldsymbol{\Omega})}{\partial \boldsymbol{\Omega}} + \frac{\partial \mathbf{u}(\boldsymbol{\mu}, \boldsymbol{\Omega})^{\top}}{\partial \boldsymbol{\Omega}}\mathbf{b}(\mathbf{y}), \tag{19}$$

and since the expectation of the scoring vector is equal to $\mathbf{0}$,

$$\mathrm{E}\{\mathbf{b}(\mathbf{y})|\boldsymbol{\mu}, \boldsymbol{\Omega}\} = -\left\{\frac{\partial \mathbf{u}(\boldsymbol{\mu}, \boldsymbol{\Omega})^{\top}}{\partial \boldsymbol{\Omega}^{-1}}\right\}\frac{\partial a_c(\boldsymbol{\Omega})}{\partial \boldsymbol{\Omega}};$$

in other words, $\mathrm{E}\{\mathbf{b}(\mathbf{y})|\boldsymbol{\mu}, \boldsymbol{\Omega}\}$ is the root of (19). If the maximum likelihood estimate $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Omega}})$ is in the interior of the parameter space, then it is also a root of (19), and so necessarily $\mathbf{b}(\mathbf{y}) = \mathrm{E}\{\mathbf{b}(\mathbf{y}|\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Omega}})\}$. Hence, maximum likelihood estimation is equivalent to moment matching. In particular, the expectation of any statistic that is a linear function of $\mathbf{b}(\mathbf{y})$, such as $\mathbf{T}_1$ and $\mathbf{T}_2 = \Sigma_j \mathbf{D}_j$, evaluated at the maximum likelihood solution, is equal to this statistic. In addition,

$$\mathrm{E}\{\mathbf{T}_1|\boldsymbol{\mu}, \boldsymbol{\Omega}\} = (N - M)\mathbf{V}_1, \text{ and}$$

$$\mathrm{E}\{\mathbf{T}_2|\boldsymbol{\mu}, \boldsymbol{\Omega}\} = M\mathbf{V}_1 + \left(N - \frac{\Sigma_j n_j^2}{N}\right)\mathbf{V}_2.$$

Hence, the maximum likelihood estimators are given by the formulas:

$$\hat{\mathbf{V}}_1 = (N - M)^{-1}\mathbf{T}_1, \tag{20}$$

$$\hat{\mathbf{V}}_2 = \left(N - \frac{\Sigma_j n_j^2}{N}\right)^{-1} (\mathbf{T}_2 - M\hat{\mathbf{V}}_1). \tag{21}$$

Models with a saturated factor structure at one level but a restricted structure at the other level can be efficiently fitted by combining moment matching (for the covariance matrix at the saturated level) and the scoring algorithm (for the parameters at the level with restrictions). For example, a model with restricted factor structure for the individual level and a saturated one at the group level could be fitted by the following iterative procedure: Start with the initial solution given by (20) and the decomposition of (21), iteratively estimate the individual-level parameters by the Fisher-scoring algorithm at each iteration, and obtain the fitted variance matrix for the group level from (21). Note that $\hat{\mathbf{V}}_1$ does not depend on $\hat{\mathbf{V}}_2$, and so when fitting models with saturated factor structure at the individual level, $\hat{\mathbf{V}}_1$ does not change with the iterations for $\hat{\mathbf{V}}_2$.

Restricted maximum likelihood estimates (REML, Patterson & Thompson, 1971) for any factor analysis model can be computed by adjusting the scoring vector by the partial derivatives of $\frac{1}{2}\log\{\det(\mathbf{X}^\top\mathbf{W}^{-1}\mathbf{X})\}$ (see Harville, 1974). REML is preferable to ML when the location parameters are nuisance parameters, as is usually the case when no explanatory variables are considered.

Two-Stage Procedures

The problem of different covariance structures for individual and aggregate data has been well understood and various two-stage procedures have been used to estimate separate covariance structures for each level of aggregation. These procedures rely on a decomposition of the total variation into within- and between-group covariance components (matrices), which are then subjected to separate factor analyses (see Cronbach, 1976; and Härnquist, 1978, for examples). The estimators (20) and (21) provide such a decomposition. Although such a procedure is not fully efficient, the loss of efficiency may be quite modest: The within-group statistic $\mathbf{T}_1$ has the Wishart distribution $\mathbf{W}_p(\mathbf{V}_1, N - M)$, and since our data contain less information than N independent observations from $\mathbf{N}_p(\boldsymbol{\mu}, \mathbf{V}_1)$, the efficiency of the two-stage estimator for the within-group factor structure is at least $1 - M/(N - 1)$. Since $\mathbf{T}_1$ contains no information about $\mathbf{V}_2$, for balanced data the two-stage estimator for $\mathbf{V}_2$ is almost fully efficient. Loss of efficiency of this estimator is likely to be most serious for extremely unbalanced data although the analysis of the SIMS dataset in section 6 is an example to the contrary. In any case, this two-stage estimator for the within- and between-group factor structures is a suitable initial solution for the scoring (or any other iterative) algorithm.

5. Multiple Levels of Nesting

Extensions of the factor analysis model to multiple layers of nesting present no conceptual difficulties. For example, in a three-level model assume that each *district* (with groups, and individuals within groups) has the two-level factor structure (4), with common factor structure parameters $\boldsymbol{\Lambda}_h$, $\boldsymbol{\Psi}_h$, and $\boldsymbol{\Theta}_h$, $(h = 1, 2)$; but the (conditional) mean, $\boldsymbol{\mu}$ in (4), is a random vector with the distribution $\mathbf{N}_p(\boldsymbol{\mu}^*, \mathbf{V}_3)$, where $\mathbf{V}_3 = \boldsymbol{\Theta}_3 + \boldsymbol{\Lambda}_3\boldsymbol{\Psi}_3\boldsymbol{\Lambda}_3^\top$. The parameters contained in the matrices $\boldsymbol{\Lambda}_3$, $\boldsymbol{\Psi}_3$, and $\boldsymbol{\Theta}_3$, describe the between-district factor structure.

The two-stage algorithm has a straightforward extension for estimation with the

three-level model; the decomposition of the total sum of squares and crossproducts into its within-group, within-district, and between-district components, and the moment-matching equations analogous to (20) and (21) yield:

$$\hat{\mathbf{V}}_1 = (N - M)^{-1}\mathbf{T}_1,$$

$$\hat{\mathbf{V}}_2 = \frac{\mathbf{T}_2 - (M - Q)\hat{\mathbf{V}}_1}{N - \sum_k \frac{\Sigma_j n_{jk}^2}{N_k}},$$

$$\hat{\mathbf{V}}_3 = \frac{\mathbf{T}_3 - \left(\sum_k \frac{\Sigma_j n_{jk}^2}{N_k} - \frac{\Sigma_k \Sigma_j n_{jk}^2}{N}\right)\hat{\mathbf{V}}_2 - (Q - 1)\hat{\mathbf{V}}_1}{N - \frac{\Sigma_k N_k^2}{N}},$$

where $N_k = \Sigma_j n_{jk}$ is the number of observations in the district k, N, M and Q are the respective numbers of individuals, groups and districts, and $\mathbf{T}_h$ ($\hat{\mathbf{V}}_h$), $h = 1, 2, 3$, are the within-group, between-group, and between-district matrices of the sums of squares and crossproducts (unbiased estimates of covariance matrices), respectively.

The Fisher scoring algorithm can be extended for the three-level analysis. The formulas for the scoring vector and the information matrix become more extensive but inversion of matrices can still be restricted to matrices of small sizes, with the exception of one inversion of the expected information matrix per iteration. To avoid the latter, a quasi-Newton method or various adaptations of the Fisher-scoring algorithm can be used.

The covariance matrix for all the observations has the form

$$(\mathbf{W}_3 =)\ \mathrm{diag}_k(\mathbf{W}_{2,k} + \mathbf{V}_3 \otimes \mathbf{J}_{M_k}).$$

where $\mathbf{W}_{2,k} = \mathbf{V}_1 \otimes \mathbf{I}_{N_k} + \mathbf{V}_2 \otimes \mathbf{J}_{N_k}$ is the (conditional) within-district covariance matrix for the district k. The inverse and the determinant of $\mathbf{W}_3$ are given by the formulas

$$\mathbf{W}_3^{-1} = \mathrm{diag}_k(\mathbf{W}_{2,k}^{-1}) - \mathrm{diag}_k(\mathbf{W}_{2,k}^{-1})\mathbf{K}_k\mathbf{V}_3\mathbf{G}_{3,k}^{-1}\mathbf{K}_k^{\top}\ \mathrm{diag}_k(\mathbf{W}_{2,k}^{-1}), \tag{22}$$

$$\det \mathbf{W}_3 = \det \mathbf{W}_{2,k}\ \det \mathbf{G}_{3,k},$$

where $\mathbf{K}_k = \mathbf{I}_p \otimes \mathbf{1}_{N_k}$, and $\mathbf{G}_{3,k} = \mathbf{I}_p + \mathbf{K}_k^{\top}\mathbf{W}_{2,k}^{-1}\mathbf{K}_k$. The inverses and the determinants of the matrices $\mathbf{W}_{2,k}$ are given by (6) and (7), respectively. The scoring function is now obtained by formal differentiation of the log-likelihood (5), with $\mathbf{W} = \mathbf{W}_3$, and the derivatives can be economically evaluated using (22). Even the simplest versions of these models are likely to contain large numbers of parameters that render the use of the Fisher scoring method in its original form computationally inefficient, and in general it is preferable to use a quasi-Newton method. A natural starting solution for any iterative fitting algorithm is provided by the means (ordinary regression solution) and the classical factor analysis solutions for the moment estimates of the subject-, group- and district-level covariance matrices $\hat{\mathbf{V}}_1$, $\hat{\mathbf{V}}_2$, and $\hat{\mathbf{V}}_3$, ignoring any constraints for covariance structure parameters across the levels of nesting.

6. Examples

A Simulated Dataset

A simulation of a balanced 5-variate dataset with 50 groups with 10 subjects in each group, according to the model in (4), with

$$\boldsymbol{\mu} = \mathbf{0},\ \boldsymbol{\Lambda}_1 = \boldsymbol{\Lambda}_2 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 0 & -1 & 1 \end{pmatrix}^{\top},$$

$$\boldsymbol{\Psi}_1 = \mathbf{I}_2,\ \boldsymbol{\Psi}_2 = \begin{pmatrix} 1.25 & 1 \\ 1 & 1.25 \end{pmatrix},$$

and $\boldsymbol{\Theta}_1 = \mathbf{I}$, and $\boldsymbol{\Theta}_2 = 0.2\mathbf{I}$ was used to demonstrate the scoring algorithm with constrained maximization. In the fitted models $\boldsymbol{\Theta}_1$, $\boldsymbol{\Theta}_2$ and $\boldsymbol{\mu}$ are always estimated. When no constraints on $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2$ are imposed $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$ are both constrained to the identity matrix. When a common set of factor loadings is assumed ($\boldsymbol{\Lambda}_1 = \boldsymbol{\Lambda}_2$, and the common value is denoted by $\boldsymbol{\Lambda}$), $\boldsymbol{\Psi}_1 = \mathbf{I}$ was set and $\boldsymbol{\Psi}_2$ was estimated as an arbitrary covariance matrix. To ensure identifiability of $\boldsymbol{\Lambda}$ the constraint that $\boldsymbol{\Lambda}^{\top}\boldsymbol{\Theta}_1^{-1}\boldsymbol{\Lambda}$ be diagonal was imposed.

The deviance of the saturated model solution given by (20) and (21) is 9199.54. The level-wise decomposition of this solution for one factor each ($r_1 = r_2 = 1$, $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2$ unrelated, and $\boldsymbol{\Psi}_1 = \boldsymbol{\Psi}_2 = 1$) has the deviance 9622.85, and the deviance for the corresponding maximum likelihood estimate is 9609.54. If the cross-level constraint $\boldsymbol{\Lambda}_1 = \boldsymbol{\Lambda}_2$ is imposed and $\boldsymbol{\Psi}_2$ estimated, the deviance increases to 9785.15. Clearly, no model with one factor at each level provides a satisfactory fit for the data. Also, relationship of the factor structures at the individual and group levels using two factors cannot be assessed from a single factor solution.

The top panel of Table 1 gives the maximum likelihood solution for the exploratory model with two factors at each level ($\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2$ unrelated). The deviance of this model, 9199.75, is only 0.21 higher than for the saturated model which involves two additional free parameters. The bottom panel of Table 1 contains the maximum likelihood solution for the two-factor model with a common matrix of factor loadings $\boldsymbol{\Lambda}(\boldsymbol{\Psi}_1 = \mathbf{I}, \boldsymbol{\Psi}_2$ estimated). The associated deviance is 9202.04, 2.50 higher than the deviance for the saturated model (30 parameters), but the model contains only 22 free parameters (9 involved in $\boldsymbol{\Lambda}$, 10 in $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ and 3 in $\boldsymbol{\Psi}_2$).

Initial estimates for the variances in $\boldsymbol{\Psi}_2$ were obtained by a naive noniterative method, and the initial value for the covariance was set to 0. For the single factor models the scoring algorithm, starting with the moment solution described in section 4, required 6 iterations, and one or two additional iterations when starting with more distant solutions. Convergence for two-factor models was much slower, requiring up to 20 iterations. No multiple local maxima of the log-likelihood function were found for any of these models. Iterations were terminated when the change in the value of the deviance was less than 0.001. In each instance, the corrections of the estimated parameters as well as the scoring vector had a norm smaller than 10^{-4}. With very rare exceptions the deviance decreased at every iteration. For multifactor models the constraint of orthogonality of the factor loadings was implemented by the Lagrange multiplier method.

Comparison of the deviances indicates that the model used for generating data would be selected by the likelihood ratio criterion. The elements of the "fitted − observed" difference matrices are in the ranges $-0.08 - 0.06$ for the within-group ($\hat{\mathbf{V}}_1$

TABLE 1

Maximum Likelihood Solutions for Two-Factor Models with Unrelated Factor Structures and Common Factor Loadings

$\mathrm{diag}(\Theta_1)$					
	1.026	1.048	0.975	0.908	0.967
Λ_1^T					
	0.916	1.216	1.029	1.210	0.919
	1.138	-0.876	0.097	-0.893	1.143
$\mathrm{diag}(\Theta_2)$					
	0.197	0.200	0.206	0.137	0.271
Λ_2^T					
	-2.268	0.313	-0.936	0.476	-2.215
	-1.287	-0.454	-0.854	-0.401	1.080

Note: Deviance is 9199.75.

$\mathrm{diag}(\Theta_1)$					
	0.975	1.065	0.968	0.890	1.033
Λ^T					
	0.988	1.052	0.988	1.026	0.947
	0.976	-0.933	-0.003	-0.985	0.945
$\mathrm{diag}(\Theta_2)$					
	0.208	0.188	0.216	0.150	0.274
Ψ_2	$\begin{pmatrix} 1.157 & 1.269 \\ 1.269 & 1.686 \end{pmatrix}$				

Note: Deviance is 9202.04.

$-\hat{\Sigma}_1$), $-0.16 - 0.13$ for between-group ($\hat{V}_2 - \hat{\Sigma}_2$), and $-0.08 - 0.10$ for the total covariance matrices ($\hat{V}_1 + \hat{V}_2 - T_1 - T_2$). Asymptotic standard errors (referring to $M \to \infty$) for all the estimated parameters can be obtained from the scoring method. The standard errors for the parameters in Λ are in the range $0.05 - 0.08$, for Θ_1 $0.08 - 0.13$, and for Θ_2 $0.08 - 0.15$. The standard errors for the elements of Ψ_2, (1.157, 1.269, 1.686), are (0.273, 0.325, 0.291). Note that although Ψ_2 and $\hat{\Psi}_2$ are almost singular, the information submatrix for the unique elements of Ψ_2 is well-conditioned.

Second International Mathematics Study

The Mathematics achievement data on U.S. eighth-grade students from the Second International Mathematics Study (Crosswhite, Dossey, Swafford, McKnight, & Cooney, 1985) was used to further illustrate the two-level factor analysis techniques. In the study, two classes were randomly selected within each of a set of schools and school districts. An achievement test was administered at the end of Spring 1982 containing 75 items in the areas of arithmetic, algebra, geometry, and measurement. Our analysis considers the items of the core and a rotated form (Form A) taken by 819 students from 179 different classrooms (within 112 schools). A simple random sampling of classrooms is assumed, ignoring the clustering of classes within schools and school districts. Eight achievement variable totals were created from the 75 right-wrong scored items corresponding to two sums for each of arithmetic, algebra, measurement, and geometry. The topic classifications used in the study were: ratio, proportion, percent (11 items); common and decimal fractions (12 items); equations and expressions (10 items); integers, numbers (9 items); standard units, estimation (6 items); area, volume (5 items); coordinates, visualization (9 items); plane figures (8 items). Note that the discrete nature of these subscores renders the validity of the factor analysis contentious.

The U.S. eighth grade mathematics curriculum varies a great deal across classrooms due to tracking (selection) into remedial, typical, algebra, and enriched classes. In each type of class a different mix of topics is taught. In more advanced classes algebra and geometry are taught earlier than in typical classes that focus more on arithmetic. In this way, opportunity to learn varies for the eight studied achievement variables from classroom to classroom. Assuming that within-class correlations follow a unidimensional model and that the between-classroom variation is to a large extent due to tracking based on previous overall mathematics performance, a one-factor model at each level is applied. The size of the between-class variation of the individual variables and its factor structure are of principal interest. Due to the variable-specific variation in learning opportunities, the group-level variances Θ_2 were expected to be positive and large.

Over 70% of the classrooms (129) have 3–6 students in the sample. There are 10 classrooms with a single student in the sample, while the classroom with the largest representation has 10 students in the sample. The maximum likelihood solution is given in the top panel of Table 2. The goodness of fit statistic, comparing this solution ($\hat{V}_1$, $\hat{V}_2$) with the saturated model fit (20) and (21), ($\hat{V}_{1s}$, $\hat{V}_{2s}$), is equal to 60.4 (χ^2 null-distribution with 40 df), indicating that the model fit could be improved by including another factor at either or both levels. Inspection of the differences of the saturated and the fitted (single factor) variance matrices suggests that the model fit can be significantly improved at both levels. In particular, most of the entries in the matrix of differences $\hat{V}_{2s} - \hat{V}_2$ are positive. The maximum likelihood solution for the (exploratory) model with two factors at each level is displayed in the bottom panel of Table 2. Note that a Heywood case is obtained; the solution at an iteration is such that the matrices H_{2j} for

TABLE 2

Maximum Likelihood Solution for the Single Factors Model and the Model with Two Factors at Both Levels; SIMS Dataset.

$\mathbf{diag}(\Theta_1)$								
	2.103	2.502	1.668	1.625	1.002	0.885	1.502	1.666
Λ_1^T								
	1.725	1.641	1.133	1.236	0.794	0.722	0.865	0.984
$\mathbf{diag}(\Theta_2)$								
	0.095	0.231	0.297	0.078	0.074	0.108	0.113	0.179
Λ_2^T								
	1.728	1.943	1.770	1.551	0.898	0.707	0.879	1.121

Note: Deviance is 23,834.65.

$\mathbf{diag}(\Theta_1)$								
	2.115	2.218	1.579	1.622	1.004	0.878	1.396	1.647
Λ_1^T								
	1.705	1.666	1.150	1.245	0.790	0.723	0.876	0.971
	0.147	-0.406	-0.242	0.104	0.047	0.102	0.336	0.176
$\mathbf{diag}(\Theta_2)$								
	0.000[1]	0.151	0.100	0.064	0.074	0.108	0.075	0.164
Λ_2^T								
	1.757	1.965	1.774	1.536	0.902	0.701	0.866	1.137
	0.269	0.263	-0.407	-0.158	0.036	-0.049	-0.167	0.107

Note: Deviance is 23,805.74.

[1] Constrained to zero.

the largest groups j have a negative eigenvalue. In this case, a variance has a negative estimate at an iteration; it is set to zero, and from then on its value remains unaltered.

The difference of the deviances for the two discussed factor analysis models is 31.1. Because of the Heywood case, the null-distribution of this statistic does not have a χ^2 distribution with 14 degrees of freedom, but the improvement in the model fit can be assessed informally. While the fit for the student level appears to be very good, the differences $\hat{\mathbf{V}}_{2s} - \hat{\mathbf{V}}_2$ are reduced only marginally. Addition of a third factor at the classroom level leads to further negative estimated variances in $\boldsymbol{\Theta}_2$, and smaller reduction of the deviance, but the differences $\hat{\mathbf{V}}_{2s} - \hat{\mathbf{V}}_2$ get reduced substantially. Similarly, addition of another factor at the student level improves the model fit only marginally; the deviance is reduced by less than the associated number of degrees of freedom. It appears that for the classroom level, the informal model checking procedure is in conflict with the likelihood ratio criterion. (We conjecture that this is caused by the severely unbalanced nature of the data.) In the analysis of the simulated dataset the likelihood ratio criterion appears to be in agreement with the informal model checking. Lack of balance and association of outcomes with cluster size are illustrated by substantial differences between the arithmetic mean $\bar{\mathbf{y}}$,

$$\bar{\mathbf{y}} = (5.22,\ 6.86,\ 4.77,\ 4.82,\ 3.84,\ 2.00,\ 3.93,\ 3.68)$$

and the maximum likelihood estimate $\hat{\boldsymbol{\mu}}$ (which is essentially independent of the factor structure specification),

$$\hat{\boldsymbol{\mu}} = (5.07,\ 6.68,\ 4.62,\ 4.67,\ 3.76,\ 1.94,\ 3.86,\ 3.58).$$

Note that $\hat{\boldsymbol{\mu}}$ is uniformly smaller than $\bar{\mathbf{y}}$.

In conclusion, the first factor at both levels can be described as weighted mean scores, with weights approximately proportional to the number of items in each topic classification, that is, the total score is a suitable description for the first factor at each level. Second factors are significant at both levels, but they appear to be mutually different. Search for a suitable interpretation would proceed by rotation of the factors, in analogy with the classical exploratory factor analysis, and therefore it is not pursued here. Note that the expectation that $\boldsymbol{\Theta}_2$ is positive and large is not fulfilled by the two-factors solution. These results are in agreement with those obtained by Muthén (in press). See Muthén (1991) for a fuller analysis of the dataset.

The factor analyses reported above have been replicated for the same dataset with the 112 schools as clusters, ignoring classroom identification. Very similar results were obtained. A three-level analysis, accounting for clustering of students within classrooms within schools is likely to result in severe confounding of the parameter estimates because most schools have only one classroom in the data. For both datasets, and all the models fitted, the deviances of the initial solutions based on the moment estimates $\hat{\mathbf{V}}_{1s}$ and $\hat{\mathbf{V}}_{2s}$ are only 3–7 points higher than the corresponding maximum likelihood solutions.

Approximate methods, such as the exploratory factor analysis of the pooled set of observations (ignoring the clustering of students within classes), and the factor analysis of the within- and between-group statistics $\mathbf{T}_1$ and $\mathbf{T}_2$ succeed in recovering the total score as the first factor. The within-group means and the pooled-data analyses produce only one set of factors, and neither of them agrees with either second factor of the maximum likelihood solution. The factor analyses of the within- and between-group statistics $\mathbf{T}_1$ and $\mathbf{T}_2$ appear to be satisfactory, though.

7. Conclusion

Factor analysis of multilevel data can be motivated in complete analogy with classical factor analysis. Importance of the maximum likelihood estimation has been discussed by McDonald and Goldstein (1989) and Muthén (1989). The algorithm presented in this paper is computationally feasible for datasets with several components and models with a moderate number of factors; the limitation is given by the number of estimated parameters, but even that can be overcome by application of quasi-Newton methods. Analysis of two examples indicates that a noniterative algorithm based on the method of moments provides a good approximation to the maximum likelihood solution when no cross-level constraints are used.

References

Aitkin, M., & Longford, N. T. (1986). Statistical modeling issues in school effectiveness studies. *Journal of the Royal Statistical Society, Series A 149*, 419–461.

Cronbach, L. J. (1976). *Research on classrooms and schools: Formulation of questions, design, and analysis*. Unpublished manuscript, Stanford University, Stanford Evaluation Consortium, School of Education.

Crosswhite, F. J., Dossey, J. A., Swafford, J. O., McKnight, C. C., & Cooney, T. J. (1985). *Second International Mathematics Study: Summary report for the United States*. Champaign, IL: Stipes.

de Leeuw, J., & Kreft, I. (1986). Random coefficient models for multilevel analysis. *Journal of Educational Statistics, 11*, 57–85.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika, 73*, 43–56.

Goldstein, H. (1987). *Multilevel models in educational and social research*. London, U.K.: C. Griffin & Co.

Goldstein, H., & McDonald, R. P. (1988). A general model for the analysis of multilevel data. *Psychometrika, 53*, 455–467.

Härnquist, K. (1978). Primary mental abilities of collective and individual levels. *Journal of Educational Psychology, 70*, 706–716.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika, 61*, 383–385.

Jamshidian, M., & Jennrich, R. I. (1988). *Conjugate gradient methods in confirmatory factor analysis* (UCLA Statistics Series 8). Los Angeles: UCLA.

Jennrich, R. I., & Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics, 42*, 805–820.

Jöreskog, K. G. (1977). Factor analysis by least-squares and maximum-likelihood methods. In K. Enstein, A. Ralston, & H. S. Wilf (Eds.), *Statistical methods for digital computers* (pp. 125–153). New York: John Wiley & Sons.

Jöreskog, K. G., & Sörbom, D. (1979). *Advances in factor analysis and structural equation models*. Cambridge, MA: Abt Books.

Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method* (2nd ed.). London: Butterworth & Co.

Lee, S. Y. (1990). Multilevel analysis of structural equation models. *Biometrika, 772*, 763–772.

Lee, S. Y., & Jennrich, R. I. (1979). A study of algorithms for covariance structure analysis with specific comparisons using factor analysis. *Psychometrika, 43*, 99–113.

Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika, 74*, 817–827.

Longford, N. T. (1988). *VARCL. Software for variance component analysis of data with hierarchically nested random effects (maximum likelihood)*. Princeton, NJ: Educational Testing Service.

Longford, N. T. (1990). Multivariate variance component analysis: An application in test development. *Journal of Educational Statistics, 15*, 91–112.

Luenberger, D. G. (1984). *Linear and nonlinear programming* (2nd ed.). Reading, MA: Addison-Wesley.

Mason, W. M., Wong, G. Y., & Entwisle, B. (1984). Contextual analysis through the multilevel linear model. In S. Leinhardt (Ed.), *Sociological methodology (1983–84)* (pp. 72–103). San Francisco: Jossey Bass.

McDonald, R. P., & Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data. *British Journal Mathematical and Statistical Psychology, 42*, 215–232.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*, 557–585.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28*, 338–354.

Muthén, B. O. (in press). Mean and covariance structure analysis of hierarchical data. *Journal of Educational Statistics*.

Muthén, B. O., & Satorra, A. (1989). Multilevel aspects of varying parameters in structural models. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 87–99). New York: Academic Press.

Patterson, H. D., & Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika, 58*, 545–554.

Raudenbush, S. W., & Bryk, A. S. (1985). A hierarchical model for studying school effects. *Sociology of Education, 59*, 1–17.

Manuscript received 11/1/90
Final version received 1/14/92