Scand J Statist 8: 65–92, 1981

Analysis of Covariance Structures[1]

KARL G. JÖRESKOG

University of Uppsala

ABSTRACT. Most of the important models and techniques for analysis of covariance structures are reviewed and illustrated with a few simple examples. Some general approaches to the problems of identification, estimation and testing of covariance structures are considered. The ACOVS-model and the LISREL-model are treated in some detail. Specific types of covariance structures described are variance and covariance components models, measurement models, path analysis models and simplex models. Simultaneous analysis of mean and covariance structures is also considered as well as the simultaneous analysis of data from several populations.

Key words: covariance structure analysis, factor analysis, variance components, path analysis, structural equation models, autoregressive models

1. Introduction

Analysis of covariance structures is the common term for a number of techniques for analyzing multivariate data in order to detect and assess latent (unobserved) sources of variation and covariation in the observed measurements. The techniques of covariance structure analysis are general and flexible in that they can handle many types of covariance structures useful especially in the behavioral and social sciences. Although these techniques can be used for exploratory analysis they have been most successfully applied to confirmatory analysis where the type of covariance structure is specified in advance. A covariance structure of a specified kind may arise because of a specified substantive theory or hypothesis, a given classificatory design for the measures, known experimental conditions or because of results from previous studies based on extensive data. Sometimes the observed variables are ordered through time, as in longitudinal studies, or according to linear or circular patterns, as in Guttman's (1954) simplex and circumplex models or according to a given causal scheme, as in path analysis.

This paper reviews most of the important models and techniques for analysis of covariance structures and illustrates them with a few simple examples. The examples are introduced in section 2 together with the main types of covariance structures. General covariance and correlation structures are defined in section 3. Special cases of general covariance structures are: the ACOVS-model (Jöreskog, 1970*a*, 1973*a*, 1974) and the LISREL-model (Jöreskog, 1973*b*, 1977; Jöreskog & Sörbom, 1978) and these are briefly described also. Approaches to the statistical problems of identification, estimation and testing are also considered in section 3. The analysis of the examples are continued in section 4. Section 5 discusses generalizations which permit simultaneous analysis of mean and covariance structures and simultaneous analysis of data from several populations.

This paper draws on various material published previously by the author, in particular Jöreskog (1973*a*, 1974, 1978). Other important related material is contained in Bock & Bargmann (1966), Browne (1974, 1977), McDonald (1974, 1975, 1978) and Bentler & Weeks (1981).

2. Some types of covariance structures

2.1. *Variance and covariance components*

Several authors (Bock, 1960; Bock & Bargmann, 1966; Wiley et al., 1973) have considered covariance structure analysis as an approach to study differences in test performance when the tests have been constructed by assigning items or subtests according to objective features of content or format to subclasses of a factorial or hierarchical classification.

Bock (1960) suggested that the scores of N subjects on a set of tests classified in a 2^n factorial design may be viewed as data from an $N \times 2^n$ experimental design, where the subjects represent a random mode

[1] The contents of this paper were presented in lectures held at the 8th Nordic Conference on Mathematical Statistics, Mariehamn, Finland, May 1980. The lectures were concluded with the discussion which is reported at the end of the paper.

Table 1. *Inter trial covariance matrix*

	1	2	3	4	5	6	7	8	9	10	11	12
1	51.6											
2	−27.7	72.1										
3	38.9	−41.1	69.9									
4	−36.4	40.7	−39.1	75.8								
5	13.8	−5.2	17.9	1.9	84.8							
6	−13.6	10.9	9.5	17.8	−37.4	91.1						
7	21.5	−9.4	8.5	−13.1	59.7	−54.4	79.9					
8	−12.8	17.2	−3.1	22.0	−43.3	52.7	−49.9	87.2				
9	11.0	−8.9	19.2	−11.2	−12.6	21.9	−10.6	17.5	27.6			
10	−4.5	10.2	−7.6	12.7	20.4	−11.5	16.5	−14.8	−8.8	19.9		
11	9.2	−0.3	18.9	−13.6	−3.9	19.0	−8.3	13.1	17.7	−2.8	27.3	
12	−3.7	7.5	−4.5	12.8	19.9	−8.8	15.5	−8.6	−5.4	13.3	−1.0	16.0

of classification and the tests represent n fixed modes of classification. Bock pointed out that conventional mixed-model analysis of variance gives useful information about the psychometric properties of the tests. In particular, the presence of non-zero variance components for the random mode of classification and for the interaction of the random and fixed modes of classification provides information about the number of dimensions in which the tests are able to discriminate among subjects. The relative size of these components measure the power of the tests to discriminate among subjects along the respective dimensions.

The following example was given in an unpublished paper by Browne (1970).

Example 1: The Rod and Frame (RF) tcst is used as a measure of field dependence. A subject is seated in a darkened room on a chair which may be tilted to the left or to the right. In front of him is a luminous rod located in a luminous square frame. The chair, frame, and rod are tilted to prespecified positions. By operating push buttons connected to an electric motor the subject is to move the rod to the vertical position. The score on the trial is the angle of the rod from the vertical. This can assume positive and negative values. Each subject undergoes 12 trials. The last two columns of the design matrix **A** *below give initial positions of the frame and chair for each trial. A value of +1 denotes that the position of the frame or chair was at +28° from the vertical, a value of −1 denotes that the angle was −28° and a value of 0 denotes that the initial position was vertical.*

Table 1 shows a covariance matrix between trials of the RF-test obtained from a sample of 107 eighteen year old males.

One would like to estimate the variance components associated with the general bias, frame effect, chair effect and error.

$$\mathbf{A}=\begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \\ 1 & 1 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \\ 1 & 1 & 0 \\ 1 & -1 & 0 \\ 1 & 1 & 0 \\ 1 & -1 & 0 \end{bmatrix}$$

Let a, b and c be uncorrelated random components associated with general bias, frame effect and chair effect, respectively, and let e denote an error component uncorrelated with a, b and c and uncorrelated over trials. Let

$$\mathbf{u}_\nu' = (a_\nu, b_\nu, c_\nu)$$

be the values of a, b and c for subject ν. Then the scores on the twelve trials for subject ν is

$$\mathbf{x}_\nu = \mathbf{A}\mathbf{u}_\nu + \mathbf{e}_\nu$$

with covariance matrix

$$\mathbf{\Sigma} = \mathbf{A}\mathbf{\Phi}\mathbf{A}' + \sigma_e^2\mathbf{I}, \tag{1}$$

where

$$\mathbf{\Phi} \simeq \text{diag}\,(\sigma_a^2, \sigma_b^2, \sigma_c^2)$$

Equation (1) shows that all the 78 variances and covariances in $\mathbf{\Sigma}$ are *linear* functions of the four parameters σ_a^2, σ_b^2, σ_c^2 and σ_e^2. To see this explicitly, consider the covariance matrix generated by trials 1, 2, 5 and 9:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_a^2+\sigma_b^2+\sigma_c^2+\sigma_e^2 & & & \\ \sigma_a^2-\sigma_b^2-\sigma_c^2 & \sigma_a^2+\sigma_b^2+\sigma_c^2+\sigma_e^2 & & \\ \sigma_a^2-\sigma_b^2+\sigma_c^2 & \sigma_a^2+\sigma_b^2-\sigma_c^2 & \sigma_a^2+\sigma_b^2+\sigma_c^2+\sigma_e^2 & \\ \sigma_a^2+\sigma_b^2 & \sigma_a^2-\sigma_b^2 & \sigma_a^2-\sigma_b^2 & \sigma_a^2+\sigma_b^2+\sigma_e^2 \end{bmatrix}$$

This is an example of a *linear covariance structure*. If this structure holds, the four parameters can be solved in terms of the elements of $\boldsymbol{\Sigma}$. For example, $\sigma_a^2 = \frac{1}{2}(\sigma_{41}+\sigma_{42})$, $\sigma_b^2 = \frac{1}{2}(\sigma_{41}+\sigma_{42}-\sigma_{21}-\sigma_{31})$, $\sigma_c^2 = \sigma_{31}-\sigma_{42}$, etc. There are many ways in which the four parameters can be solved in terms of the σ's. If the ten equations are consistent, however, all solutions are identical. In this case the parameters are *over-identified.*

Consider the covariance structure generated by the first four rows of **A**:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_a^2+\sigma_b^2+\sigma_c^2+\sigma_e^2 & & & \\ \sigma_a^2-\sigma_b^2-\sigma_c^2 & \sigma_a^2+\sigma_b^2+\sigma_c^2+\sigma_e^2 & & \\ \sigma_a^2+\sigma_b^2+\sigma_c^2 & \sigma_a^2-\sigma_b^2-\sigma_c^2 & \sigma_a^2+\sigma_b^2+\sigma_c^2+\sigma_e^2 & \\ \sigma_a^2-\sigma_b^2-\sigma_c^2 & \sigma_a^2+\sigma_b^2+\sigma_c^2 & \sigma_a^2-\sigma_b^2-\sigma_c^2 & \sigma_a^2+\sigma_b^2+\sigma_c^2+\sigma_e^2 \end{bmatrix}$$

In this case we can solve for $\sigma_a^2 = \frac{1}{2}(\sigma_{21}+\sigma_{31})$, say, but it is impossible to solve for σ_b^2 and σ_c^2 separately. Only the sum $\sigma_b^2+\sigma_c^2$ is identified. This is an example of a *non-identified model* in which some parameters are underidentified and others are not. The reason for this is that the matrix **A** has rank 2 and not rank 3 as in the previous case. The example will be continued in section 4.

In general, if **A** is of order $p \times r$ and of rank k, one may choose k independent linear functions, each one linearly dependent on the rows of **A** and estimate the covariance matrix of these functions. It is customary to choose linear combinations that are mutually uncorrelated but this is not necessary. Let **L** be the matrix of coefficients of the chosen linear functions and let **K** be any matrix such that $\mathbf{A} = \mathbf{KL}$. For example, **K** may be obtained from

$$\mathbf{K} = \mathbf{AL}'(\mathbf{LL}')^{-1}.$$

The model may then be reparameterized to full rank by defining $\mathbf{u}^* = \mathbf{Lu}$. We then have $\mathbf{x} = \mathbf{Au}+\mathbf{e} = \mathbf{KLu}+\mathbf{e} = \mathbf{Ku}^*+\mathbf{e}$. The covariance matrix of **x** is represented as

$$\boldsymbol{\Sigma} = \mathbf{K}\boldsymbol{\Phi}^*\mathbf{K}' + \boldsymbol{\Psi}$$

where $\boldsymbol{\Phi}^*$, the covariance matrix of $\mathbf{u}^*$ is not necessarily diagonal and $\boldsymbol{\Psi}$ is the diagonal covariance matrix of **e**. The latter may be taken to be homogenous, if desired.

The above model assumes that all measurements are on the same scale. Wiley, Schmidt and Bramble (1973) suggested the study of a general class of components of covariance models which would allow different variables to be on different scales. The covariance matrix $\boldsymbol{\Sigma}$ will then be of the form

$$\boldsymbol{\Sigma} = \boldsymbol{\Delta}\mathbf{A}\boldsymbol{\Phi}\mathbf{A}'\boldsymbol{\Delta}+\boldsymbol{\Theta} \quad \text{or} \quad \boldsymbol{\Sigma} = \boldsymbol{\Delta}(\mathbf{A}\boldsymbol{\Phi}\mathbf{A}'+\boldsymbol{\Psi})\boldsymbol{\Delta} \tag{2a–b}$$

The matrix $\mathbf{A}(p \times k)$ is assumed to be known and gives the coefficient of the linear functions connecting the manifest and latent variables, $\boldsymbol{\Delta}$ is a $p \times p$ diagonal matrix of unknown scale factors, $\boldsymbol{\Phi}$ is the $k \times k$ symmetric and positive definite covariance matrix of the latent variables and $\boldsymbol{\Psi}$ and $\boldsymbol{\Theta}$ are $p \times p$ diagonal matrices of error variances.

Within this class of models eight different special cases are of interest. These are generated by the combination of the following set of conditions:

$$\text{on } \boldsymbol{\Delta}: \begin{cases} \boldsymbol{\Delta} = \mathbf{I} \\ \boldsymbol{\Delta} \neq \mathbf{I} \end{cases}$$

$$\text{on } \boldsymbol{\Phi}: \begin{cases} \boldsymbol{\Phi} \text{ is diagonal} \\ \boldsymbol{\Phi} \text{ is not diagonal} \end{cases}$$

$$\text{on } \boldsymbol{\Psi} \text{ or } \boldsymbol{\Theta}: \begin{cases} \boldsymbol{\Psi} \text{ or } \boldsymbol{\Theta} = \sigma^2\,\mathbf{I} \\ \boldsymbol{\Psi} \text{ or } \boldsymbol{\Theta} \text{ general diagonal.} \end{cases}$$

The classical formulation of the mixed model and its generalizations assume that $\boldsymbol{\Delta} = \mathbf{I}$. This is appropriate if the observed variables are in the same metric as for example when the observed variables represent physical measurements, time to criterion measures, reaction times or items similarly scaled

such as semantic differential responses. However, if the observed variables are measured in different metrics then the classical model would not fit. In such cases the inclusion of Δ in the model as a general diagonal matrix of scaling factors would provide a useful alternative specification. It should be pointed out that the elements of Δ do not have to be related to the variances of the variables.

The classical components of variance model assume that Φ is diagonal. However, there are usually no substantive reasons for assuming this.

The two conditions on Ψ or Θ correspond to homogenous and heterogeneous error variances. If the variables are in the same metric and if the measurement situation is sufficiently similar from variable to variable then it would seem reasonable to hypothesize that the variances of the errors of measurement ought to be homogeneous, i.e., in (2a) we take $\Delta = \mathbf{I}$ and $\Theta = \sigma^2\mathbf{I}$.

If, on the other hand, the scale of measurement is the same but the meaurement situation from variable to variable is different enough to generate different kinds of error structures, then the variances of the errors of measurement might differ systematically from variable to variable. For this situation it would seem best to take $\Delta = \mathbf{I}$ but leave Θ free in (2a). If the manifest fariables were in different metrics then clearly the error variances in the observed metric will most likely be heterogeneous. One useful hypothesis to test in this context would be that the standard deviations of the errors of measurement are proportional to the rescaling factors. This would correspond to taking $\Psi = \sigma^2\mathbf{I}$ in (2b). When both Δ and Ψ are free, (2a) and (2b) are equivalent.

2.2. *Measurement models*

Most measurements employed in the behavioral and social sciences contain sizeable errors of measurements and any adequate theory or model must take this fact into account. Of particular importance is the study of congeneric measurements, i.e., those measurements that are assumed to measure the same thing.

Classical test theory (Lord & Novick, 1968) assumes that a test score x is the sum of a true score τ and an error score e, where e and τ are uncorrelated. A set of test scores $x_1, ..., x_p$ with true scores $\tau_1, ..., \tau_p$ is said to be congeneric if every pair of true scores τ_i and τ_j have unit correlation. Such a set of test scores can be represented as

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\beta}\tau + \mathbf{e},$$

where $\mathbf{x}' = (x_1, ..., x_p)$, $\boldsymbol{\beta}' = (\beta_1, ..., \beta_p)$ is a vector of regression coefficients, $\mathbf{e}' = (e_1, ..., e_p)$ is the vector of error scores, $\boldsymbol{\mu}$ is the mean vector of $\mathbf{x}$ and τ is a true score, for convenience scaled to zero mean and unit variance. The elements of $\mathbf{x}$, $\mathbf{e}$ and τ are regarded as random variables for a population of examinees. Let $\theta_1, ..., \theta_p$ be the variances of $e_1, ..., e_p$, respectively, i.e., the error variances. The corresponding true score variances are $\beta_1^2, ..., \beta_p^2$. One important problem is that of estimating these quantities. The covariance matrix of $\mathbf{x}$ is

$$\boldsymbol{\Sigma} = \boldsymbol{\beta}\boldsymbol{\beta}' + \boldsymbol{\Theta}, \tag{3}$$

where $\boldsymbol{\Theta} = \text{diag}\,(\theta_1, ..., \theta_p)$.

Parallel tests and tau-equivalent tests, in the sense of Lord & Novick (1968), are special cases of congeneric tests. Parallel tests have equal true score variances and equal error variances, i.e.,

$$\beta_1^2 = ... = \beta_p^2, \quad \theta_1 = ... = \theta_p.$$

Tau-equivalent tests have equal true score variances but possibly different error variances.

Parallel and tau-equivalent tests are homogenous in the sense that all covariances between pairs of test scores are equal. For parallel tests the variances are also equal. Scores on such tests are directly comparable, i.e., they represent measurements on the same scale. For tests composed of binary items this can hold only if the tests have the same number of items and are administered under the same time limits. Congeneric tests, on the other hand, need not satisfy such strong restrictions. They need not even be tests consisting of items but can be ratings, for example, or even measurements produced by different measuring instruments.

The previous model generalizes immediately to several sets of congeneric test scores. If there are q sets of such tests, with $m_1, m_2, ..., m_q$ tests respectively, we write $\mathbf{x}' = (\mathbf{x}_1', \mathbf{x}_2', ..., \mathbf{x}_q')$ where $\mathbf{x}_g'$, $g = 1, 2, ..., q$ is the vector of observed scores for the gth set. Associated with the vector $\mathbf{x}_g$ there is a true score τ_g and vectors $\boldsymbol{\mu}_g$ and $\boldsymbol{\beta}_g$ defined as before so that

$$\mathbf{x}_g = \boldsymbol{\mu}_g + \boldsymbol{\beta}_g\tau_g + \mathbf{e}_g.$$

As before we may, without loss of generality, assume that τ_g is scaled to zero mean and unit variance. If the different true scores $\tau_1, \tau_2, ..., \tau_q$ are all mutually uncorrelated, then each set of tests can be analyzed separately. However, in most cases these true scores correlate with each other and an overall analysis of the entire set of tests must be made. Let $p = m_1 + m_2 + ... m_q$ be the total number of tests. Then $\mathbf{x}$ is of order p. Let $\boldsymbol{\mu}$ be the mean vector of $\mathbf{x}$, and let $\mathbf{e}$ be the vector of error scores. Furthermore, let

$$\boldsymbol{\tau}' = (\tau_1, \tau_2, ..., \tau_q)$$

Table 2. *Lord's vocabulary test data covariance matrix* ($N = 649$)

	x_1	x_2	y_1	y_2
x_1	86.3979			
x_2	57.7751	86.2632		
y_1	56.8651	59.3177	97.2850	
y_2	58.8986	59.6683	73.8201	97.8192

and let $\mathbf{B}$ be the matrix of order $p \times q$, paritioned as

$$\mathbf{B} = \begin{bmatrix} \beta_1 & 0 & \dots & 0 \\ 0 & \beta_2 & \dots & 0 \\ . & . & \dots & . \\ . & . & \dots & . \\ 0 & 0 & & \beta_q \end{bmatrix}.$$

Then $\mathbf{x}$ is represented as

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{B}\boldsymbol{\tau} + \mathbf{e}.$$

Let $\boldsymbol{\Gamma}$ be the correlation matrix of $\boldsymbol{\tau}$. Then the covariance matrix $\boldsymbol{\Sigma}$ of $\mathbf{x}$ is

$$\boldsymbol{\Sigma} = \mathbf{B}\boldsymbol{\Gamma}\mathbf{B}' + \boldsymbol{\Theta}, \tag{4}$$

where $\boldsymbol{\Theta}$ is a diagonal matrix of order p containing the error variances.

The correlation coefficient corrected for attenuation between two tests x and y is the correlation between their true scores. If, on the basis of a sample of examinees, the disattenuated correlation is near unity, the experimenter concludes that the two tests are measuring the same trait.

The following example is based on some data from Lord (1957).

Example 2: Two tests x_1 and x_2 are 15-item vocabulary tests administered under liberal time limits. Two other tests y_1 and y_2 are highly speeded 75-item vocabulary tests. The covariance matrix is given in Table 2. One would like to estimate the disattenuated correlation between x and y and to test whether this is one. One would also like to test whether the two pairs of tests are parallel.

We set up the following measurement model

$$\begin{pmatrix} x_1 \\ x_2 \\ y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \beta_1 & 0 \\ \beta_2 & 0 \\ 0 & \beta_3 \\ 0 & \beta_4 \end{pmatrix} \begin{pmatrix} \tau_x \\ \tau_y \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix}$$

with covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \beta_1 & 0 \\ \beta_2 & 0 \\ 0 & \beta_3 \\ 0 & \beta_4 \end{pmatrix} \begin{pmatrix} 1 & \varrho \\ \varrho & 1 \end{pmatrix} \begin{pmatrix} \beta_1 & \beta_2 & 0 & 0 \\ 0 & 0 & \beta_3 & \beta_4 \end{pmatrix} + \begin{pmatrix} \theta_1 & 0 & 0 & 0 \\ 0 & \theta_2 & 0 & 0 \\ 0 & 0 & \theta_3 & 0 \\ 0 & 0 & 0 & \theta_4 \end{pmatrix}$$

$$= \begin{bmatrix} \beta_1^2+\theta_1 & & & \\ \beta_1\beta_2 & \beta_2^2+\theta_2 & & \\ \beta_1\beta_3\varrho & \beta_2\beta_3\varrho & \beta_3^2+\theta_3 & \\ \beta_1\beta_4\varrho & \beta_2\beta_4\varrho & \beta_3\beta_4 & \beta_4^2+\theta_4 \end{bmatrix} \tag{5}$$

This is an example of a *non-linear covariance structure* in which the ten variances and covariances of the observed variables are non-linear functions of nine parameters. Each of these nine parameters are identified in terms of the σ's, except possibly for the sign in some of them, as can easily be verified. For example, $\beta_1^2 = (\sigma_{21}\sigma_{31}/\sigma_{32})$, $\varrho^2 = (\sigma_{31}\sigma_{42}/\sigma_{21}\sigma_{43})$, $\theta_1 = \sigma_{11} - \beta_1^2$, etc.

In this model, x_1 and x_2 are congeneric measures of τ_x and y_1 and y_2 are congeneric measures of τ_y. The disattenuated correlation ϱ is the correlation between τ_x and τ_y. To analyze the data one can set up the four hypotheses:

H_1: $\beta_1 = \beta_2$, $\beta_3 = \beta_4$, $\theta_1 = \theta_2$, $\theta_3 = \theta_4$, $\varrho = 1$

H_2: $\beta_1 = \beta_2$, $\beta_3 = \beta_4$, $\theta_1 = \theta_2$, $\theta_3 = \theta_4$

H_3: $\sigma = 1$

H_4: $\beta_1, \beta_2, \beta_3, \beta_4, \theta_1, \theta_2, \theta_3, \theta_4$, and ϱ unconstrained.

and estimate the model under each of these. Under hypotheses H_1, H_2 and H_3 the model involves *equality constraints* imposed on the parameters of the base model H_4. The analysis of the example will be continued in section 4.

2.3. *Path analysis models*

Path analysis, due to Wright (1934), is a technique to assess the direct causal contribution of one variable to another in non-experimental investigations. The problem in general is that of estimating the parameters of a set of linear structural equations representing the cause and effect relationships hypothesized by the investigator. Traditionally the vari-

ables in the structural equation system were directly observed variables but recently several models have been studied which involve hypothetical constructs, i.e., latent variables, which, while not directly observed, have operational implications for relationships among observable variables (see, e.g., Werts & Linn, 1970; Hauser & Goldberger, 1971; Jöreskog & Goldberger, 1975). In some models, the observed variables appear only as effects (indicators) of the hypothetical constructs, while in others, the observed variables appear as causes or as both causes and effects of latent variables.

Suppose that two variables are used on two occasions, i.e., in a two-wave longitudinal design. Assume that the two variables measure the same latent variable η on two different occasions, i.e., y_1 and y_2 measure η_1 on the first occasion and y_3 and y_4 measure η_2 on the second occasion.

The equations defining the measurement relations are

$$y_1 = \eta_1 + \varepsilon_1$$

$$y_2 = \lambda_1 \eta_1 + \varepsilon_2$$

$$y_3 = \eta_2 + \varepsilon_3$$

$$y_4 = \lambda_2 \eta_2 + \varepsilon_4$$

The main interest is in the stability of η over time. This can be studied by means of the structural equation

$$\eta_2 = \beta \eta_1 + \zeta,$$

the regression of η_2 on η_1. In particular, one is interested in whether β is close to one and ζ is small.

Let $\boldsymbol{\Omega}$ be the covariance matrix of (η_1, η_2) and let $\boldsymbol{\Theta}$ be the covariance matrix of $(\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4)$. If all the ε's are uncorrelated so that $\boldsymbol{\Theta}$ is diagonal, the covariance matrix of (y_1, y_2, y_3, y_4) is

$$\boldsymbol{\Sigma} = \begin{bmatrix} \omega_{11}+\theta_{11} & & & \\ \lambda_1\omega_{11} & \lambda_1^2\omega_{11}+\theta_{22} & & \\ \omega_{21} & \lambda_1\omega_{21} & \omega_{22}+\theta_{33} & \\ \lambda_2\omega_{21} & \lambda_1\lambda_2\omega_{21} & \lambda_2\omega_{22} & \lambda_2^2\omega_{22}+\theta_{44} \end{bmatrix}$$

The matrix $\boldsymbol{\Sigma}$ has ten variances and covariances which are functions of nine parameters. The model is in fact a reparameterization of that discussed in the previous subsection.

Often when the same variables are used repeatedly there is a tendency for the corresponding errors (the ε's) to correlate over time because of memory and other retest effects. Hence there is a need to generalize the above model to allow for correlations between ε_1 and ε_3 and also between ε_2 and ε_4. This means that there will be two non-zero covariances θ_{31} and θ_{42} in $\boldsymbol{\Theta}$. The covariance matrix of the observed variables will now be

$$\boldsymbol{\Sigma} = \begin{bmatrix} \omega_{11}+\theta_{11} & & & \\ \lambda_1\omega_{11} & \lambda_1^2\omega_{11}+\theta_{22} & & \\ \omega_{21}+\theta_{31} & \lambda_1\omega_{21} & \omega_{22}+\theta_{33} & \\ \lambda_2\omega_{21} & \lambda_1\lambda_2\omega_{21}+\theta_{42} & \lambda_2\omega_{22} & \lambda_2^2\omega_{22}+\theta_{44} \end{bmatrix} \tag{6}$$

This $\boldsymbol{\Sigma}$ has its ten independent elements expressed in terms of eleven parameters. Hence it is clear that the model is not identified. In fact, none of the eleven parameters are identified without further conditions imposed. The loadings λ_1 and λ_2 may be multiplied by a constant and the ω's divided by the same constant. This does not change σ_{21}, σ_{32}, σ_{41} and σ_{43}. The change in the other σ's may be compensated by adjusting the θ's additively. Hence to make the model identified one must fix one λ or one ω at a non-zero value or one θ at some arbitrary value. However, the *correlation* between η_1 and η_2 is identified without any restrictions, since

$$\text{Corr}\,(\eta_1, \eta_2) = (\omega_{21}^2/\omega_{11}\omega_{22})^{\frac{1}{2}} = [(\sigma_{32}\sigma_{41})/(\sigma_{21}\sigma_{43})]^{\frac{1}{2}}.$$

This model may therefore be used to estimate this correlation coefficient and to test whether this is one. To make further use of the model it is necessary to make some assumption about the nature of the variables. For example, if it can be assumed that the two variables on each occasion are tauequivalent we can set both λ_1 and λ_2 equal to one. Then the model can be estimated and tested with one degree of freedom. If $\lambda_1 = \lambda_2$ the model is just identified.

While the above model is not identified as it stands it becomes so as soon as there is information about one or more background variables affecting η_1 or η_2 or both. To illustrate this an example of a longitudinal study analyzed in more detail by Wheaton et al. (1977) will be used.

Example 3: This study was concerned with the stability over time of attitudes such as alienation and the relation to background variables such as education and occupation. Data on attitude scales were collected from 932 persons in two rural regions in Illinois at three points in time: 1966, 1967 and 1971. The variables used for the present illustration are the Anomia *subscale and the* Powerlessness *subscale, taken to be indicators of* Alienation. *This example uses data from 1967 and 1971 only. The background variables are*

the respondent's education (years of schooling completed) and Duncan's Socioeconomic Index (SEI). These are taken to be indicators of the respondent's socioeconomic status (SES). The sample covariance matrix of the six observed variables is given in Table 3.

Let

y_1 = Anomia 67
y_2 = Powerlessness 67
x_1 = Education
ξ = SES
y_3 = Anomia 71
y_4 = Powerlessness 71
x_2 = SEI
η_1 = Alienation 67
η_2 = Alienation 71

The model is then specified as

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \lambda_1 & 0 \\ 0 & 1 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{pmatrix},$$

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ \lambda_3 \end{pmatrix} \zeta + \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix},$$

$$\begin{pmatrix} 1 & 0 \\ -\beta & 1 \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} \zeta + \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix}, \qquad (7)$$

It is assumed that ζ_1 and ζ_2 are uncorrelated and that the scales for η_1, η_2 and ξ have been chosen to be the same as for y_1, y_3 and x_1, respectively.

Let $\phi = \mathrm{Var}(\xi)$ and $\psi_i = \mathrm{Var}(\zeta_i)$, $i = 1, 2$ and let $\mathbf{\Omega}$ be the covariance matrix of (η_1, η_2, ξ). It is obvious that there is a one-to-one correspondence between the six ω's in $\mathbf{\Omega}$ and $(\phi, \beta, \gamma_1, \gamma_2, \psi_1, \psi_2)$. In terms of $\mathbf{\Omega}$ the covariance matrix of $(y_1, y_2, y_3, y_4, x_1, x_2)$ is

$$\begin{bmatrix} \omega_{11}+\theta_{11} & & & & & \\ \lambda_1\omega_{11} & \lambda_1^2\omega_{11}+\theta_{22} & & & & \\ \omega_{21}+\theta_{31} & \lambda_1\omega_{21} & \omega_{22}+\theta_{33} & & & \\ \lambda_2\omega_{21} & \lambda_1\lambda_2\omega_{21}+\theta_{42} & \lambda_2\omega_{22} & \lambda_2^2\omega_{22}+\theta_{44} & & \\ \omega_{31} & \lambda_1\omega_{31} & \omega_{32} & \lambda_2\omega_{32} & \omega_{33}+\theta_{55} & \\ \lambda_3\omega_{31} & \lambda_1\lambda_3\omega_{31} & \lambda_3\omega_{32} & \lambda_2\lambda_3\omega_{32} & \lambda_3\omega_{33} & \lambda_3^2\omega_{33}+\theta_{66} \end{bmatrix} \qquad (8)$$

Table 3. *Covariance matrix for variables in the stability of alienation example*

	y_1	y_2	y_3	y_4	x_1	x_2
y_1	11.834					
y_2	6.947	9.364				
y_3	6.819	5.091	12.532			
y_4	4.783	5.028	7.495	9.986		
x_1	−3.839	−3.889	−3.841	−3.625	9.610	
x_2	−21.899	−18.831	−21.748	−18.775	35.522	450.288

The upper left 4×4 part is the same as (6). It is clear that the two last rows of $\mathbf{\Sigma}$ determines λ_1, λ_2, λ_3, ω_{31}, ω_{32}, ω_{33}, θ_{55} and θ_{66}. With λ_1 and λ_2 determined, the other parameters are determined by the upper left part. Altogether there are seventeen parameters to estimate. The example is continued in section 4.

2.4. *Simplex models*

Simplex models is a type of covariance structure which often occurs in longitudinal studies when the same variable is measured repeatedly on the same people over several occasions. The simplex model is equivalent to the covariance structure generated by a first-order non-stationary autoregressive process. Guttman (1954) used the term simplex also for variables which are not ordered through time but by other criteria. One of his examples concerns tests of verbal ability ordered according to increasing complexity. The typical feature of a simplex correlation structure is that the correlations decrease as one moves away from the main diagonal.

Jöreskog (1970*b*) formulated various simplex models in terms of the well-known Wiener and Markov stochastic processes. A distinction was made between a perfect simplex and a quasi-simplex.

A perfect simplex is reasonable only if the measurement errors in the test scores are negligible. A quasi-simplex, on the other hand, allows for sizeable errors of measurement.

Consider p fallible variables $y_1, y_2, \ldots, y_p$. The unit of measurement in the true variables η_i may be chosen to be the same as in the observed variables y_i. The equations defining the model are then, taking all variables as deviations from their means,

$$y_i = \eta_i + \varepsilon_i, \quad i = 1, 2, \ldots, p,$$
$$\eta_i = \beta_i \eta_{i-1} + \zeta_i, \quad i = 2, 3, \ldots, p,$$

Table 4. *Correlations among grade point averages, high school rank and an aptitude test*

	y_0	y_0'	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
y_0	1.000									
y_0'	.393	1.000								
y_1	.387	.375	1.000							
y_2	.341	.298	.556	1.000						
y_3	.278	.237	.456	.490	1.000					
y_4	.270	.255	.439	.445	.562	1.000				
y_5	.240	.238	.415	.418	.496	.512	1.000			
y_6	.256	.252	.399	.383	.456	.469	.551	1.000		
y_7	.240	.219	.387	.364	.445	.442	.500	.544	1.000	
y_8	.222	.173	.342	.339	.354	.416	.453	.482	.541	1.000

Note: y_0 is high school rank, y_0' ACT composite score, and y_1 through y_8 are eight semesters grade-point averages.

where the ε_i are uncorrelated among themselves and uncorrelated with all the η_i and where ζ_{i+1} is uncorrelated with η_i, $i=1, 2, ..., p-1$. The parameters of the model are $\omega_i = \text{Var}(\eta_i)$, $\theta_i = \text{Var}(\varepsilon_i)$, $i=1, 2, ..., p$ and $\beta_2, \beta_3, ..., \beta_p$. The residual variance $\text{Var}(\zeta_{i+1})$ is a function of ω_{i+1}, ω_i and β_{i+1}, namely $\text{Var}(\zeta_{i+1}) = \omega_{i+1} - \beta_{i+1}^2 \omega_i$, $i=1, 2, ..., p-1$. The covariance matrix of $y_1, y_2, ..., y_p$ is of the form, here illustrated with $p=4$.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \omega_1+\theta_1 & & & \\ \beta_2\omega_1 & \omega_2+\theta_2 & & \\ \beta_2\beta_3\omega_1 & \beta_3\omega_2 & \omega_3+\theta_3 & \\ \beta_2\beta_3\beta_4\omega_1 & \beta_3\beta_4\omega_2 & \beta_4\omega_3 & \omega_4+\theta_4 \end{bmatrix}, \quad (9)$$

Consider first the perfect simplex, i.e., the case $\theta_i = 0$, $i=1, 2, ..., p$. Then

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \alpha_1 & & & \\ \gamma_2 & \alpha_2 & & \\ 0 & \lambda_3 & \alpha_3 & \\ 0 & 0 & \lambda_4 & \alpha_4 \end{bmatrix}$$

is tri-diagonal and it may be readily verified that there is a one-to-one correspondence between the parameters $(\omega_1, \omega_2, \omega_3, \omega_4, \beta_2, \beta_3, \beta_4)$ of $\boldsymbol{\Sigma}$ and the parameters $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \gamma_2, \gamma_3, \gamma_4)$ of $\boldsymbol{\Sigma}^{-1}$. Although $\boldsymbol{\Sigma}$ is non-linear, $\boldsymbol{\Sigma}^{-1}$ is linear. Covariance structures in which $\boldsymbol{\Sigma}$ or $\boldsymbol{\Sigma}^{-1}$ has a linear structure was considered by Anderson (1969, 1970).

Now consider the quasi-simplex represented by (9). It is seen from (9) that although the product $\beta_2\omega_1 = \sigma_{21}$ is identified, β_2 and ω_1 are not separately identified. The product $\beta_2\omega_1$ is involved in the off-diagonal elements in the first column (and row) only. One can multiply β_2 by a constant and divide ω_1 by the same constant without changing the product. The change induced by ω_1 in σ_{11} can be absorbed in θ_1 in such a way that σ_{11} remains unchanged. Hence $\theta_1 = \text{Var}(\varepsilon_1)$ is not identified. For η_2 and η_3 we have

$$\omega_2 = \frac{\sigma_{32}\sigma_{21}}{\sigma_{31}},$$

$$\omega_3 = \frac{\sigma_{43}\sigma_{32}}{\sigma_{42}},$$

so that ω_2 and ω_3, and hence also θ_2 and θ_3, are identified. With ω_2 and ω_3 identified, β_3 and β_4 are identified by σ_{32} and σ_{43}. The middle coefficient β_3 is overidentified since

$$\beta_3\omega_2 = \frac{\sigma_{31}\sigma_{42}}{\sigma_{41}} = \sigma_{32}.$$

Since both ω_4 and θ_4 are involved in σ_{44} only, these are not identified but their sum σ_{44} is.

This analysis of the identification problem shows that for the "inner" variables y_2 and y_3, the parameters ω_2, ω_3, θ_2, θ_3 and β_3 are identified, whereas there is an indeterminacy associated with each of the "outer" variables y_1 and y_4. To eliminate these indeterminacies one condition must be imposed on the parameters ω_1, θ_1 and β_2, and another on the parameters ω_4 and θ_4. Perhaps the most natural way of eliminating the indeterminacies is to set $\theta_1 = \theta_2$ and $\theta_4 = \theta_3$. To illustrate a simplex model, some data published by Humphreys (1968) and analyzed by Werts et al. (1978) will be used.

Example 4: The variables include eight semesters of gradepoint averages, high school rank and a composite score on the American College Testing tests for approximately 1 600 undergraduate students at the University of Illinois. The correlation matrix is given in Table 4. One would like to estimate the reliabilities of the gradepoint averages and to test whether the auto-regressive process is stationary.

The example is analyzed in section 4.

3. General approaches to analysis of covariance structures

3.1. *General covariance and correlation structures*

Any covariance structure may be defined by specifying that the population variances and covariances of the observed variables are certain functions of parameters $\theta_1, \theta_2, ..., \theta_t$ to be estimated from data:

$\sigma_{ij} = \sigma_{ij}(\boldsymbol{\theta})$, or in matrix form $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$. It is assumed that the functions $\sigma_{ij}(\boldsymbol{\theta})$ are continuous and have continuous first derivatives and that $\boldsymbol{\Sigma}$ is positive definite at every point $\boldsymbol{\theta}$ of the admissible parameter space. The distribution of the observed variables is assumed to be multivariate with an unconstrained mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and is assumed to be sufficiently well described by the moments of first and second order, so that additional information about $\boldsymbol{\theta}$ contained in moments of higher order may be ignored. In particular this will hold if the distribution is multivariate normal. The condition that the mean vector $\boldsymbol{\mu}$ is unconstrained will be relaxed in section 5 where also mean structures will be considered.

A correlation structure is defined by specifying that the population correlations ϱ_{ij} of the observed variables are functions $\varrho_{ij} = \varrho_{ij}(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$. Such a correlation structure is treated as a covariance structure by specifying that

$$\boldsymbol{\Sigma} = \mathbf{D}_\sigma \mathbf{P}(\boldsymbol{\theta}) \mathbf{D}_\sigma \tag{10}$$

where $\mathbf{D}_\sigma$ is a diagonal matrix of population standard deviations $\sigma_1, \sigma_2, \ldots, \sigma_p$ of the observed variables, which are regarded as free parameters, and $\mathbf{P}(\boldsymbol{\theta})$ is the correlation matrix. The covariance structure (10) has parameters $\sigma_1, \sigma_2, \ldots, \sigma_p, \theta_1, \theta_2, \ldots, \theta_t$. The standard deviations $\sigma_1, \sigma_2, \ldots, \sigma_p$ as well as $\boldsymbol{\theta}$ must be estimated from data and the estimate of σ_i does not necessarily equal the corresponding standard deviation in the sample.

3.2. *Identification*

Before an attempt is made to estimate the parameters $\boldsymbol{\theta}$, the identification problem must be resolved. The identification problem is essentially whether or not $\boldsymbol{\theta}$ is uniquely determined by $\boldsymbol{\Sigma}$. Every $\boldsymbol{\theta}$ in the admissible parameter space generates a $\boldsymbol{\Sigma}$ but two or more $\boldsymbol{\theta}$'s may possibly generate the same $\boldsymbol{\Sigma}$. The whole model is said to be identified if for any two vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in a region of the parameter space, locally or globally, $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$ implies that $\boldsymbol{\Sigma}(\boldsymbol{\theta}_1) \neq \boldsymbol{\Sigma}(\boldsymbol{\theta}_2)$, i.e., if $\boldsymbol{\Sigma}$ is generated by one and only one $\boldsymbol{\theta}$. This means that all parameters are identified. However, even if the whole model is not identified some parameters can still be identified. Consider the set of all parameter vectors $\boldsymbol{\theta}$ generating the same $\boldsymbol{\Sigma}$. If a parameter $\boldsymbol{\theta}_i$ has the same value in all such vectors, this parameter is identified. For parameters which are identified the methods to be described will yield consistent estimators. If a model is not completely identified, appropriate restrictions may be imposed on $\boldsymbol{\theta}$ to make it so, and the choice of restrictions may affect the interpretation of the results of an estimated model.

Identifiability depends on the choice of model. To examine the identification problem for a particular model consider the equations

$$\sigma_{ij} = \sigma_{ij}(\boldsymbol{\theta}), \quad i \leqslant j. \tag{11}$$

There are $(\frac{1}{2})p(p+1)$ equations in t unknown parameters $\boldsymbol{\theta}$. Hence a necessary condition for identification of all parameters is that

$$t \leqslant (\tfrac{1}{2})p(p+1).$$

If a parameter θ can be determined from $\boldsymbol{\Sigma}$ by solving the equations (11) or a subset of them, this parameter is identified; otherwise it is not. Often some parameters can be determined from $\boldsymbol{\Sigma}$ in several ways, i.e., by using different sets of equations. This gives rise to overidentifying conditions on $\boldsymbol{\Sigma}$ which must hold if the model is true. Since the equations (11) are often non-linear, the solution of the equations is often complicated and tedious and explicit solutions for all θ's seldom exist. In the previous section we discussed the identification problem in terms of some specific examples.

There are various ways in which the computer program may be used to check the identification status of the model. If the GLS or ML methods are used for estimation (see the next subsection) the information matrix may be obtained and checked for positive definiteness. If the model is identified then the information matrix is almost certainly positive definite. If the information matrix is singular, the model is not identified and the rank of the information matrix may indicate which parameters are not identified. Another procedure which may also be used when other methods of estimation are used is the following. Choose a set of reasonable values for the parameters and compute $\boldsymbol{\Sigma}$. Then run the program with this $\boldsymbol{\Sigma}$ as input matrix and estimate $\boldsymbol{\theta}$. If this results in the same estimated values as those used to generate $\boldsymbol{\Sigma}$, then it is most likely that the model is identified. Otherwise, those parameters which gave a different value are probably not identified.

3.3. *Estimation*

The population is characterized by the mean vector $\boldsymbol{\mu}$, which is unconstrained, and the covariance matrix $\boldsymbol{\Sigma}$ which is a function of $\boldsymbol{\theta}$. In practice $\boldsymbol{\theta}$ is unknown and must be estimated from a sample of N independent observations on the random vector $\mathbf{x}$ of order p. Let $\mathbf{S} = (s_{ij})$ be the usual sample covariance matrix of order $p \times p$, based on $n = N - 1$ degrees of freedom. The information provided by $\mathbf{S}$ may also be represented by a correlation matrix $\mathbf{R} = (r_{ij})$ and a set of standard deviations $s_1, s_2, \ldots, s_p$,

where $s_i = (s_{ii})^{\frac{1}{2}}$ and $r_{ij} = s_{ij}/s_i s_j$. In many applications both the origin and the unit in the scales of measurement are arbitrary or irrelevant and then only the correlation matrix may be of interest. In such cases one takes $\mathbf{S}$ to be the correlation matrix $\mathbf{R}$ in what follows.

Since the mean vector is unconstrained, and higher moments are ignored, the estimation problem may be regarded as a problem of how to fit a matrix $\boldsymbol{\Sigma}$ of the form $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ to the observed covariance matrix $\mathbf{S}$. Although a number of different methods of an ad hoc nature have been used in specific cases there appears to be only three methods which can be used in general. These are the *unweighted least squares* (ULS) method, which minimizes

$$U = (\tfrac{1}{2}) \operatorname{tr} (\mathbf{S} - \boldsymbol{\Sigma})^2, \tag{12}$$

the *generalized least squares* (GLS) method, which minimizes

$$G = (\tfrac{1}{2}) \operatorname{tr} (\mathbf{I} - \mathbf{S}^{-1}\boldsymbol{\Sigma})^2, \tag{13}$$

and the *maximum likelihood* (ML) method, which minimizes

$$M = \operatorname{tr} (\boldsymbol{\Sigma}^{-1}\mathbf{S}) - \log |\boldsymbol{\Sigma}^{-1}\mathbf{S}| - p. \tag{14}$$

Each function is to be minimized with respect to $\boldsymbol{\theta}$.

All three functions U, G and M may be minimized by basically the same algorithm. The notation $F = F(\mathbf{S}, \boldsymbol{\Sigma})$ will be used for any one of the three functions. The GLS and ML methods are scale-free in the sense that $F(\mathbf{S}, \boldsymbol{\Sigma}) = F(\mathbf{DSD}, \mathbf{D\Sigma D})$ for any diagonal matrix of positive scale-factors; ULS does not have this property. With ULS, an analysis of $\mathbf{S}$ and of $\mathbf{DSD}$ yield results which may not be properly related. When $\mathbf{x}$ has a multivariate normal distribution both GLS and ML yield estimates that are efficient in large samples. Both GLS and ML require a positive definite covariance matrix $\mathbf{S}$ or correlation matrix $\mathbf{R}$; ULS will work even on a matrix which is non-Gramian.

Under the assumption that $\mathbf{x}$ has a multinormal distribution or that $\mathbf{S}$ has a Wishart distribution, M in (6) is a transform of the log-likelihood function for the sample, hence its association to the maximum likelihood method. Jöreskog and Goldberger [1972] derived the expression for G from Aitken's [1934–35] principle of generalized least squares using estimated asymptotic variances and covariances of the elements of $\mathbf{S}$ under multinormality of $\mathbf{x}$. Browne [1974] justified GLS under the slightly more general assumption that the elements of $\mathbf{S}$ have an asymptotic normal distribution. Since the variances and covariances in $\mathbf{S}$ are generally correlated and have unequal variances, it would seem that ULS uses the wrong metric in measuring deviations between $\mathbf{S}$ and $\boldsymbol{\Sigma}$. Nevertheless, ULS produces consistent estimators under more general assumptions than those which have been used to justify ML and GLS.

All three functions U, G and M are members of a general class of weighted least squares (WLS) functions (see Browne, 1974 and Lee, 1979). Let $\mathbf{A}$ be a symmetric positive definite weight matrix. Then WLS minimizes the function

$$F = (\tfrac{1}{2}) \operatorname{tr} \{[(\mathbf{S} - \boldsymbol{\Sigma}(\boldsymbol{\theta})]\mathbf{A}\}^2$$

with respect to $\boldsymbol{\theta}$.

In ULS, $\mathbf{A} = \mathbf{I}$. Browne (1974) showed, under general regularity conditions, that if $\mathbf{A}$ is a random matrix converging in probability to $\boldsymbol{\Sigma}^{-1}$ then the estimator $\hat{\boldsymbol{\theta}}$ which minimizes F is consistent and asymptotically efficient. A convenient choice of $\mathbf{A}$ is $\mathbf{A} = \mathbf{S}^{-1}$ yielding the GLS function G in (13). Maximum likelihood estimators will be obtained if $\mathbf{A} = \boldsymbol{\Sigma}_0^{-1}$ where $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$, $\boldsymbol{\theta}_0$ being the true $\boldsymbol{\theta}$. Since $\boldsymbol{\theta}_0$ is unknown this is not operational. However, ML estimates will be obtained if F is minimized iteratively and $\mathbf{A} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})$ is updated in each iteration. The estimator of $\boldsymbol{\theta}$ which minimizes F in this way is identical to that which minimizes M in (14). Browne (1974) showed that the GLS and ML estimators are asymptotically equivalent. Computationally there is a slight advantage with GLS compared with ML since the weight matrix is constant through the iterations. When the covariance structure is linear, GLS has a great advantage for then the function G is exactly quadratic in $\boldsymbol{\theta}$ and can therefore be minimized in one step.

The derivatives of F are

$$\partial F/\partial \theta_i = \operatorname{tr} [\mathbf{A}(\boldsymbol{\Sigma} - \mathbf{S})\mathbf{A}\,\partial\boldsymbol{\Sigma}/\partial\theta_i], \tag{15}$$

where $\mathbf{A} = \mathbf{I}$ in ULS, $\mathbf{A} = \mathbf{S}^{-1}$ in GLS and $\mathbf{A} = \boldsymbol{\Sigma}^{-1}$ in ML. Assuming that $\mathbf{S}$ converges in probability to $\boldsymbol{\Sigma}$ and ignoring terms of order $\boldsymbol{\Sigma} - \mathbf{S}$, the second derivatives are approximately

$$\partial^2 F/\partial\theta_i \partial\theta_j = \operatorname{tr} [\mathbf{A}\,\partial\boldsymbol{\Sigma}/\partial\theta_i \mathbf{A}\,\partial\boldsymbol{\Sigma}/\partial\theta_j]. \tag{16}$$

Note that (15) and (16) require only the first derivatives of the covariance structure functions $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. In ML and GLS, (16) yields the elements of the information matrix which is positive definite at every point $\boldsymbol{\theta}$ of the admissible parameter space, if $\boldsymbol{\theta}$ is identified.

The function $F(\boldsymbol{\theta})$ may be minimized numerically by Fisher's scoring method (see e.g. Rao, 1973, Section 5g) or the method of Davidon–Fletcher–

Powell (see Fletcher & Powell, 1963; see also Gruvæus & Jöreskog, 1970). Lee & Jennrich (1979) and Lee (1979) suggested that F can be minimized by the Gauss–Newton algorithm but this is equivalent to Fisher's scoring algorithm when the second-order derivatives are approximated by (16).

The minimization starts at an arbitrary starting point $\boldsymbol{\theta}^{(1)}$ and generates successively new points $\boldsymbol{\theta}^{(2)}$, $\boldsymbol{\theta}^{(3)}$, ..., such that $F(\boldsymbol{\theta}^{(s+1)}) < F(\boldsymbol{\theta}^{(s)})$ until convergence is obtained.

Let $\mathbf{g}^{(s)}$ be the gradient vector $\partial F/\partial\boldsymbol{\theta}$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(s)}$ and let $\mathbf{E}^{(s)}$ be the matrix whose elements are given by (16) evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(s)}$. Then Fisher's scoring method computes a correction vector $\boldsymbol{\delta}^{(s)}$ by solving the symmetric equation system

$$\mathbf{E}^{(s)}\boldsymbol{\delta}^{(s)} = \mathbf{g}^{(s)} \tag{17}$$

and then computes the new point as

$$\boldsymbol{\theta}^{(s+1)} = \boldsymbol{\theta}^{(s)} - \boldsymbol{\delta}^{(s)}. \tag{18}$$

This requires the computation of $\mathbf{E}^{(s)}$ and the solution of (17) in each iteration which is often quite time consuming. An alternative is to use the method of Davidon–Fletcher–Powell, which evaluates only the inverse of $\mathbf{E}^{(1)}$ and in subsequent iterations $\mathbf{E}^{-1}$ is improved, using information built up about the function, so that ultimately $\mathbf{E}^{-1}$ converges to an approximation of the inverse of $\partial^2 F/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$ at the minimum.

In GLS and ML, (2/N) times the inverse of the information matrix $\mathbf{E}$, given by (8) and evaluated at the minimum of F, provides an estimate of the asymptotic covariance matrix of the estimators $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. The square root of the diagonal elements of $(2/N)\mathbf{E}^{-1}$ are large-sample estimates of the standard errors of the $\hat{\theta}$'s.

Unfortunately no statistical theory is available for computing standard errors for ULS estimators. Such standard errors may be obtained by jackknifing but this requires extensive computation.

3.4. *Assessment of fit*

When the number of independent parameters in $\boldsymbol{\theta}$ is less than the total number of variances and covariances in $\boldsymbol{\Sigma}$, i.e., when $t < (\frac{1}{2})p(p+1)$, the model imposes conditions on $\boldsymbol{\Sigma}$ which must hold if the model is true. In GLS and ML, the validity of these conditions, i.e., the validity of the model, may be tested by a likelihood ratio test. The logarithm of the likelihood ratio is simply $(N/2)$ times the minimum value of the function F. Under the model, this is distributed, in large samples, as a χ^2 distribution with degrees of freedom equal to

$$d = (\tfrac{1}{2})p(p+1) - t. \tag{19}$$

3.5. *Tests of structural hypotheses*

Once the validity of the model has been reasonably well established, various structural hypotheses within the model may be tested. One can test hypotheses of the forms

(i) that certain θ's are fixed equal to assigned values and/or

(ii) that certain θ's are equal in groups.

Each of these two types of hypotheses leads to a covariance structure $\boldsymbol{\Sigma}(\boldsymbol{\nu})$ where $\boldsymbol{\nu}$ is a subset of $u < t$ elements of $\boldsymbol{\theta}$. Let F_ν be the minimum of F under the structural hypothesis and let F_θ be the minimum of F under the general model. Then $(N/2)(F_\nu - F_\theta)$ is approximately distributed as χ^2 with $t-u$ degrees of freedom.

3.6. *The use of χ^2 in exploratory studies*

The values of χ^2 should be interpreted very cautiously because of the sensitivity of χ^2 to various model assumptions such as linearity, additivity, multinormality, etc., but also for other reasons. In most empirical work many of the models considered may not be very realistic. If a sufficiently large sample were obtained, the test statistic would, no doubt, indicate that any such model is statistically untenable. The model should rather be that $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ represents a reasonable approximation to the population covariance matrix. From this point of view the statistical problem may not be one of testing a given hypothesis (which a priori may be considered false) but rather one of fitting various models with different numbers of parameters and to decide when to stop fitting. In other words, the problem is to extract as much information as possible out of a sample of given size without going so far that the result is affected to a large extent by "noise". It is reasonable and likely that more information can be extracted from a large sample than from a small one. In such a problem it is the difference between χ^2 values that matters rather than the χ^2 values themselves. In an exploratory study, if a value of χ^2 is obtained which is large compared to the number of degrees of freedom, the fit may be examined by an inspection of the residuals, i.e., the discrepancies between observed and reproduced values. Often the results of an analysis, an inspection of residuals or other considerations will suggest ways to relax the model somewhat by introducing more parameters. The new model usually yields a smaller χ^2. If the drop in χ^2 is large compared to the difference in degrees of freedom, this is an indication that the change made in the model represents a real improvement. If, on the other hand, the drop in χ^2 is close

to the difference in number of degrees of freedom, this is an indication that the improvement in fit is obtained by "capitalizing on chance" and the added parameters may not have any real significance or meaning.

Often it is not possible, or even desirable, to specify the model completely since there may be other models which are equally plausible. In such a situation it is necessary to have a technique of analysis which will give information about which of a number of alternative models is (are) the most reasonable. Also, if there is sufficient evidence to reject a given model due to poor fit to the data, the technique should be designed to suggest which part of the model is causing the poor fit. The examples of section 4 illustrate the assessment of fit of a model and strategies for model modification.

3.7. *The ACOVS model*

The approach to analysis of covariance structures described in sections 3.1–3.6 is completely general in the sense that *any* covariance structure can be handled. This approach has not been used much in practice because it requires the specification, by means of programmed subroutines, of the functions $\sigma_{ij}(\boldsymbol{\theta})$ and $\partial\sigma_{ij}/\partial\boldsymbol{\theta}$ for each application. Another approach, which has been found to be extremely useful in practice, is to assume a definite form for $\boldsymbol{\Sigma}$ but one which is still so general and flexible that it can handle most problems arising in practice. Two such approaches have been developed: the ACOVS-model and the LISREL-model.

The ACOVS-model (Jöreskog, 1970*a*, 1973*a*, 1974) assumes that $\boldsymbol{\Sigma}$ has the following form

$$\boldsymbol{\Sigma} = \mathbf{B}(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}^2)\mathbf{B}' + \boldsymbol{\Theta}^2, \tag{20}$$

where $\mathbf{B}(p \times q)$, $\boldsymbol{\Lambda}(q \times r)$, the symmetric matrix $\boldsymbol{\Phi}(r \times r)$ and the diagonal matrices $\boldsymbol{\Psi}(q \times q)$ and $\boldsymbol{\Theta}(p \times p)$ are parameter matrices. The covariance structure (20) will arise if the vector of observed variables $\mathbf{x}(p \times 1)$ has the form

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{B}\boldsymbol{\Lambda}\boldsymbol{\xi} + \mathbf{B}\boldsymbol{\zeta} + \mathbf{e} \tag{21}$$

where $\boldsymbol{\mu}$ is the mean vector of $\mathbf{x}$ and $\boldsymbol{\xi}$, $\boldsymbol{\zeta}$, and $\mathbf{e}$ are uncorrelated random vectors of latent (unobserved) variables with zero means and covariance matrices $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}^2$ and $\boldsymbol{\Theta}^2$, respectively.

When applying the model (20), the number of variables p is given by the data, and q and r are given by the particular application that the investigator has in mind. In any such application any parameter in $\mathbf{B}$, $\boldsymbol{\Lambda}$, $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$, or $\boldsymbol{\Theta}$ may be known a priori and one or more subsets of the remaining parameters may have identical but unknown values. Thus, parameters are of three kinds: (a) *fixed parameters* that have been assigned given values, (b) *constrained parameters* that are unknown, but equal to one or more other parameters, and (c) *free parameters* that are unknown and not constrained to be equal to any other parameter. The advantage of such an approach is the great generality and flexibility obtained by the various specifications that may be imposed. Thus the general model contains a wide range of specific models. The examples considered in section 2 are all of the form of (20) or can be reparameterized to be of this form. Model (1) uses $\mathbf{B} = \mathbf{I}$, $\boldsymbol{\Theta} = \mathbf{0}$, $\boldsymbol{\Lambda} = \mathbf{A}$ and equality constraints on the diagonal elements of $\boldsymbol{\Psi}$. Models (2a–b) use $\mathbf{B} = \boldsymbol{\Delta}$ in addition. Models (3) and (4) use $\boldsymbol{\Lambda} = \mathbf{I}$ and $\boldsymbol{\Psi} = \mathbf{0}$. Model (6) is of the form (20) with $\mathbf{B} = \mathbf{I}$, $\boldsymbol{\Psi} = \mathbf{0}$ and

$$\boldsymbol{\Lambda} = \begin{pmatrix} 1 & 0 \\ \lambda_1 & 0 \\ 0 & 1 \\ 0 & \lambda_2 \end{pmatrix} \qquad \boldsymbol{\Phi} = \begin{pmatrix} \omega_{11} & \\ \omega_{21} & \omega_{22} \end{pmatrix}$$

$$\boldsymbol{\Theta} = \begin{pmatrix} \theta_{11} & & & \\ 0 & \theta_{22} & & \\ \theta_{31} & 0 & \theta_{33} & \\ 0 & \theta_{42} & 0 & \theta_{44} \end{pmatrix}$$

The path analysis model (8) and the simplex model (9) require reparameterization before they can be written in the form of (20), see Jöreskog (1970*b*). One disadvantage with the reparameterization is that one will not get estimates of the original parameters. Although estimates of the original parameters can easily be computed afterwards it is more difficult to obtain standard errors of the original parameters. For this reason the LISREL model has been developed to accomodate the original parameters of path models and simplex models, but as will be seen, the ACOVS-model is a submodel of the LISREL-model, so that many other models may also be fitted.

3.8. *The LISREL model*

The LISREL model (Jöreskog, 1973*b*, 1977; Jöreskog & Sörbom, 1978) considers random vectors $\boldsymbol{\eta}' = (\eta_1, \eta_2, \ldots, \eta_m)$ and $\boldsymbol{\xi}' = (\xi_1, \xi_2, \ldots, \xi_n)$ of latent dependent and independent variables, respectively and the following system of linear structural relations

$$\mathbf{B}\boldsymbol{\eta} = \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \tag{22}$$

where $\mathbf{B}(m \times m)$ and $\mathbf{\Gamma}(m \times n)$ are coefficient matrices and $E(\boldsymbol{\zeta}) = \mathbf{0}$. It is furthermore assumed that $\boldsymbol{\zeta}$ is uncorrelated with $\boldsymbol{\xi}$ and that $\mathbf{B}$ is nonsingular.

The vectors $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ are not observed but instead vectors $\mathbf{y}' = (y_1, y_2, \ldots, y_p)$ and $\mathbf{x}' = (x_1, x_2, \ldots, x_q)$ are observed such that

$$\mathbf{y} = \mathbf{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{23}$$

$$\mathbf{x} = \mathbf{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta} \tag{24}$$

where $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$ are vectors of errors of measurement in $\mathbf{y}$ and $\mathbf{x}$, respectively. We take $\mathbf{y}$ and $\mathbf{x}$ to be measured as deviations from their means. The matrices $\mathbf{\Lambda}_y(p \times m)$ and $\mathbf{\Lambda}_x(q \times n)$ are regression matrices of $\mathbf{y}$ on $\boldsymbol{\eta}$ and of $\mathbf{x}$ on $\boldsymbol{\xi}$, respectively. It is convenient to refer to $\mathbf{y}$ and $\mathbf{x}$ as the observed variables and $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ as the latent variables. The errors of measurement are assumed to be uncorrelated with $\boldsymbol{\eta}$, $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ but may be correlated among themselves.

Let $\mathbf{\Phi}(n \times n)$ and $\mathbf{\Psi}(m \times m)$ be the covariance matrices of $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$, respectively, and let $\mathbf{\Theta}_\varepsilon$ and $\mathbf{\Theta}_\delta$ be the covariance matrices of $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$, respectively. Then it follows, from the above assumptions, that the covariance matrix $\mathbf{\Sigma}[(p+q) \times (p+q)]$ of $\mathbf{z} = (\mathbf{y}', \mathbf{x}')'$ is

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{Q} & \mathbf{\Lambda}_y \mathbf{B}^{-1}\mathbf{\Gamma}\mathbf{\Phi}\mathbf{\Lambda}_x' \\ \mathbf{\Lambda}_x \mathbf{\Phi}\mathbf{\Gamma}'\mathbf{B}'^{-1}\mathbf{\Lambda}_y' & \mathbf{\Lambda}_x \mathbf{\Phi}\mathbf{\Lambda}_x' + \mathbf{\Theta}_\delta \end{pmatrix} \tag{25}$$

where

$$\mathbf{Q} = \mathbf{\Lambda}_y(\mathbf{B}^{-1}\mathbf{\Gamma}\mathbf{\Phi}\mathbf{\Gamma}'\mathbf{B}'^{-1} + \mathbf{B}^{-1}\mathbf{\Psi}\mathbf{B}'^{-1})\mathbf{\Lambda}_y' + \mathbf{\Theta}_\varepsilon$$

The elements of $\mathbf{\Sigma}$ are functions of the elements of $\mathbf{\Lambda}_y$, $\mathbf{\Lambda}_x$, $\mathbf{B}$, $\mathbf{\Gamma}$, $\mathbf{\Phi}$, $\mathbf{\Psi}$, $\mathbf{\Theta}_\varepsilon$ and $\mathbf{\Theta}_\delta$. In applications some of these elements are fixed and equal to assigned values. In particular, this is so for elements of $\mathbf{\Lambda}_y$, $\mathbf{\Lambda}_x$, $\mathbf{B}$ and $\mathbf{\Gamma}$, but we shall allow for fixed values in the other matrices also. For the remaining nonfixed elements of the eight parameter matrices one or more subsets may have identical but unknown values. Thus, the elements in $\mathbf{\Lambda}_y$, $\mathbf{\Lambda}_x$, $\mathbf{B}$, $\mathbf{\Gamma}$, $\mathbf{\Phi}$, $\mathbf{\Psi}$, $\mathbf{\Theta}_\varepsilon$ and $\mathbf{\Theta}_\delta$ are of three kinds:

(i) *fixed parameters* that have been assigned given values

(ii) *constrained parameters* that are unknown but equal to one or more other parameters and

(iii) *free parameters* that are unknown and not constrained to be equal to any other parameter.

Equations (22), (23) and (24), with the accompanying assumptions, definee the general LISREL model. Equations (23) and (24) constitute the measurement model and equation (22) constitutes the structural equation model.

The computer program for LISREL (see Jöreskog & Sörbom, 1978) is structured in such a way that one can handle all kinds of reasonable submodels in a simple manner. For example, by specifying that there are no y- and η-variables, the model becomes

$$\mathbf{x} = \mathbf{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta}, \tag{26}$$

which is the classical factor analysis model. By specifying that there are no x- and no ξ-variables, the model reduces to

$$\left.\begin{aligned} \mathbf{y} &= \mathbf{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\epsilon} \\ \mathbf{B}\boldsymbol{\eta} &= \boldsymbol{\zeta} \end{aligned}\right\}. \tag{27}$$

With this model we also have an ordinary factor analysis model, but in this case we can handle relations among the factors by specifying a structure of the $\mathbf{B}$-matrix. For example, by specifying a $\mathbf{B}$ which generates an autoregressive structure one can handle simplex models of the form (9) and more general ones, see e.g. Jöreskog & Sörbom (1977) and Jöreskog (1979).

By specifying that there are no x-variables and that $\mathbf{B}$ is an identity matrix we get

$$\left.\begin{aligned} \boldsymbol{\eta} &= \mathbf{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \\ \mathbf{y} &= \mathbf{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\epsilon} \end{aligned}\right\} \tag{28}$$

or equivalently, cf. (21)

$$\mathbf{y} = \mathbf{\Lambda}_y(\mathbf{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}) + \boldsymbol{\epsilon}, \tag{29}$$

which is a second-order factor analysis model equivalent to the ACOVS model (Jöreskog, 1974). In a similar manner we can get a model for interdependent systems by specifying that the $\mathbf{\Lambda}_y$ and $\mathbf{\Lambda}_x$ matrices are identity matrices and that $\boldsymbol{\delta}$ and $\boldsymbol{\epsilon}$ are zero, i.e., the model

$$\mathbf{B}\mathbf{y} = \mathbf{\Gamma}\mathbf{x} + \boldsymbol{\zeta}, \tag{30}$$

often used in econometrics.

4. Analysis of the examples

4.1. *Example 1: Repeated trials of the rod and frame test*

Estimation of the variance components according to model (1) gives

$$\hat{\sigma}_a^2 = 4.08 \quad \hat{\sigma}_b^2 = 13.83 \quad \hat{\sigma}_c^2 = 28.64 \quad \hat{\sigma}_\varepsilon^2 = 21.60$$

However, examination of the fit of the model to the data, reveals that the fit is very poor: $\chi^2 = 319.4$ with 74 degrees of freedom. We shall therefore seek

Table 5. *Summary of analyses*

Hypothesis	No. par	χ^2	d.f.	P
H_1	4	37.33	6	0.00
H_2	5	1.93	5	0.86
H_3	8	36.21	2	0.00
H_4	9	0.70	1	0.40

Table 6. *Tests of hypotheses*

	Parallel	Congeneric	
$\varrho = 1$	$\chi^2_6 = 37.33$	$\chi^2_2 = 36.21$	$\chi^2_4 = 1.12$
$\varrho \neq 1$	$\chi^2_5 = 1.93$	$\chi^2_2 = 0.70$	$\chi^2_4 = 1.23$
	$\chi^2_1 = 35.40$	$\chi^2_1 = 35.51$	

an alternative model which accounts for the data better. This is obtained by structuring the error component **e**.

There are six distinct experimental conditions among the twelve trials, each one repeated twice. Let τ_i, $i = 1, 1, \ldots, 6$ be random components associated with these experimental conditions. Then

$$x_{i\alpha} = \tau_i + e_{i\alpha}$$

where $\alpha = 1, 2$ indexes the two replications. This simply means that one should allow the error variances to be different for different experimental conditions but still equal within replications of the same condition. An analysis according to this model gives $\chi^2 = 179.6$ with 69 degrees of freedom. The reduction in χ^2 clearly indicates that the error variances depend on the experimental condition.

The ML-estimates of the variance components are now, with standard errors below the estimates,

$$\hat{\sigma}^2_a = 4.18 \quad \hat{\sigma}^2_b = 11.26 \quad \hat{\sigma}^2_c = 27.10$$
$$(0.75) \qquad (1.72) \qquad (4.24)$$

$$\hat{\sigma}^2_{e_1} = 22.17 \quad \hat{\sigma}^2_{e_2} = 34.61 \quad \hat{\sigma}^2_{e_3} = 29.07 \quad \hat{\sigma}^2_{e_4} = 37.32$$
$$(2.71) \qquad (3.89) \qquad (3.34) \qquad (4.30)$$

$$\hat{\sigma}^2_{e_5} = 11.74 \quad \hat{\sigma}^2_{e_6} = 5.09$$
$$(1.41) \qquad (0.67)$$

The results indicate that most of the variance in the trials is associated with the chair effect. The variance due to the frame effect is less than half of this and the variance due to general bias is still smaller. The error variances are generally quite large except for the two experimental conditions in which the chair is already vertical.

4.2. *Example 2: The disattenuated correlation between two vocabulary tests*

The measurement model set up in section 2 is specified as a LISREL model with no y- and no η-variables as in (26). All four models H_1, H_2, H_3 and H_4 can be estimated in one run.

The results are shown in Table 5. Each hypothesis is tested against the general alternative that $\boldsymbol{\Sigma}$ is unconstrained. To consider various hypotheses that can be tested, the four χ^2 values of Table 5 are recorded in a 2×2 table as in Table 6. Test of H_1 against H_2 gives $\chi^2 = 35.40$ with 1 degree of freedom. An alternative test is H_3 against H_4, which gives $\chi^2 = 35.51$ with 1 degree of freedom. Thus, regardless of whether we treat the two pairs of tests as parallel or congeneric, the hypothesis $\varrho = 1$ is rejected. There is strong evidence that the unspeeded and speeded tests do not measure the same trait. The hypothesis of parallelness of the two pairs of tests can also be tested by means of Table 6. This gives $\chi^2 = 1.12$ or $\chi^2 = 1.23$ with 4 degrees of freedom, depending on whether we assume $\varrho = 1$ or $\varrho \neq 1$. Thus we cannot reject the hypothesis that the two pairs of tests are parallel. It appears that H_2 is the most reasonable of the four hypotheses. The maximum likelihood estimate of ϱ under H_2 is $\hat{\varrho} = 0.899$ with a standard error of 0.019. An approximate 95% confidence interval for ϱ is $0.86 < \varrho < 0.94$. The substantive matter is discussed further in Jöreskog (1974).

4.3. *Example 3: The stability of alienation*

The model (7) is an example of the general LISREL form (22), (23) and (24). We distinguish between the two models

(A) $\theta_{31} = \theta_{42} = 0$

(B) θ_{31} and θ_{42} free

The maximum likelihood estimates of the parameters with standard errors in parenthesis are given in Table 7. The stability of alienation over time is reflected in the parameter β. The influence of SES on Alienation at the two occasions is significant in model A. The coefficient for 1967, γ_1, is -0.614 with a standard error of 0.056 and for 1971, γ_2, it is -0.174 with a standard error equal to 0.054. The negative sign of the λ-coefficients indicates that for high socioeconomic status the alienation is low and vice versa. However, the overall fit of the model A is not acceptable: χ^2 with six degrees of freedom equals 71.5. Model B is intuitively more plausible. As can be seen from Table 7 the inclusion of θ_{31} and θ_{42} results in a model with an acceptable overall fit. A test of the hypothesis that both θ_{31} and θ_{42} are

Table 7. *Maximum likelihood estimates for models A and B*

The standard errors of the estimates are given within parenthesis

Parameter	Model A	Model B
λ_1	0.889 (.041)	0.979 (.062)
λ_2	0.849 (.040)	0.922 (.060)
λ_3	5.329 (.430)	5.221 (.422)
β	0.705 (.054)	0.607 (.051)
γ_1	−0.614 (.056)	−0.575 (.056)
γ_2	−0.174 (.054)	−0.227 (.052)
ψ_1	5.307 (.473)	4.846 (.468)
ψ_2	3.742 (.388)	4.089 (.405)
ϕ	6.666 (.641)	6.803 (.650)
θ_{11}	4.015 (.343)	4.735 (.454)
θ_{22}	3.192 (.271)	2.566 (.404)
θ_{33}	3.701 (.373)	4.403 (.516)
θ_{44}	3.625 (.292)	3.074 (.435)
θ_{31}	—	1.624 (.314)
θ_{42}	—	0.339 (.261)
θ_{55}	2.944 (.500)	2.807 (.508)
θ_{66}	260.982 (18.242)	264.809 (18.154)
χ^2	71.470	4.730
d.f.	6	4

zero yields $\chi^2 = 66.8$ with 2 degrees of freedom so that this hypothesis must be rejected.

4.4. *Example 4: A simplex model for academic performance*

The quasi-simplex model underlying the covariance structure (9) is a LISREL model of the form (27) with no x- and no ξ-variables and with $\mathbf{\Lambda}_y = \mathbf{I}$ and

$$\mathbf{B} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -\beta_2 & 1 & 0 & 0 \\ 0 & -\beta_3 & 1 & 0 \\ 0 & 0 & -\beta_4 & 1 \end{pmatrix}$$

Using first only the measures of grade-point averages in Table 4, estimation of a quasi-Markov simplex gives the following correlations between the true academic achievements $\eta_2, \eta_3, \ldots, \eta_7$. These correlations are

	η_2	η_3	η_4	η_5	η_6	η_7
η_2	1.000					
η_3	0.838	1.000				
η_4	0.812	0.969	1.000			
η_5	0.724	0.865	0.892	1.000		
η_6	0.677	0.809	0.834	0.935	1.000	
η_7	0.619	0.740	0.763	0.855	0.914	1.000

Here every correlation ϱ_{ij} with $|i-j| > 1$ is the product of the correlations just below the diagonal. For example, $\varrho(\eta_5, \eta_2) = 0.838 \times 0.969 \times 0.892 = 0.724$. These correlations form a perfect Markov simplex. The goodness of fit test of the model gives $\chi^2 = 23.91$ with 15 degrees of freedom. This represents reasonably good fit considering the large sample size. The reliabilities of the semester grades $y_2, y_3, \ldots, y_7$ can also be obtained directly from the solution in which the η's are standardized. The reliabilities are

y_2	y_3	y_4	y_5	y_6	y_7
0.569	0.575	0.562	0.584	0.581	0.608

A test of the hypothesis that these are equal gives $\chi^2 = 2.17$ with 5 degrees of freedom, so that this hypothesis is not rejected by the data despite the large sample size.

In this example the correlations $\varrho(\eta_1, \eta_j)$, $j \neq 1$ and $\varrho(\eta_i, \eta_8)$, $i \neq 8$ and the reliabilities of y_1 and y_8 are not identified. However, in view of the above test of equality of reliabilities it seems reasonable to assume that all reliabilities or equivalently all error variances in the standardized solution are equal for y_1 through y_8. This assumption makes it possible to estimate the intercorrelations among all the η's.

Assuming that x_0 and x_0' are indicators of pre-college academic achievement η_0 which is assumed to influence the true academic achievement in the first semester η_1, one can estimate again the quasi-Markov simplex and show how this use of x_0 and x_0' helps identify the parameters of the model. The only parameter which is now not identified is θ_8, the error variance in y_8. This gives a $\chi^2 = 36.92$ with 28 degrees of freedom. If we assume that the reliabilities of all the semester grades are equal, all parameters are identified and the goodness of fit becomes 45.22 with 34 degrees of freedom. The difference 8.30 with 6 degrees of freedom provides another test of equality of the reliabilities. Finally a test of the hypothesis that the whole process is stationary, i.e., that

$$\beta_2 = \beta_3 = \ldots = \beta_8$$

$$\theta_2 = \theta_3 = \ldots = \theta_8$$

gives $\chi^2 = 12.82$ with 11 degrees of freedom so that this hypothesis cannot be rejected. There is good evidence that the whole Markov process is stable over time.

5. Generalizations

In this section we consider two different generalizations of the models and methods discussed in the previous sections. One generalization allows for mean structures as well as covariance structures; the

other concerns the analysis of data from several independent samples.

5.1. *Mean structures as well*

Suppose that not only the covariance matrix $\boldsymbol{\Sigma}$ but also the mean vector $\boldsymbol{\mu}$ is a function of the parameter vector $\boldsymbol{\theta}$. The easiest way to generalize the development 3.1–3.5 to this situation is to consider the augmented vector $\mathbf{x}^{*\prime} = (\mathbf{x}', 1)$ where the last variable is a fixed variable which is constant equal to one for all observations, and use matrices of moments about zero instead of covariance matrices. Let $\mathbf{M}[(p+1)\times(p+1)]$ be the sample moment matrix of $\mathbf{x}^*$, i.e.,

$$\mathbf{M} = \begin{pmatrix} \mathbf{M}_{xx} & \bar{\mathbf{x}} \\ \bar{\mathbf{x}}' & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{S} + \bar{\mathbf{x}}\bar{\mathbf{x}}' & \bar{\mathbf{x}} \\ \bar{\mathbf{x}}' & 1 \end{pmatrix}$$

where, as before, $\bar{\mathbf{x}}$ is the sample mean vector of $\mathbf{x}$ and $\mathbf{S}$ the sample covariance matrix of $\mathbf{x}$ and

$$\mathbf{M}_{xx} = (1/N)\sum_{\alpha=1}^{N} \mathbf{x}_\alpha \mathbf{x}_\alpha',$$

$\mathbf{x}_\alpha$ being the αth observation of $\mathbf{x}$.

The corresponding population moment matrix is

$$\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}' & \boldsymbol{\mu} \\ \boldsymbol{\mu}' & 1 \end{pmatrix}$$

Estimates of $\boldsymbol{\theta}$ may be obtained by fitting $\boldsymbol{\Omega}$ to $\mathbf{M}$ using $F(\boldsymbol{\Omega}, \mathbf{M})$ where F is anyone of the three fitting functions (12), (13) and (14). Jöreskog and Sörbom (1980) showed that (14) leads to maximum likelihood estimates in the usual sense.

5.2. *The LISREL-model with structured means*

In section 3.8 the LISREL model was defined by (22), (23) and (24) in which all random variables were assumed to have zero means. This assumption will now be relaxed and it will be shown that the LISREL IV computer program can be used also to estimate the same three equations even if they include constant intercept terms. This is possible if we introduce a fixed variable whose observations are all equal to one and analyze the sample moment matrix instead of the sample covariance matrix.

The LISREL model is now defined by the following three equations corresponding to (22), (23) and (24), respectively:

$$\mathbf{B}\boldsymbol{\eta} = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}, \tag{31}$$

$$\mathbf{y} = \boldsymbol{\nu}_y + \boldsymbol{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{32}$$

$$\mathbf{x} = \boldsymbol{\nu}_x + \boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta}, \tag{33}$$

where $\boldsymbol{\alpha}$, $\boldsymbol{\nu}_y$ and $\boldsymbol{\nu}_x$ are vectors of constant intercept terms. As before, we assume that $\boldsymbol{\zeta}$ is uncorrelated with $\boldsymbol{\xi}$, $\boldsymbol{\epsilon}$ is uncorrelated with $\boldsymbol{\eta}$ and that $\boldsymbol{\delta}$ is uncorrelated with $\boldsymbol{\xi}$. However, it will not be necessary to assume that $\boldsymbol{\epsilon}$ is uncorrelated with $\boldsymbol{\delta}$ as we did in section 3.8. We also assume, as before, that $E(\boldsymbol{\zeta}) = \mathbf{0}$, $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $E(\boldsymbol{\delta}) = \mathbf{0}$, but it is not assumed that $E(\boldsymbol{\xi})$ and $E(\boldsymbol{\eta})$ are zero. The mean of $\boldsymbol{\xi}$, $E(\boldsymbol{\xi})$ will be a parameter denoted by $\boldsymbol{\kappa}$. The mean of $\boldsymbol{\eta}$, $E(\boldsymbol{\eta})$ is obtained from (31) as

$$E(\boldsymbol{\eta}) = \mathbf{B}^{-1}(\boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\kappa}) \tag{34}$$

By taking expectation of (32) and (33) we find the mean vectors of the observed variables to be

$$E(\mathbf{y}) = \boldsymbol{\nu}_y + \boldsymbol{\Lambda}_y\mathbf{B}^{-1}(\boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\kappa}) \tag{35}$$

$$E(\mathbf{x}) = \boldsymbol{\nu}_x + \boldsymbol{\Lambda}_x\boldsymbol{\kappa}. \tag{36}$$

In general, in a single population, all the mean parameters $\boldsymbol{\nu}_y$, $\boldsymbol{\nu}_x$, $\boldsymbol{\alpha}$ and $\boldsymbol{\kappa}$ will not be identified without further conditions imposed. However, in simultaneous analysis of data from several groups (see next subsection), simple conditions (see e.g. Jöreskog & Sörbom (1980)) can be imposed to make all the mean parameters identified. We shall not be concerned with identification here but merely confine ourselves to show how the model (31)–(33) can be written in the form of (22)–(24) which is the form used by the program.

The LISREL specification of (31)–(33) is as follows. We treat $\mathbf{y}$ and $\mathbf{x}$ as y-variables and $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ as η-variables in LISREL sense. In addition, we use a single fixed x-variable equal to 1. This variable 1 is also used as the first η-variable. We can then write the model in the form of (22) and (23) as

$$\begin{pmatrix} 1 & \mathbf{0}' & \mathbf{0}' \\ \mathbf{0} & \mathbf{B} & -\boldsymbol{\Gamma} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} 1 \\ \boldsymbol{\eta} \\ \boldsymbol{\xi} \end{pmatrix} = \begin{pmatrix} 1 \\ \boldsymbol{\alpha} \\ \boldsymbol{\kappa} \end{pmatrix} 1 + \begin{pmatrix} 0 \\ \boldsymbol{\zeta} \\ \boldsymbol{\xi} - E(\boldsymbol{\xi}) \end{pmatrix} \tag{37}$$

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{x} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\nu}_y & \boldsymbol{\Lambda}_y & \mathbf{0} \\ \boldsymbol{\nu}_x & \mathbf{0} & \boldsymbol{\Lambda}_x \end{pmatrix} \begin{pmatrix} 1 \\ \boldsymbol{\eta} \\ \boldsymbol{\xi} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon} \\ \boldsymbol{\delta} \end{pmatrix} \tag{38}$$

Note that the mean parameter vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\kappa}$ appear in the $\boldsymbol{\Gamma}$-matrix in LISREL and $\boldsymbol{\nu}_y$ and $\boldsymbol{\nu}_x$ appear in the first column of $\boldsymbol{\Lambda}_y$ in LISREL.

The matrix $\boldsymbol{\Sigma}$ in (25) should now be interpreted as the population moment matrix of the vector $(\mathbf{y}', \mathbf{x}', 1) = (\mathbf{z}', 1)$, say, where $\mathbf{z}$ corresponds to y-variables in LISREL and 1 corresponds to an x-variable. The LISREL estimates are obtained by minimizing the function $F(\boldsymbol{\theta})$, where $\mathbf{S}$ is now the sample moment matrix.

5.3. *Simultaneous analysis in several groups*

The LISREL model (31), (32) and (33) can be used to analyze data from several groups simultaneously according to LISREL models for each group with some or all parameters constrained to be equal over groups. Examples of such simultaneous analysis have been given by Jöreskog (1971), McGaw & Jöreskog (1971), Sörbom (1974, 1975, 1976, 1981) and Jöreskog & Sörbom (1980).

Consider a set of G populations. These may be different nations, states or regions, culturally or socioeconomically different groups, groups of individuals selected on the basis of some known or unknown selection variable, groups receiving different treatments, etc. In fact, they may be any set of mutually exclusive groups of individuals which are clearly defined. It is assumed that a number of variables have been measured in a random sample of individuals from each population.

It is assumed that a LISREL model of the form (31), (32) and (33) holds in each group. The model for group g is defined by the parameter matrices $\mathbf{\Lambda}_y^{(g)}$, $\mathbf{\Lambda}_x^{(g)}$, $\mathbf{B}^{(g)}$, $\mathbf{\Gamma}^{(g)}$, $\mathbf{\Phi}^{(g)}$, $\mathbf{\Psi}^{(g)}$, $\mathbf{\Theta}_\varepsilon^{(g)}$, $\mathbf{\Theta}_\delta^{(g)}$, where the superscript (g) refers to the gth group, $g = 1, 2, ..., G$. Each of these matrices may contain fixed, free and constrained parameters as before. If there are no constraints between groups, each group can be analyzed separately. However, if there are constraints between groups, the data from all groups must be analyzed simultaneously to get efficient estimates of the parameters. For example, if the measurement properties of the observed variables are the same in all groups one would postulate that

$$\mathbf{\Lambda}_y^{(1)} = \mathbf{\Lambda}_y^{(2)} = \ldots = \mathbf{\Lambda}_y^{(G)}$$

$$\mathbf{\Lambda}_x^{(1)} = \mathbf{\Lambda}_x^{(2)} = \ldots = \mathbf{\Lambda}_x^{(G)}$$

and perhaps also that

$$\mathbf{\Theta}_\varepsilon^{(1)} = \mathbf{\Theta}_\varepsilon^{(2)} = \ldots = \mathbf{\Theta}_\varepsilon^{(G)}$$

$$\mathbf{\Theta}_\delta^{(1)} = \mathbf{\Theta}_\delta^{(2)} = \ldots = \mathbf{\Theta}_\delta^{(G)}$$

The possible differences between groups would then be represented by differences in the distributions of the latent variables, i.e., by $\mathbf{\Theta}^{(g)}$ and $\mathbf{\Psi}^{(g)}$. By postulating

$$\mathbf{B}^{(1)} = \mathbf{B}^{(2)} = \ldots = \mathbf{B}^{(G)}$$

$$\mathbf{\Gamma}^{(1)} = \mathbf{\Gamma}^{(2)} = \ldots = \mathbf{\Gamma}^{(G)}$$

one can test the hypotheses that also the structural relations are invariant over groups.

In general, any degree of invariance can be tested, from the one extreme where all parameters are assumed to be invariant over groups to the other extreme when there are no constraints between groups.

To estimate the models simultaneously we minimize the fitting function

$$F = \sum_{g=1}^{G} (N_g/N)\, F(\mathbf{S}_g, \mathbf{\Sigma}_g)$$

where F, as before, is anyone of the functions (12), (13) and (14) and where N_g is the number of observations in group g and $N = N_1 + N_2 + \ldots + N_G$. When $G = 1$, this reduces to the same fitting function as before. When the observed variables have a multinormal distribution, F is minus $(N/2)$ times the logarithm of the likelihood function except for an additive constant.

The χ^2 goodness-of-fit measure is defined as before. This is a test of the hypotheses that the LISREL model holds in each group including all constraints, against the general alternative that all $\mathbf{\Sigma}^{(g)}$ are unconstrained positive definite, $g = 1, 2, ..., G$. The degrees of freedom are

$$d = (\tfrac{1}{2})G(p+q)(p+q+1) - t,$$

where t is the total number of independent parameters estimated in all groups.

While this approach to simultaneous analysis in several groups has been formulated in terms of LISREL models with unconstrained mean vectors it can be generalized in the same way to models with constraints on the means. In fact, the LISREL IV computer program (Jöreskog & Sörbom, 1978) cand handle models combining the features of sections 5.2 and 5.3.

Acknowledgement

The research reported in this paper has been supported by the Bank of Sweden Tercentenary Foundation under project *Methodology of Evaluation Research* (project director: Karl G. Jöreskog).

References

Aitken, A. C. (1934–35). On least squares and the linear combination of observations. *Proceedings of the Royal Society of Edinburgh* **55**, 42–48.

Anderson, T. W. (1969). Statistical inference for covariance matrices with linear structure. In *Multivariate analysis* II, (ed. P. R. Krishnaiah). pp. 55–66. Academic Press, New York.

Anderson, T. W. (1970). Estimation of covariance matrices which are linear combinations or whose inverses are

linear combinations of given matrices. In *Essays in probability and statistics*, pp. 1–24 (ed. R. C. Bose et al.). University of North Carolina Press, Chapel Hill, North Carolina.

Bentler, P. M. & Weks, D. G. Multivariate analysis with latent variables. In *Handbook of statistics, vol.* **2** (ed. P. R. Krishnaiah & L. Kanal). North Holland Publishing Co., Amsterdam (in press).

Bock, R. D. (1960). Components of variance analysis as a structural and discriminal analysis for phychological tests. *British Journal of Statistical Psychology* **13**, 151–163.

Bock, R. D. & Bargmann, R. E. (1966). Analysis of covariance structures. *Psychometrika* **31**, 507–534.

Browne, M. W. (1970). Analysis of covariance structures. Paper presented at the annual conference of the South African Statistical Association.

Browne, M. W. (1964). Generalized least squares estimators in the analysis of covariance structures. *South African Statistical Journal* **8**, 1–24. (Reprinted in D. J. Aigner & A. S. Goldberger (Eds.) (1977). *Latent variables in socioeconomic models*, pp. 302–226. North Holland Publishing Co., Amsterdam.

Browne, M. W. (1977) The analysis of patterned correlation matrices by generalized least squares. *British Journal of Mathematical and Statistical Psychology* **30**, 113–124.

Fletcher, R. & Powell, M. J. D. (1963). A rapidly convergent descent method for minimization. *The Computer Journal* **6**, 163–168.

Gruveaus, G. & Jöreskog, K. G. (1970). *A computer program for minimizing a function of several variables*. Research Bulletin 70–14. Educational Testing Service, Princeton N. J.

Guttman, L. (1954). A new approach to factor analysis: the radex. In *Mathematical thinking in the social sciences* (ed. P. F. Lazarsfeld), pp. 258–348. Columbia University Press, New York.

Hauser, R. M. & Goldberger, A. S. (1971). The treatment of unobservable variables in path analysis. In *Sociological methodology 1971* (ed. H. L. Costner), pp. 81–117. Jossey-Bass, London.

Humphreys, L. G. (1968). The fleeting nature of college academic success. *Journal of Educational Psychology* **59**, 375–380.

Jöreskog, K. G. (1970*a*). A general method for analysis of covariance structures. *Biometrika* **57**, 239–251.

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

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika* **36**, 409–426.

Jöreskog, K. G. (1973*a*). Analysis of covariance structures. In *Multivariate analysis. III* (ed. P. R. Krishnaiah), pp. 263–285. Academic Press, New York.

Jöreskog, K. G. (1973*b*). A general method for estimating a linear structural equation system. In *Structural equation models in the social sciences* (ed. A. S. Goldberger & O. D. Duncan), pp. 85–112. Seminar Press, New York.

Jöreskog, K. G. (1974). Analyzing psychological data by structural analysis of covariance matrices. In *Contemporary developments in mathematical psychology. II* (ed. R. C. Atkinson, D. H. Krantz, R. D. Luce & P. Suppes), pp. 1–56. W. H. Freeman & Co., San Francisco.

Jöreskog, K. G. (1977). Structural equation models in the social sciences: Specification, estimation and testing. In *Applications of statistics* (ed. P. R. Krishnaiah), pp. 265–286. North Holland Publishing Co., Amsterdam.

Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika* **43**, 443–477.

Jöreskog, K. G. (1979). Statistical estimation of structural models in longitudinal-developmental investigations. In *Longitudinal research in the study of behavior and development* (ed. J. R. Nesselroade & P. B. Baltes), pp. 303–352. Academic Press, New York.

Jöreskog, K. G. & Goldberger, A. S. (1972). Factor analysis by generalized least squares. *Psychometrika* **37**, 243–260.

Jöreskog, K. G. & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association* **10**, 631–639.

Jöreskog, K. G. & Sörbom, D. (1977). Statistical models and methods for analysis of longitudinal data. In *Latent variables in socio-economic* (ed. D. J. Aigner & A. S. Goldberger), pp. 235–285. North Holland Publishing Co., Amsterdam.

Jöreskog, K. G. & Sörbom, D. (1978). *LISREL IV—A general computer program for estimation of a linear structural equation system by maximum likelihood methods*. International Educational Services. Chicago.

Jöreskog, K. G. & Sörbom, D. (1980). *Simultaneous analysis of longitudinal data from several cohorts*. Research Report 80–5. Department of Statistics, University of Uppsala.

Lee, S. Y. (1979). The Gauss-Newton algorithm for the the weighted least squares factor analysis. *The Statistician* **27**, 103–114.

Lee, S. Y. & Jennrich, R. I. (1979). A study of algorithms for covariance structure analysis with specific comparisons using factor analysis. *Psychometrika* **44**, 99–113.

Lord, F. M. (1957). A significance test for the hypothesis that two variables measure the same trait except for errors of measurement. *Psychometrika* **22**, 207–220.

Lord, F. M. & Novick, M. E. (1968). *Statistical theories of mental test scores*. Addison-Wesley Publishing Co. Reading.

McDonald, R. P. (1974). Testing pattern hypotheses for covariance matrices. *Psychometrika* **39**, 189–201.

McDonald, R. P. (1975). Testing pattern hypotheses for correlation matrices. *Psychometrika* **40**, 253–255.

McDonald, R. P. (1978). A simple comprehensive model for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology* **31**, 59–72.

McGaw, B. & Jöreskog, K. G. (1971). Factorial invariance of ability measures in groups differing in intelligence and socioeconomic status. *British Journal of Mathematical and Statistical Psychology* **24**, 154–168.

Rao, C. R. (1973). *Linear statistical inference and its applications*. 2nd ed. Wiley, New York.

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.

Sörbom, D. (1975). Detection of correlated errors in longitudinal data. *British Journal of Mathematical and Statistical Psychology* **28**, 138–151.

Sörbom, D. (1976). A statistical model for the measurement of change in true scores. In *Advances in psychological and educational measurement* (ed. D. N. M. de Gruijter and L. J. Th. van der Kamp), pp. 159–169. Wiley, New York.

Sörbom, D. (1978). An alternative to the methodology for analysis of covariance. *Psychometrika* **43**, 381–396.

Sörbom, D. (1981). Structural equation models with structured means. To appear in *Systems under indirect observation: Causality, structure, prediction* (ed. K. G. Jöreskog and H. Wold). North-Holland Publishing Co., Amsterdam.

Werts, C. E. & Linn, R. L. (1970). Path analysis: Psychological examples. *Psychological Bulletin* **67**, 193–212.

Werts, C. E., Linn, R. L. & Jöreskog, K. G. (1978). Reliability of college grades from longitudinal data. *Educational and Psychological Measurement* **38**, 89–95.

Wheaton, B., Muthén, B., Alwin, D. & Summers, G. (1977). Assessing reliability and stability in panel models. In *Sociological methodology 1977* (ed. D. R. Heise), pp. 84–136. Jossey-Bass, San Francisco.

Wiley, D. E., Schmidt, W. H. & Bramble, W. J. (1973). Studies of a class of covariance structure models. *Journal of the American Statistical Association* **68**, 317–323.

Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics* **5**, 161–215.

Karl G. Jöreskog
Department of Statistics
University of Uppsala
P.O. Box 513
S-751 20 Uppsala
Sweden

DISCUSSION OF K. G. JÖRESKOG'S PAPER

Erling B. Andersen (University of Copenhagen

My comments on Professor Jöreskogs excellent and clear presentation will consist of three parts. Firstly I shall briefly mentioned the parallel theory for discrete data, with is now known under the name "latent structure analysis". Secondly I shall discuss a number of problems connected with model selection and model checking. Thirdly and finally I shall present two examples, which I hope can give rise to a fruitful discussion.

1. Several of Professor Jöreskogs models are based on the interplay between latent variables and observed or manifest variables. The statistical problems is characterizes by the fact, the we have a relatively simple structure in the latent space and a somewhat more complex structure in the observed variables.

Without pretending to be a complete list, we have at least four important type of models, which are essential both for the continuous models, as for example LISREL, and their discrete counterparts.

(i) Simple latent structure models; for example one-factor models.

(ii) Models for comparison of several populations. (Section 5.3 of the paper.)

(iii) Models for correlated latent variables. (Examples 2 and 3 of the paper.)

(iv) Longitudinal models. (Example 4 of the paper.)

All these models are also treated in latent structure analysis for discrete data. Andersen (1980) gives a survey of the field.

2. Two quotations from Professor Jöreskogs paper seem to suggest that the model is primarily checked by checking that the empirical covariance matrix has the structure prescribed by the model. On page 75 it is claimed that: if the covariance structure is accepted, the model is valid. And on p. 75 it is said that: "The result of an analysis, an inspection of the residuals or other considerations suggest ways to relax the model." (Here an analysis means an analysis of the covariance matrix.) Thus main emphasize seems to be on the covariance matrix. This among other things raises the question of where the modelling take place in the covariance matrix or in the original data. Not only where the modelling start but also where it ends.

From the examples given in the paper one gets the impression that remodelling to give a better fit is performed as reformulations of the covariance matrix. Many applied statisticians would undoubtedly feel it more relevant to go back to the original data for example by means of a analysis of the residuals in order to discover possible model reformulations.

The ambiguity in the formulation of a model is illustrated by the heavy attention one has to pay to identificationproblems. Personally I do not like the idea of computerized identification as described on p. 73 and feel that the identification of the model should be an integral and quite natural part of the process of formulating the model.

That the models are closely connected to the assumption of a normal distribution is very obvious To claim that only the first two moments are considered is in reality the same thing. In some cases this assumption does not seem appropriate. One point in case in example 3, where one could question whether the variable "Education" is normally distributed.

A relevant question could be: Is the model choice based on the type of data at hand? Has it thus been considered to transform the data before the covariance matrix is computed?

3. Finally the two examples: The *first* example are constructed data, where two of the models fit the same data. (The data is constructed in such a way that we have a complete fit.) In Table 1 is shown the covariance matrix between the amount of iron, copper, silver and gold found in the tombs after 37 of Nebuchadnezzar's generals.

Table 1. *Covariances between the metal found in 37 ancient tombs*

Variable		Iron	Copper	Silver	Gold
1	Iron	104.4	81.9	46.1	62.9
2	Copper	81.9	109.8	47.6	65.0
3	Silver	46.1	47.6	80.0	74.7
4	Gold	62.9	65.0	74.7	127.2

These data are fit by the model with two correlated latent variables p. 69 ($\beta_1 = 8.9$, $\beta_2 = 9.2$, $\beta_3 = 7.4$, $\beta_4 = 10.1$, $\theta_1 = \theta_2 = \theta_3 = \theta_4 = 25.2$, $\varrho = 0.7$). The theory is here that the activities of war and peace take place at different times and require different—although correlated-talents. Since iron and cobber are usually found in weapons and silver and gold in coins or juwelry, variables 1 and 2 relate to war and variables 3 and 4 to peace. The data thus bears evidence that this theory is correct.

But the data are also fitted by the longitudinal model p. 71 ($\beta_2 = 1.03$, $\beta_3 = 0.56$, $\beta_4 = 1.36$, $\theta_1 = \theta_2 = \theta_3 = \theta_4 = 25.2$, $\omega_1 = 79.2$, $\omega_2 = 84.6$, $\omega_3 = 54.8$, $\omega_4 = 102.0$). Here one can theorize that a general goes through four phases: First a lowel rank officer, then a high ranking officer, then a business man and finally a influential person at court. The four types of metal seem to associate naturally to these four phases. Since the model fit the data, also this theory is confirmed by the data. If the two theories are alternative and the metal found is the evidence available, how should we then choose or judge between the two theories.

The point I would like to make is that different models may fit the same set of data and statisticians would general tend to claim that they have verified or confirmed a theory if the derived statistical model fit the data collected to check the theory.

The *second* example is given in Table 2. The table shows the empirical covariances between examination average after high school and examination average after one and two years of study in economy at the University of Copenhagen. The figures are based on a random sample of 27 students.

For these data I have tried to fit the longitudinal simplex model of p. 71. Since we have 6 variances/

Table 2. *Covariances between examination averages for 27 economy students*

	Average		
Average	High school	1st year	2nd year
High school	0.792	0.600	0.137
1st year	0.600	1.265	0.600
2nd year	0.137	0.600	1.030

covariances and 6 parameters in the model (assuming $\theta_1 = \theta_2 = \theta_3$) we can fit directly. We get, when the model is

$$\begin{bmatrix} \omega_1 + \theta & & \\ \beta_2 \omega_1 & \omega_2 + \theta & \\ \beta_2 \beta_3 \omega_1 & \beta_3 \omega_2 & \omega_3 + \theta \end{bmatrix}$$

the following estimates

$\hat{\beta}_3 = 0.137/0.600 = 0.228$

$\hat{\omega}_2 = 0.600/0.228 = 2.631$

$\hat{\theta} = 1.265 - 2.631 = -1.367$!!

Hence the estimates does not belong to the range of the parameters pace.

The maximum likelihood estimates are undoubtedly given by (although I did not check)

$\hat{\theta} = 0$

$\hat{\omega}_1 = 0.792$

$\hat{\beta}_2 = 0.600/0.792 = 0.758$

$\hat{\omega}_2 = 1.265$

$\hat{\omega}_3 = 1.030$

$\hat{\beta}_3 = 0.600/1.265 = 0.474.$

Since these are real data I do not know what to do and how to estimate in, and, therefore, how to check my model. The example also raises the question on what to do when the likelihood function is maximizes (or equivalently some convenient distance measure is minimizes) at the boundary of the permissible range of the parameters.

Reference

Andersen, E. B. (1980): Latent Structure Analysis. A Survey, Research Report. Department of Statistics University of Copenhagen.

Harri Kiiveri (University of Western Australia)

Most of my comments will be concerned with covariance structures which arise from models with unobserved variables.

1. The Lisrel model without means can be written as

$$\begin{bmatrix} \mathbf{I} & \mathbf{0} & -\Lambda_y & \mathbf{0} \\ & \mathbf{I} & \mathbf{0} & -\Lambda_x \\ & \mathbf{0} & \mathbf{B} & -\Gamma \\ & & & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{Y} \\ \mathbf{X} \\ \eta \\ \xi \end{bmatrix} = \begin{bmatrix} \epsilon \\ \delta \\ \zeta \\ \xi \end{bmatrix}$$

i.e. $\mathbf{U}^T\mathbf{Z} = \mathbf{e}$

and $$\text{Cov}(\mathbf{e}) = \begin{bmatrix} \Theta_\varepsilon & & 0 \\ & \Theta_\delta & \\ & & \psi \\ 0 & & \Phi \end{bmatrix}$$

$$= \mathbf{A} \text{ (say)}.$$

Assuming that **B** and **A** are nonsingular and denoting the covariance matrix of **Z** by $\mathbf{\Sigma}$ we get:

$$\mathbf{\Sigma}^{-1} = \mathbf{U}\mathbf{A}^{-1}\mathbf{U}^T.$$

If all the variables were observed and had a joint Gaussian distribution, it is not difficult to see that the structure on $\mathbf{\Sigma}^{-1}$ gives rise to the following factorization of the joint density.

$$f(\mathbf{z}) = g_0(\xi)g_1(\mathbf{x}, \xi)g_2(\eta, \xi)g_3(\mathbf{y}, \eta)$$

which can be put into the form

$$f(\mathbf{z}) = f_0(\xi)f_1(\mathbf{x}|\xi)f_2(\eta|\xi)f_3(\mathbf{y}|\eta)$$

where the terms on the right hand side are conditional densities. This will be abbreviated as

$$f = (\xi)(\mathbf{x}|\xi)(\eta|\xi)(\mathbf{y}|\eta).$$

Depending on the structure in **A** and **U** further factorization is possible. For example, the model for study of stability of alienation with correlated "errors" has a structure which gives

$$f = (\xi)(x_1|\xi)(x_2|\xi)(\eta_1|\xi)(\eta_2|\eta_1\xi)(y_1y_3|\eta_1\eta_2) \times (y_2y_4|\eta_1\eta_2).$$

Such factorizations can be described by means of graphs and rules given for reading off the resulting conditional independences (Kiiveri & Speed, 1980).

There is also a connection with other conditional independence models for contingency tables and Gaussian data (Speed, 1978); Wermuth, 1976; Dempster, 1972). Viewing the Lisrel model in this light gives a means of extending the model to other types of data (i.e. discrete or mixtures of discrete and continuous). For example, the discrete version of a simple test score model for tests $X_1, \ldots, X_p$ (having levels $r_1, r_2, \ldots, r_p$ respectively) assumed to be measuring an underlying trait τ might be specified as

$$p(X_1 = x_1, \ldots, X_p = x_p, \tau = t) = p_0(t) \prod_{i=1}^{p} p_i(X_i = x_i|t)$$

where τ could be discrete or continuous. This type of model for discrete data has been discussed by Goodman (1973, 1974).

2. The added complication of unobserved variables could be handled using incomplete data theory (Sundberg, 1974; Dempster et al., 1977). The complete data would be

$$\begin{bmatrix} \mathbf{S}_{YY} & & & \\ \mathbf{S}_{XY} & \mathbf{S}_{XX} & & \\ \mathbf{S}_{\eta Y} & \mathbf{S}_{\eta X} & \mathbf{S}_{\eta\eta} & \\ \mathbf{S}_{\xi Y} & \mathbf{S}_{\xi X} & \mathbf{S}_{\xi\eta} & \mathbf{S}_{\xi\xi} \end{bmatrix} \text{ with } \begin{bmatrix} \bar{\mathbf{Y}} \\ \bar{\mathbf{X}} \\ \bar{\eta} \\ \bar{\xi} \end{bmatrix}$$

if the means were included. Only $\mathbf{S}_{YY}, \mathbf{S}_{XY}, \mathbf{S}_{XX}\bar{\mathbf{X}}, \bar{\mathbf{Y}}$ are observed. Function evaluations and derivative calculations can be done in a nice way using incomplete data theory and expressions for 2nd derivatives can also be given. The *E–M* algorithm (Dempster et al., 1977) although slow can be used to find a local maximum of the likelihood function of the *observed* data even if all the parameters are not identified. However it would be quicker to use this to produce initial values for some other algorithm.

3. From the examples given by Professor Jöreskog it appears that the one to oneness of the mapping $\boldsymbol{\theta} \to \mathbf{\Sigma}(\boldsymbol{\theta})$ must be checked on a case by case basis or in a crude way numerically. It would be nice to have a more general procedure and perhaps the matrix calculus of (McDonald & Swaminathan, 1973) along with incomplete data theory can be exploited in this direction. A numerical determination of the rank of $\partial\mathbf{\Sigma}/\partial\boldsymbol{\theta}$ can determine a set of parameters which need to be fixed to obtain (local) identifiability (McDonald & Krane, 1979.)

4. Finally I would like to touch on some theoretical points. There do not seem to be any existence or uniqueness results for the maximum likelihood estimates, and one wonders whether the special sort of incompleteness and structure could be used to produce some results in a curved exponential family context.

It seems as if testing a model should be possible without arbitrarily fixing unidentified parameters, since any parameter in the equivalence class of the maximum point gives the same value of the likelihood. Also I am not clear on the effect of the particular fixed values chosen on the relative magnitudes of the parameters or the standard errors.

In closing I would like to thank Professor Jöreskog for his survey and particularly for his examples on the application of the theory to data. I like the trick used to estimate means in Lisrel and wonder what other "devious" means may be employed to fit seemingly "non-Lisrel" models.

References

Darroch, J. N., Lauritzen, S. L. & Speed, T. P. (1980). Log-linear models for contingency tables and Markov fields over graphs. *Annals of Statistics* **8** (to appear).

Dempster, A. P. (1972). Covariance selection. *Biometrics* **28**, 157–175.

Dempster, A. P., Laird, N. M. & Rubin, P. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. Ser. B.* **39**, 1–21.

Goodman, L. A. (1973). The analysis of multidimensional contingency tables when some variables are posterior to others: a modified path analysis approach. *Biometrika* **60**, 179–192. G

Goodman, L. A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I—a modified latent structure approach. *American Journal of Sociology* **79**, 1179–1259.

Kiiveri, H. T. & Speed, T. P. (1980). The structural analysis of multivariate data: a review. Submitted to *Sociological methodology*.

McDonald, R. P. & Krane, W. R. (1979). A Monte-Carlo study of local identifiability and degrees of freedom in the asymptotic likelihood ratio test. *British Journal of Mathematical and Statistical Psychology* **32**, 121–132.

McDonald, R. P. & Swaminathan, H. (1973). A simple matrix calculus with applications to multivariate analysis. *General Systems* **18**, 37–54.

Speed, T. P. (1978). Relations between models for spatial data, contingency tables and Markov fields on graphs. *Proceedings of the conference on spatial patterns and processes* (ed. R. L. Tweedie). Supplement to *Advances in Applied Probability*.

Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. *Scand. J. Statist.* **1**, 49–58.

Wermuth, N. (1976). Analogies between multiplicative models in contingency tables and covariance selection. *Biometrics* **32**, 95–108.

Wermuth, N. (1980). Linear recursive equations, covariance selection and path analysis. *J. Amer. Statist. Assoc.* (to appear).

Petter Laake (University of Oslo)

1. First of all I want to thank Professor Jöreskog for his inspiring lectures on application and estimation in covariance structures. They certainly contained a lot of interesting and useful ideas. I have got my interest and experience in covariance structure analysis from cooperation with social scientists, and my discussion is probably influenced by that.

2. There can be no doubt that measurements in the social sciences may contain measurement errors. The statistical model ought to take this into consideration, especially, of course, when the parameters of interest are related to the true variables. Models which relate every measurement with its true, latent, variable and an error term, and relate the latent variables in a structural form are extremely useful. This is for instance the case in the quasi-simplex models. Then, the interpretation of the results is easy since the units of measurement in the observed variable are the same as in the latent variable, and there is no question of validity. This is not necessarily the case in the more advanced covariance structures.

To be more specific I will comment on one of Professor Jöreskog's examples, namely Example 3 which is an application of the LISREL-model. The model specifies the relationship between latent variables (socioeconomic status (SES) and alienation) and relates the latent variables and the measurements in a factor analysis model. The parameters of main interest are those of the relation between the latent variables, namely the parameters describing the influence of SES on alienation in 1967 and 1971 and the parameter describing the influence of alienation 1967 on alienation 1971. Since all the relations are given explicitly the analysis seems to be confirmatory or in statistical terms that we want to estimate parameters of a given structural relationship. Then, the social scientists will be interested in the estimated size of the parameters regardless of their statistical significance. This is so because the parameters are included in the model according to our interest in them and because the question of statistical significance may be a question of a sufficiently large sample size. To interpret the estimated size of the parameter we usually relate it to the scale of the variables and their distributions and not to the standard error of the estimate. But latent variables are in general not related to any scale. So how do we interpret some of the main results of Example 3, for instance that the coefficients for SES on alienation 1967 and 1971 are—0.575 and—0.227, respectively? One possibility is to make the scale of the latent variable equal to the one of a measurement by fixing one element of the factor loading matrix equal to 1. This solves the problem of identification, but the problem of interpretation is not fully solved since the latent variable still influences other measurements. Another possibility, which could be studied in more detail, is to use the factor model for estimation of factor scores (see for instance Lawley & Maxwell (1971, p. 106)). Then, the interpretation of the coefficients can be related to the empirical distribution of the factor scores. Further, it is possible to relate the coefficients directly to the measurements through the factor model.

Professor Jöreskog's opinion on the application of covariance structure analysis seems to be that they most successfully have been applied to confirmatory analysis, but that they also can be used for exploratory analysis (Jöreskog, p. 75). I certainly agree with him that for instance the LISREL—model is useful when a well-established theory of the problem exists. This is so because we in some way must have demonstrated the validity of the factor model, for instance by earlier use of the latent variables in similar context. If the validity fails, the analysis of the structural relationship among the latent variables may be meaningless. On the other hand it is

indicated that LISREL may be used for exploratory analysis. This use of it may be doubtful. Let us say that the relationship among the latent variables is given, but for a given vector of measurements we are free to choose the factor model. It is well known that every orthogonally rotated factor loading matrices are statistically identical with regard to the model. The interpretation of the factors and the coefficients of the main relationship will, however, depend upon the chosen factor loading matrix. This can make it difficult for the social scientist to get an understanding of the structure of the measurements. Further, the factor model itself will not demonstrate the validity of the latent variables resulting from the exploratory analysis. To give a latent variable a name is, of course, not necessarily enough to make it useful.

Social scientists often deal with complex social theories and try to build complex statistical models. I believe that it not necessarily will give us an uninteresting or unrealistic analysis of a problem even if we disregard structural relationships which are known to exist. This may be especially relevant in an exploratory analysis, when the aim is to get an understanding of the structure of a given set of measurements. Interesting information that is easy to interpret may, for instance, be gained if multivariate regression analysis, which disregards measurement errors, is considered. If we go back to Example 3, it will be interesting to know if conclusions similar to those of Jöreskog are gained if a multivariate regression model is studied instead of the LISREL-model?

3. *Example: Predicting the labour force in Norway* (*Helgeland, 1980*). Next I will give an example of the use of a covariance structured model related to the simplex model which is being studied at the Central Bureau of Statistics of Norway. To be specific let η_t be the unobservable labour force at time t. In the Bureau η_t has to be estimated by a sample survey. Let therefore

$$y_t = \eta_t + \varepsilon_t,$$

where ε_t is the sampling error, be the estimated labour force. Var ε_t can be found using the design of the sample survey. We now assume that $\{\eta_t\}$ is a ARMA-process

$$\eta_t = \sum_{i=1}^{p} \beta_i \eta_{t-i} + \sum_{j=1}^{q} \alpha_j a_{t-j},$$

with $\alpha_0 = 1$ and a_t's independent. The observations are the estimated labour force $y_1, y_2, \ldots, y_T$. Since the units of measurement in the latent variable η is the same as in the observed variable y, the interpretation is straightforward. The aim of the study is not only estimate the α's and β's, but to predict

$$E\{\eta_T | y_1, y_2, \ldots, y_T\}.$$

This is actually a LISREL-specified model, and numerical methods similar to those described by Jöreskog can be used for estimation. The asymptotic theory does not carry over, however, since no replications are made at time t. It is easy to imagine that models like this may be relevant in Central Bureaus of Statistics.

4. I have now a few more technical comments on Jöreskog's lectures. An important problem for the applicability of the LISREL specification is the efficiency of the numerical technique for maximum likelihood estimation of the parameters. It is well known that the efficiency depends upon the start value for iteration, but we may never know if we actually have reached a global maximum of the likelihood function. It is easy to imagine situations where this may be difficult, especially when the sample sizes are small and multiple solutions to the likelihood equations exist. In situations like these I think it may be a good idea to try different start values for iteration and hope for every iteration to give the same solution.

Another problem is the one of restricted parameter estimation. Some parameters of the covariance structure are known to be greater or equal to zero. A numerical technique can in general produce an iterate which lies outside the constraint space. When for instance a variance estimate tends to zero the approach is usually to fix the variance at a given level during the following iterations. This may be a questionable procedure since it may cause the procedure to converge to a point that is not even a local maximum (Bard, 1974, p. 151). Thus, the problem is whether it is possible to consider numerical techniques that take the constraint into consideration and still iterate efficiently.

5. There exist alternatives for parameter estimations in the models. The most commonly used seem to be maximum likelihood estimation (ML) and generalized least squares estimation (GLS). I certainly agree with Professor Jöreskog in that GLS is computationally easier and thus probably more attractive from a practical point of view. They both produce asymptotically efficient estimates. Both Jöreskog & Goldberger (1972) and later Browne (1974) report numerical results that could indicate that there is a bias in the GLS-estimates compared to the ML-estimates. This could indicate that GLS-estimates really are biased in small samples or that they may have larger variances. To make the compu-

tationally more attractive GLS-method even more attractive to the user I think their small samples properties ought to be studied in some more detail.

6. The choice of a model is always a difficult problem. In a confirmatory analysis the model fitting is to me of minor interest and the only relevant testing seems to be the one of testing for correlation in residuals to detect any possible shortcoming of the model. In an exploratory analysis the model fitting is important and controversial. Professor Jöreskog uses the likelihood ratio test in his model choice. The properties of the test may be questionable. The test is likely, for example, even in large samples, to be sensitive to departures from normality. This is in my opinion not necessarily an important objection since the teststatistic is mainly used to study the drop in fit when alternative models are introduced. I think, however, there may be some objections to it from a practical point of view.

Firstly, a straightforward use of the likelihood ratio test gives preference to models with many parameters. A model with many parameters will give a nice fit, which may be of importance to the statistician, but not necessarily to the social scientist who then will get a more complex structure to explain. Thus, I think, some kind of penalty function to keep the number of parameters at a low level could be introduced in addition to the maximum likelihood function. This is done by for instance Hannan & Quinn (1979) and Schwarz (1978), mainly for application in time series analysis. The procedures can, however, be used for any model choice when maximum likelihood estimation is used.

Secondly, the likelihood ratio test is not diagnostic which may be an important drawback to the use of it. When we reject a model we usually want to know where it is misspecified. In Example 3 we will, for instance, probably observe that the elements of $(\mathbf{S}-\hat{\boldsymbol{\Sigma}})$ at the position of the covariances of y_1, y_3 and y_2, y_4 are large. Thus, without predistinguishing between the models A and B, the analysis of the residuals will guide us to respecify to model B. To conclude, I think other test statistic which may be based on the residuals could contain more information on why models are rejected.

7. Finally, I will once more thank Professor Jöreskog for his fine overview. In my daily work I find it extremely useful that Professor Jöreskog has cleared up concepts which were in actual use in sociology and psychology and expressed them in a language natural to the statisticians. Together with his work on estimation methods, Professor Jöreskog has certainly made the covariance structured models valuable for statistical analysis in the social sciences.

References

Bard, Y. (1974). *Nonlinear parameter estimation.* Academic Press, New York.

Browne, M. W. (1974). Generalized least squares estimators in the analysis of covariance structures. *South African Statistical Journal* **8**, 1–24.

Jöreskog, K. G. & Goldberger, A. S. (1972). Factor analysis by generalized least squares. *Psychometrika* **37**, 243–260.

Hannan, E. J. & Quinn, B. G. (1979). The determination of the order of an autoregression. *J. Roy. Stat. Soc. B* **41**, 190–195.

Helgeland, J. (1980). Unpublished manuscript. Central Bureau of Statistics of Norway.

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

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.

D. R. Cox (Imperial College, London)

I congratulate Professor Jöreskog both on this impressively lucid and comprehensive paper and also on the important body of work which it summarizes.

Three questions follow. Have Bartlett correction factors been evaluated for any of the likelihood ratio tests involved? Such corrections may appreciably improve the large-scale chisquared approximations. Can marginal likelihoods be found for any of the problems with structure in the means as well as in the covariances? Has Professor Jöreskog suggestions on how to examine whether information on dependency is adequately summarized in a covariance matrix? The suggestion of Cox & Small (1978) for finding two derived directions corresponding to maximum curvature seems reasonable although its implementation is cumbersome computationally.

If the full data for Example 1 became available the possibility would arise of an analysis based on standard factorial contrasts and their interaction with subjects.

Reference

Cox, D. R. & Small, N. J. H. (1978). Testing multivariate normality. *Biometrika* **65**, 263–272.

Tore Schweder (University of Tromsø)

The following comment is not ment to reduce the value of the very clear and informative exposition Jöreskog has given of the analysis of covariance structure. My intention is actually not to focus on the analysis of covariance structure as such, but rather to fix the attention for a while on the difficult concept of "causality" which you have touched upon repeatedly.

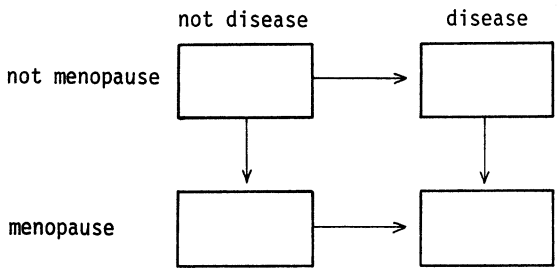

Fig. 1

The causal relationships which are modelled in path analysis may be called static as opposed to dynamic causal relationships. The causal variables in your models have for each individual a fixed value. The variation is over individuals, not also over time. And your concern is in some sense to estimate the correlation between the cause variable and the effect variable. Typical such static causal variables are sex and intelligense, the latter being a latent variable.

This static concept of causality is quite limited. In ordinary language, causality is often used in a more dynamic sense. And also in a sense more open for human action. The type of concept I have in mind is exemplified in the statement "smoking does cause lung cancer", rephrased "the probability of getting lung cancer increases when smoking is increased". It is hard to change your own sex or intelligence, but your level of smoking should be a matter of choice.

Such a dynamic concept of causality must be defined in the framework of stochastic processes. This has been done by Aalen et al. (1979) in their study of the relationship between menopause and the skin disease pustulosis planetaris. They modeled causality in terms of the transition intensities in the two-dimensional stochastic process with state space indicated in Fig. 1.

Menopause being a cause of the disease is defined as the transition intensity of getting the disease being dependent on having had menopause or not.

As opposed to correlation, the concept of transition intensity—or drift in a continuous state space—is defined in a time context, and is a non-symmetric concept. When the situation may be modelled as a multivariate stochastic process—or rather one separate multivariate stochastic process for each individual of the population, it may be possible to formulate various hypothesis of causation in terms of the component transition intensities. In Schweder (1970) the concept of local dependence was introduced for this purpose. Statistical methodology has been developed by Aalen et al. (1979) for the two-dimensional situation mentioned above, and may be further developed along those lines.

As mentioned above, my intention with this comment has been to focus attention on a conceptual framework and a statistical methodology for "causal analysis" which in many situations—I believe, is more appropriate than path analysis or other methods based on correlation.

References

Aalen, O. et al. (1980). Interaction between life history events. Nonparametric analysis for prospective and retrospective data in the presence of censoring. *Scand. J Statist.* **7**, 161–171.

Schweder, T. (1970), Composable Markov processes. *Journal of Applied Probability* **7** (2), 400–10.

Reply by K. G. Jöreskog

I want to thank all the discussants for their constructive comments on my paper. The discussants have raised several issues which are important in the practical applications of covariance structure analysis and I will therefore state my position on these issues here.

Model specification and purpose of analysis. The model should be based on a substantive theory or hypothesis, on a given design for known experimental conditions or on known results from previous studies. The model is supposed to explain or account for the jointly dependent variables and their interrelationships. The purpose of the analysis is to test the validity and adequacy of the model and to estimate its parameters. If the model is misspecified the analysis should indicate this and suggest how the model should be modified.

Identification. Before the model is estimated by means of data the identification of the model must be examined. Although the computer program LISREL IV can estimate models which are not identified, I recommend to deal with non-identified models by adding appropriate conditions so as to make all the parameters identified. In interpreting the results of an analysis it is necessary to have a complete understanding of the identification status of each parameter of the model. Since no general and practically useful necessary and sufficient conditions for identification are available for the general LISREL model, I suggest that the identification problem be studied on a case by case basis by examining the equations (11). General necessary and sufficient conditions for identification has been given for some special classes of LISREL models, namely for factor analysis models by Howe (1955), Dunn (1973), Jennrich (1978), see also Jöreskog & Sörbom (1979) pp. 40–43, for simultaneous structural equation

models, see e.g. Goldberger (1964) pp. 306–318 and for structural equation models with measurement errors by Geraci (1976). These conditions may be applied to the measurement model part (23) and (24) and to the structural equation model part (22) separately. It is a good idea to check these conditions.

For many users of LISREL IV the identification problem may be too difficult to resolve. There may be a tendency to run the model even though the identification of it is unclear. Therefore it is a good thing that the computer program checks the positive definiteness of the information matrix. One should be aware, however, that this check is not hundred percent reliable although experience indicates that it is nearly so. The check depends on the estimated point in the parameter space at which the information matrix is evaluated and on the numerical accuracy by which the information matrix is inverted.

One and the same model, whether it is completely identified or not, may have two or more equivalent parametrizations even within the class of LISREL models. Which parametrization to use is a matter of preference and is related to the interpretation of the result. One example of this is Professor Andersen's first example and it is instructive to consider this in some detail. As demonstrated elsewhere (Jöreskog, 1974, p. 52) the two models (5) and (9) are equivalent when the number of observed variables is four. To see this let me write (5) with λ's instead of β's and ψ's instead of θ's:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \lambda_1^2+\psi_1 & & & \\ \lambda_1\lambda_2 & \lambda_2^2+\psi_2 & & \\ \lambda_1\lambda_3\varrho & \lambda_2\lambda_3\varrho & \lambda_3^2+\psi_3 & \\ \lambda_1\lambda_4\varrho & \lambda_2\lambda_4\varrho & \lambda_3\lambda_4 & \lambda_4^2+\psi_4 \end{bmatrix}$$

This model is identified and has nine independent parameters, i.e., the model has one degree of freedom. The parameter values $\lambda_1 = 8.9$, $\lambda_2 = 9.2$, $\lambda_3 = 7.4$, $\lambda_4 = 10.4$, $\psi_1 = \psi_2 = \psi_3 = \psi_4 = 25.2$, $\varrho = 0.7$ given by Professor Andersen are therefore unique.

As explained in the paper, model (9) is not identified. The three parameters β_2, ω_1 and θ_1 are only determined by two equations:

$$\sigma_{11} = \omega_1 + \theta_1$$

$$\sigma_{21} = \beta_2\omega_1$$

and the two parameters ω_4 and θ_4 are only determined by the single equation:

$$\sigma_{44} = \omega_4 + \theta_4.$$

All the other parameters are uniquely identified. The parameter values given by Professor Andersen are to some extent arbitrary. For example the following set of parameter values $\beta_2 = 1.00$, $\beta_3 = 0.56$, $\beta_4 = 1.36$, $\theta_1 = 22.5$, $\theta_2 = \theta_3 = 25.2$, $\theta_4 = 10.0$, $\omega_1 = 81.9$, $\omega_2 = 84.6$, $\omega_3 = 54.8$, $\omega_4 = 117.2$ will also reproduce identically the same $\boldsymbol{\Sigma}$.

That the two models are equivalent may be seen for example by fixing β_2 at one and absorbing θ_4 into ω_4. This gives the covariance structure

$$\boldsymbol{\Sigma} = \begin{bmatrix} \omega_1+\theta_1 & & & \\ \omega_1 & \omega_2+\theta_2 & & \\ \beta_3\omega_1 & \beta_3\omega_2 & \omega_3+\theta_3 & \\ \beta_3\beta_4\omega_1 & \beta_3\beta_4\omega_2 & \beta_4\omega_3 & \omega_4 \end{bmatrix}$$

Then it may be easily verified that there is a one-two-one relation between the two sets of parameters, assuming, of course, that the variances ω_i, $i = 1, 2, 3, 4$, are positive:

$$\beta_3 = \lambda_3\varrho/\lambda_2,\ \beta_4 = \lambda_4/\lambda_3,\ \omega_1 = \lambda_1\lambda_2,\ \omega_2 = \lambda_2^2,\ \omega_3 = \lambda_3^2,\ \omega_4 = \lambda_4^2+\psi_4,\ \theta_1 = \lambda_1^2+\psi_1-\lambda_1\lambda_2,\ \theta_2 = \psi_2,\ \theta_3 = \psi_3$$

and

$$\lambda_1 = \omega_1/\sqrt{\omega_2},\ \lambda_2 = \sqrt{\omega_2},\ \lambda_3 = \sqrt{\omega_3},\ \lambda_4 = \beta_4\sqrt{\omega_3},$$

$$\varrho = \operatorname{sgn}(\beta_3)(\omega_2|\beta_3|/\omega_3)^{\frac{1}{2}},\ \psi_1 = \omega_1+\theta_1-\omega_1^2/\omega_2,\ \psi_2 = \theta_2,$$

$$\psi_3 = \theta_3,\ \psi_4 = \omega_4 - \beta_4^2\omega_3$$

It should be noted that in both models the single overidentifying constraint is that $\sigma_{41}/\sigma_{31} = \sigma_{42}/\sigma_{32}$.

Which of the two parametrizations is to be preferred? If one is interested in the correlation ϱ between two latent variables then model 1 is the one. On the other hand, if one is interested in the autogregression coefficients β_3 and β_4 then model 2 is the one. The reason why these two seemingly different models are equivalent in this case is that the quasi-Markov-simplex model is not really meaningful until one has at least five time points. Since there is an indeterminacy associated with the first and the last variable one must have at least three variables in between these to be able to test the first-order Markov property.

Checking the data. In addition to the structural assumptions imposed by the model specification the analysis makes the ususal assumptions of linearity and additivity, independence of observtions and homoskedasticity of error terms. Another crucial question is whether second-order information is sufficient. Usually one needs to assume that the

observed variables are approximately multinormal. There is one important exception to this, however, namely when the x-variables are fixed variables. Then these variables may be any variables including dummy categorical variables. We are then only concerned with conditional distribution of **y** for given **x** which should be approximately multinormal. Before estimating the model, the assumption of multinormality should be assessed, including a check on outliers. If the observed distributions deviate far from the multinormal I recommend that the variables be transformed and/or that robustified estimates of variances and covariances be used. For methods of assessing multinormality and computing robust estimates see e.g. Gnanadesikan (1977).

Checking the model. The first and most obvious test of the model is whether the parameter estimates are reasonable. If some parameters fall outside the admissible range, either the model is fundamentally wrong or the data is not informative enough. An example of this is Professor Andersen's second example and I shall consider this below.

The fit of the model may be assessed by various means. One is the use of the overall χ^2-measure and its associated degrees of freedom and probability level. However this does not indicate which part of the model is wrong. A more detailed assessment of fit can be obtained by an inspection of the normalized residual covariances. The normalized residual covariance is $s_{ij}-\hat{\sigma}_{ij}$ divided by the square root of its asymptotic variance which is estimated as $(\hat{\sigma}_{ii}\hat{\sigma}_{jj}+\hat{\sigma}_{ij})/N$. A Q-plot of these residuals gives a very effective summary. Residuals which are larger than two in magnitude are indicative of a specification error in the model and the corresponding indices i and j usually give a hint as to where this error is.

Another alternative is to look at the first-order derivatives of the fitting function with respect to the fixed and constrained parameters, these being the Lagrangian multipliers corresponding to the constraints of the model, see Sörbom (1975). One problem with this approach, however, is that these derivatives depend on the magnitude of the data and the parameter values. In principle, one must assess each derivative in relation to its standard error. But the computation of these standard errors requires the inversion of a matrix of the order of all elements in all parameter matrices and this is obviously not feasible in most cases. A reasonable compromise may be to consider the ratio of the first-order derivative to the corresponding second-order derivative. The fixed parameter corresponding to the largest such index is probably the one which when relaxed will improve fit maximally. I recommend this procedure only when relaxing this parameter makes sense from a substantive point of view and when the values of this parameter can be clearly interpreted.

Multiple solutions and non-admissible solutions. It may happen that there are several local minima of the fitting function. The only way to avoid this is to have a model which is appropriate for the data and a large random sample. Experience indicates, however, that multiple solution seldom occur, and when they do, it is usually with solutions on the boundary of or outside the admissible parameter space, as in Professor Andersen's second example The computer program does not constrain variances to be positive, correlations to be less than one in magnitude etc. The only constraint imposed by the program is that the matrix $\boldsymbol{\Sigma}$ reproduced by the model is positive definite. Apart from this there is nothing that prevents the program from going outside the admissible parameter space. Although constrained estimation would be possible, the minimization algorithm for this would be much more time-consuming even in the case when the solution is admissible. In my opinion, this would not be worth while. If the solution is in the interior of the parameter space one will just spend more computer time to find it. If, on the other hand, the solution is inadmissible, the current program will find the solution outside the parameter space whereas a program which uses constrained estimation will find the solution on the boundary of the parameter space. In both cases the conclusion will be that the model is wrong or that the sample size is too small.

Using LISREL it is often possible, though not always, to use various tricks to force the program to stay within the admissible parameter space. Let me demonstrate this using Professor Andersen's second example. This covariance structure is generated by the following simplex model

$$y_t = \eta_t + \varepsilon_t, \quad t = 1, 2, 3 \tag{1}$$

$$\eta_t = \beta_t \eta_{t-1} + \zeta_t, \quad t = 2, 3 \tag{2}$$

with $\mathrm{Var}(\eta_t) = \omega_t$ and $\mathrm{Var}(\varepsilon_t) = \theta$, $t = 1, 2, 3$. Let $\psi_t = \mathrm{Var}(\zeta_t)$. Then there is a one-to-one correspondence between (ψ_1, ψ_2, ψ_3) and $(\omega_1, \omega_2, \omega_3)$ $(\psi_1 = \omega_1, \psi_2 = \omega_2 - \beta_2^2\omega_1, \psi_3 = \omega_3 - \beta_3^2\omega_2)$ so I will use ψ_t instead of ω_t, $t+1, 2, 3$. The trick is to write (1) and (2) for LISREL as

$$\begin{pmatrix} y_1 \\ y_3 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ -\beta_2 & 1 & 0 & 0 & 0 & 0 \\ 0 & -\beta_3 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix}$$

$$= \begin{pmatrix} \sqrt{\psi_1} & 0 & 0 & 0 & 0 & 0 \\ 0 & \sqrt{\psi_2} & 0 & 0 & 0 & 0 \\ 0 & 0 & \sqrt{\psi_3} & 0 & 0 & 0 \\ 0 & 0 & 0 & \sqrt{\theta} & 0 & 0 \\ 0 & 0 & 0 & 0 & \sqrt{\theta} & 0 \\ 0 & 0 & 0 & 0 & 0 & \sqrt{\theta} \end{pmatrix} \xi + \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

corresponding to (27) with Cov $(\xi) = I$. LISREL can then be used to estimate β_2, β_3, $\sqrt{\psi_1}$, $\sqrt{\psi_2}$, $\sqrt{\psi_3}$ and $\sqrt{\theta}$. The square roots may actually come out negative or zero but, when squared, these yield maximum likelihood estimates of ψ_1, ψ_2, ψ_3 and θ. The maximum likelihood estimates of ω_t are then computed as $\hat{\omega}_1 = \hat{\omega}_1$, $\hat{\omega}_2 = \hat{\beta}_2^2\hat{\psi}_1 + \hat{\psi}_2$ and $\hat{\omega}_3 = \hat{\beta}_3^2\hat{\psi}_2 + \hat{\psi}_3$. The actual values come out as $\hat{\omega}_1 = 0.792$, $\hat{\omega}_2 = 1.26$, $\hat{\omega}_3 = 1.03$, $\hat{\beta}_2 = 0.758$, $\hat{\beta}_3 = 0.476$, $\hat{\theta} = 0.00$ so that Professor Andersen's conjecture is indeed correct.

Causality. I agree with Dr Schweder that causality cannot be inferred from just any correlational study. On the other hand, I would not go so far as to say that causal effects can only be demonstrated by means of carefully controlled experiments. In the social sciences, notably in econometrics, causal inference is obtained by means of tightly specified models. In most cases these models are dynamic and estimated from time-series or cross-sectional data. Such models can also be estimated by LISREL. For an example where both time-series and cross-sectional data are used, see Jöreskog (1978).

Other models and methods. As pointed our by Dr Kiiveri there are alternative model formulations which are both more general and more compact in the sense of requiring fewer parameter matrices. While such an approach is mathematically elegant and makes it possible to include models for categorical variables in the same framework, I think, the computational algorithm for such an approach will be more extensive since it will have to deal with very large matrices. The *E–M* algorithm provides a mean for iterating in very simple steps but my experience with the *E–M* algorithm for factor analysis estimation is that it requires an extremely large number of iterations to converge.

References

Dunn, J. E. (1973). A note on a sufficiency condition for uniqueness of a restricted factor matrix. *Psychometrika* **38**, 141–143.

Geraci, V. J. (1976). Identification of simultaneous equation models with measurement error. *Journal of Econometrics* **4**, 263–283.

Gnanadesikan, R. (1977). *Methods for statistical data analysis of multivariate observations*. Wiley, New York.

Goldberger, A. S. (1964). *Econometric theory*. Wiley, New York.

Howe, H. G. (1955). Some contributions to factor analysis. Report No. ORNL-1919. Oak Ridge National Laboratory, Oak Ridge, Tenn.

Jennrich, R. I. (1978). Rotational equivalence of factor loading matrices with specified values. *Psychometrika* **43**, 421–426.

Jöreskog, K. G. (1974). Analyzing psychological data by structural analysis of covariance matrics. In *Contemporary developments in mathematical psychology, Vol. II* (ed. R. C. Atkinson, D. H. Krantz, R. D. Suppes), pp. 1–56. W. H. Freeman & Col., San Francisco.

Jöreskog, K. G. (1978). An econometric model for multivariate panel data. *Annales de l'INSEE*, No. 30–31.

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

Sörbom, D. (1975). Detection of correlated errors in longitudinal data. *British Journal of Mathematical and Statistical Psychology* **28**, 138–151.