DOCUMENT RESUME

ED 053 169		TM 000 688

AUTHOR Joreskog, Karl G.; Goldberger, Arthur S.
TITLE Factor Analysis by Generalized Least Squares.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO RB-71-26
PUB DATE May 71
NOTE 34p.

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS Correlation, *Factor Analysis, Factor Structure, Goodness of Fit, *Mathematical Models, *Mathematics, Probability, *Statistics, *Transformations (Mathematics)
IDENTIFIERS *Generalized Least Squares, GLS

ABSTRACT

Aitkin's generalized least squares (GLS) principle, with the inverse of the observed variance-covariance matrix as a weight matrix, is applied to estimate the factor analysis model in the exploratory (unrestricted) case. It is shown that the GLS estimates are scale free and asymptotically efficient. The estimates are computed by a rapidly converging Newton-Raphson procedure. A new technique is used to deal with Heywood cases effectively. (Author)

RB-71-26

FACTOR ANALYSIS BY GENERALIZED LEAST SQUARES

Karl G. Jöreskog
Educational Testing Service

and

Arthur S. Goldberger
University of Wisconsin

Educational Testing Service
Princeton, New Jersey
May 1971

FACTOR ANALYSIS BY GENERALIZED LEAST SQUARES

Karl G. Jöreskog

Educational Testing Service

and

Arthur S. Goldberger

University of Wisconsin

Abstract

Aitkin's generalized least squares (GLS) principle, with the inverse of the observed variance-covariance matrix as a weight matrix, is applied to estimate the factor analysis model in the exploratory (unrestricted) case. It is shown that the GLS estimates are scale free and asymptotically efficient. The estimates are computed by a rapidly converging Newton-Raphson procedure. A new technique is used to deal with Heywood cases effectively.

FACTOR ANALYSIS BY GENERALIZED LEAST SQUARES

Karl G. Jöreskog*

Educational Testing Service

and

Arthur S. Goldberger*

University of Wisconsin

1. Introduction

Consider the factor analysis model,

$$(1) \qquad y = \Lambda f + u \quad ,$$

where y is a $p \times 1$ vector of observable random variables, Λ is a $p \times k$ matrix of unknown factor loadings, f is the $k \times 1$ vector of unobservable common factors and u is a $p \times 1$ vector of unobservable unique factors or residuals. It is assumed that $\mathcal{E}(f) = 0$, $\mathcal{E}(ff') = I$, $\mathcal{E}(u) = 0$ and $\mathcal{E}(uu') = \psi^2$, where ψ^2 is a diagonal matrix. It is further assumed u and f are uncorrelated. (For convenience, a mean vector has been suppressed in (1)). From these assumptions it follows that the variance-covariance matrix Σ of y is

$$(2) \qquad \Sigma = \Lambda\Lambda' + \psi^2 \quad .$$

The force of the model when k is small relative to p lies in the constraints it imposes on this variance-covariance matrix: the $r = \frac{1}{2}p(p + 1)$

---

*The first author is Research Statistician at Educational Testing Service, Princeton, N.J. The second author is Professor of Economics at the University of Wisconsin. Work on this project was in part supported by a grant from the Research Committee of the University of Wisconsin Graduate School. The authors wish to thank Michael Browne for many helpful comments and Marielle van Thillo for valuable assistance in the numerical computations. This paper is also being distributed in the Workshop Paper Series of the Social Systems Research Institute.

distinct elements of Σ are expressed in terms of the $(k + 1)p$ unknown parameters in Λ and ψ^2. Since Λ in (1) may be postmultiplied by an arbitrary $k \times k$ orthogonal matrix without changing Σ, Λ may be chosen to satisfy $\frac{1}{2} k(k - 1)$ independent conditions. Thus, the effective number of unknown parameters are $s = (k + 1)p - \frac{1}{2} k(k - 1)$ and the degrees of freedom of the model is

(3) $$d = r - s = \frac{1}{2} [(p - k)^2 - (p - k)] \quad .$$

Let S denote the $p \times p$ sample variance-covariance matrix of y with n degrees of freedom obtained in a random sample of size $n + 1$. The estimation problem of factor analysis is to use S to develop estimates of Λ and ψ^2. The factor analysis literature contains alternative estimation procedures, many of which amount to choosing Λ and ψ^2 to make Σ close to S in some sense [cf. Anderson, 1959, pp. 19-22]. Let $\Phi(S,\Sigma)$ be a scalar measure of the distance between S and Σ to be minimized with respect to Λ and ψ. It is convenient to normalize Φ so that $\Phi = 0$ when $S = \Sigma$. A desirable property for Φ is that

$$\Phi(S,\Sigma) = \Phi(DSD,D\Sigma D)$$

for all diagonal matrices D of positive scale factors. Such a Φ will yield estimates that are scale-free.

One simple measure Φ is the unweighted sum of squares

(4) $$U = tr(S - \Sigma)^2 \quad .$$

This measure, which is minimized by the iterated principal factor method and the minres method [Harman, 1967, Chapters 8 and 9], is not scale-free and is therefore usually applied to the correlation matrix R instead of S. Another measure Φ is the function employed in maximum likelihood (ML) factor analysis [see e.g., Jöreskog, 1967]:

$$(5) \qquad F = \mathrm{tr}(\Sigma^{-1}S) - \log|\Sigma^{-1}S| - p \quad .$$

This measure is scale-free and, when y is multinormally distributed, leads to efficient estimates in large samples.

In this paper, we propose an estimation procedure which calls for minimization of the quantity

$$(6) \qquad G = \frac{1}{2}\,\mathrm{tr}(I - S^{-1}\Sigma)^2 \quad .$$

This yields a scale-free method and when normality is assumed produces estimates which have the same asymptotic properties as the maximum likelihood estimates.

2. Generalized Least Squares Principle

The background for our proposal is as follows. Assuming that y is multinormally distributed, S has the Wishart distribution with expectation Σ_0, where Σ_0 is the true population variance-covariance matrix of y. Therefore, a straigntforward application of Aitken's [1934-35] generalized least squares principle would choose parameter estimates to minimize the quantity

$$(7) \qquad \tilde{G} = \frac{1}{2}\,\mathrm{tr}[\Sigma_0^{-1}(S - \Sigma)]^2 \quad .$$

In practice, of course, Σ_0 is unknown, so that the Aitken procedure is not operational. Nevertheless S estimates Σ_0. Using the estimate S in place of Σ_0 in (7) gives

$$(8) \qquad G = \frac{1}{2}\,\mathrm{tr}[S^{-1}(S - \Sigma)]^2 = \frac{1}{2}\,\mathrm{tr}(I - S^{-1}\Sigma)^2 \quad ,$$

which is the criterion to be minimized in our modified generalized least squares (GLS) procedure.

There is an interesting connection between the ML criterion (5) and the GLS criterion (6). Let $a_1,\ldots,a_p$ denote the characteristic roots of $S^{-1}\Sigma$; they will be positive and, when S is close to Σ , lie in the neighborhood of unity. Since the trace and determinant are respectively the sum and product of the roots, we see that

$$(9) \qquad F = \sum_{m=1}^{p} (1/a_m) + \log \prod_{m=1}^{p} a_m - p = \sum_{m=1}^{p} (1/a_m - 1 + \log a_m) \quad .$$

The characteristic roots of $I - S^{-1}\Sigma$ are $1 - a_1,\ldots,1 - a_p$, so that those of $(I - S^{-1}\Sigma)^2$ are $(1 - a_1)^2,\ldots,(1 - a_p)^2$. Consequently

$$(10) \qquad G = \frac{1}{2}\sum_{m=1}^{p} (1 - a_m)^2 \quad .$$

Expanding $1/a_m$ and $\log a_m$ in a Taylor series about the point $a_m = 1$ and discarding terms of order higher than the second gives

$$1/a_m \approx 1 - (a_m - 1) + (a_m - 1)^2 \quad ,$$

$$\log a_m \approx (a_m - 1) - \frac{1}{2}(a_m - 1)^2 \quad .$$

Thus

$$(11) \qquad F \approx \frac{1}{2}\sum_{m=1}^{p}(a_m - 1)^2 = G \quad ,$$

so that the ML criterion can be viewed as an approximation to the GLS criterion.

Our proposal derives from Zellner's [1962] operational approach to generalized least squares estimation in multivariate regression models with linear constraints on the regression coefficients. Malinvaud [1966, Chapter 9] extends the approach to cover nonlinear constraints on the regression coefficients. Rothenberg [1966, p. 38] indicates a further extension to cover constraints on the disturbance variance-covariance matrix. For factor analysis with known factor loadings, Browne [1970] suggests using weighted least squares with S estimating Σ_0. Ultimately, all these procedures are applications of the minimum- χ^2 principle of estimation; cf. Neyman [1949], Taylor [1953], Ferguson [1958].

The GLS principle can be used in confirmatory (restricted) factor analysis also, but in this paper we shall consider only exploratory (unrestricted) factor analysis.

3. Reduction of G

The function G in (6) is now regarded as a function $G(\Lambda,\psi)$ of Λ and ψ and is to be minimized with respect to these matrices. The minimization will be done in two steps. We first find the conditional minimum of G for a given ψ and then find the overall minimum. To begin with we shall assume that ψ is nonsingular. The partial derivative of G with respect to Λ is (see Appendix A1)

$$(12) \qquad \partial G/\partial\Lambda = 2S^{-1}(\Sigma - S)S^{-1}\Lambda \quad ,$$

which, when set equal to zero and premultiplied by S gives

$$(13) \qquad \Sigma S^{-1}\Lambda = \Lambda \quad ,$$

or

$$(14) \qquad S^{-1}\Lambda = \Sigma^{-1}\Lambda \quad .$$

Using

$$\Sigma^{-1} = \psi^{-2} - \psi^{-2}\Lambda(I + \Lambda'\psi^{-2}\Lambda)^{-1}\Lambda'\psi^{-2} \quad ,$$

(14) simplifies to

$$S^{-1}\Lambda = \psi^{-2}\Lambda(I + \Lambda'\psi^{-2}\Lambda)^{-1} \quad ,$$

or

$$(15) \qquad (\psi S^{-1}\psi)\psi^{-1}\Lambda = \psi^{-1}\Lambda(I + \Lambda'\psi^{-2}\Lambda)^{-1} \quad .$$

The matrix $\Lambda'\psi^{-2}\Lambda$ may be assumed to be diagonal since it can always be reduced to diagonal form by a proper choice of an orthogonal post-multiplier to Λ. The columns of the matrix on the right side of (15) then become proportional to the columns of $\psi^{-1}\Lambda$. Thus the columns of $\psi^{-1}\Lambda$ are characteristic vectors of $\psi S^{-1}\psi$ and the diagonal elements of

$(I + \Lambda'\psi^{-2}\Lambda)^{-1}$ are the corresponding roots. It will be shown that the conditional minimum of G, for the given ψ, is obtained when the columns of $\psi^{-1}\Lambda$ are chosen as vectors corresponding to the k smallest characteristic roots of $\psi S^{-1}\psi$. Let $\gamma_1 \leq \gamma_2 \leq \cdots \leq \gamma_p$ be the characteristic roots of $\psi S^{-1}\psi$ and let $\omega_1, \omega_2, \cdots, \omega_p$ be an orthonormal set of corresponding characteristic vectors. Let $\Gamma = \text{diag}(\gamma_1, \gamma_2, \cdots, \gamma_p)$ be partitioned as $\Gamma = \text{diag}(\Gamma_1, \Gamma_2)$ where $\Gamma_1 = \text{diag}(\gamma_1, \gamma_2, \cdots, \gamma_k)$ and $\Gamma_2 = \text{diag}(\gamma_{k+1}, \gamma_{k+2}, \cdots, \gamma_p)$ and let $\Omega = [\omega_1, \omega_2, \cdots, \omega_p]$ be partitioned as $\Omega = [\Omega_1, \Omega_2]$ where Ω consists of the first k vectors and Ω_2 of the last $p - k$ vectors. Then

$$(16) \qquad \Omega_1'\Omega_1 = I \quad , \qquad \Omega_1'\Omega_2 = 0 \quad , \qquad \Omega_2'\Omega_2 = I \quad ;$$

$$(17) \qquad \psi S^{-1}\psi = \Omega_1\Gamma_1\Omega_1' + \Omega_2\Gamma_2\Omega_2'$$

and the conditional solution $\tilde{\Lambda}$ is given by

$$(18) \qquad \tilde{\Lambda} = \psi\Omega_1(\Gamma_1^{-1} - I)^{1/2} \quad .$$

This conditional solution is identical to that of maximum likelihood factor analysis [see e.g., Jöreskog, 1967, eq. 17, where the solution is expressed in terms of the roots and vectors of $\psi^{-1}S\psi^{-1}$].

Defining

$$(19) \qquad \tilde{\Sigma} = \tilde{\Lambda}\tilde{\Lambda}' + \psi^2 \quad ,$$

it is easily verified from (16) and (18) that

$$(20) \qquad \psi^{-1}\tilde{\Sigma}\psi^{-1} = \Omega_1\Gamma_1^{-1}\Omega_1' + \Omega_2\Omega_2'$$

and that

(21) $$I - S^{-1}\tilde{\Sigma} = \psi^{-1}[\Omega_2(I - \Gamma_2)\Omega_2']\psi$$

so that

$$\mathrm{tr}(I - S^{-1}\tilde{\Sigma})^2 = \mathrm{tr}(I - \Gamma_2)^2 = \sum_{m=k+1}^{p} (\gamma_m - 1)^2 \quad .$$

Therefore the conditional minimum of $G(\Lambda,\psi)$, with respect to Λ for a given ψ is the function $g(\psi)$ defined by

(22) $$g(\psi) = \frac{1}{2} \sum_{m=k+1}^{p} (\gamma_m - 1)^2 \quad .$$

It is now clear what the effect will be of choosing, as columns of $\psi^{-1}\Lambda$, characteristic vectors other than those corresponding to the k smallest roots. The roots not chosen would then be involved in (22) and the sum of squares would be larger than or equal to that in (22).

4. Minimization of $g(\psi)$

We now propose to minimize $g(\psi)$ numerically by the Newton-Raphson method, making use of first and second derivatives of g. The roots and vectors γ_m and ω_m, $m = 1,2,\ldots,p$, of $A(\psi) = \psi S^{-1}\psi$ are functions of ψ. The first and second derivatives of $g(\psi)$ may be obtained from the first derivatives of γ_m and ω_m. As shown in Appendix A2, the latter are

(23) $$\partial\gamma_m/\partial\psi_i = (2/\psi_i)\gamma_m\omega_{im}^2 \quad ,$$

(24) $$\partial\omega_{im}/\partial\psi_j = (1/\psi_j)\omega_{jm} \sum_{n\neq m} \frac{\gamma_m + \gamma_n}{\gamma_m - \gamma_n} \omega_{in}\omega_{jn} \quad ,$$

where ω_{im} is the i^{th} element of ω_m .

By differentiating (22) with respect to ψ_i we obtain

$$\partial g/\partial\psi_i = \sum_{m=k+1}^{p} (\gamma_m - 1)(\partial\gamma_m/\partial\psi_i)$$

which, after substitution from (23), becomes

$$\partial g/\partial\psi_i = (2/\psi_i) \sum_{m=k+1}^{p} (\gamma_m^2 - \gamma_m)\omega_{im}^2 \quad . \tag{25}$$

By differentiating (25) with respect to ψ_j , we obtain

$$\partial^2 g/\partial\psi_i\partial\psi_j = (2/\psi_i) \sum_{m=k+1}^{p} \Big\{(2\gamma_m - 1)\omega_{im}^2(\partial\gamma_m/\partial\psi_j)$$
$$+ 2(\gamma_m^2 - \gamma_m)\omega_{im}(\partial\omega_{im}/\partial\psi_j) - (1/\psi_i)\delta_{ij}(\gamma_m^2 - \gamma_m)\omega_{im}^2\Big\} \quad ,$$

which, after substitution from (23) and (24) and simplification, becomes

$$\partial^2 g/\partial\psi_i\partial\psi_j = (4/\psi_i\psi_j) \sum_{m=k+1}^{p} \Big\{(2\gamma_m^2 - \gamma_m)\omega_{im}^2\omega_{jm}^2 \tag{26}$$
$$+ (\gamma_m^2 - \gamma_m)\omega_{im}\omega_{jm} \sum_{n\neq m} \frac{\gamma_m + \gamma_n}{\gamma_m - \gamma_n} \omega_{in}\omega_{jn}$$
$$- (1/2)\delta_{ij}(\gamma_m^2 - \gamma_m)\omega_{im}\omega_{jm}\Big\} \quad .$$

In minimizing $g(\psi)$ we shall follow a procedure similar to Clarke's [1970] method for maximum likelihood factor analysis. Clarke used the roots and vectors of $\psi^{-1}S\psi^{-1}$ and minimized a function of ψ^2 rather than ψ . While this method works satisfactorily in all cases where the solution is proper, having no ψ_i very close to zero, certain improvements can be made to handle Heywood cases (improper solutions) more effectively. When one or more of the ψ_i are close to zero both first- and second-order derivatives are poorly defined and difficult to compute accurately. Jöreskog [1967] describes a procedure to deal with this difficulty, which involves (i) <u>fixing</u> ψ_i^2 at some arbitrary small positive value such as 0.001 for subsequent iterations and minimizing with respect to the remaining ψ_i and (ii) once a Heywood variable with $\psi_i^2 = 0.001$ has been found, this variable is partialed out and the minimization process repeated with fewer factors on a smaller matrix. Although this is quite correct in principle, it is somewhat time consuming. When working with the roots and vectors of $\Psi S^{-1}\Psi$, rather than those of $\Psi^{-1}S\Psi^{-1}$, the partial elimination of variables may be completely avoided. Jennrich and Robinson [1969], operating on the roots and vectors of $S^{-1/2}\Psi^2 S^{-1/2}$ instead of on those of $\Psi S^{-1}\Psi$, used a similar procedure which also does not break down when Ψ is singular. Furthermore, a transformation of variables may be made which make the derivatives stable even at $\psi_i = 0$. This transformation from ψ_i to θ_i is defined by

$$(27) \qquad \theta_i = \log \psi_i^2 \;; \qquad \psi_i = +\sqrt{e^{\theta_i}} \;.$$

We now consider g as a function of $\theta_1, \theta_2, \ldots, \theta_p$ instead of $\psi_1, \psi_2, \ldots, \psi_p$. The new function $g(\theta)$ is defined for all θ_i, $-\infty < \theta_i < +\infty$. Note that $\psi_i = 0$ corresponds to $\theta_i = -\infty$.

The derivatives $\partial g/\partial\theta_i$ and $\partial^2 g/\partial\theta_i\partial\theta_j$ are obtained from $\partial g/\partial\psi_i$ and $\partial^2 g/\partial\psi_i\partial\psi_j$ by

$$\partial g/\partial\theta_i = (\psi_i/2)(\partial g/\partial\psi_i) \quad ,$$

$$\partial^2 g/\partial\theta_i\partial\theta_j = (\psi_i\psi_j/4)(\partial^2 g/\partial\psi_i\partial\psi_j) + \delta_{ij}(\psi_i/4)(\partial g/\partial\psi_i) \quad .$$

These derivatives therefore become

$$(28) \qquad \partial g/\partial\theta_i = \sum_{m=k+1}^{p} (\gamma_m^2 - \gamma_m)\omega_{im}^2 \quad ,$$

$$(29) \qquad \partial^2 g/\partial\theta_i\partial\theta_j = \sum_{m=k+1}^{p} (2\gamma_m^2 - \gamma_m)\omega_{im}^2\omega_{jm}^2$$

$$+ \sum_{m=k+1}^{p} (\gamma_m^2 - \gamma_m)\omega_{im}\omega_{jm} \sum_{n\neq m} \frac{\gamma_m + \gamma_n}{\gamma_m - \gamma_n}\omega_{in}\omega_{jn}$$

$$= \sum_{m=k+1}^{p} (2\gamma_m^2 - \gamma_n)\omega_{im}^2\omega_{jm}^2$$

$$+ \sum_{m=k+1}^{p} (\gamma_m^2 - \gamma_m)\omega_{im}\omega_{jm} \sum_{n=1}^{k} \frac{\gamma_m + \gamma_n}{\gamma_m - \gamma_n}\omega_{in}\omega_{jn}$$

$$+ \sum_{m=k+1}^{p} (\gamma_m^2 - \gamma_m)\omega_{im}\omega_{jm} \sum_{\substack{n=k+1 \\ n\neq m}}^{p} \frac{\gamma_m + \gamma_n}{\gamma_m - \gamma_n}\omega_{in}\omega_{jn} \quad .$$

The last term may be written

$$\sum_{m=k+1}^{p}\sum_{n=k+1}^{m-1}\left[\frac{(\gamma_m^2-\gamma_m)(\gamma_m+\gamma_n)}{\gamma_m-\gamma_n}+\frac{(\gamma_n^2-\gamma_n)(\gamma_m+\gamma_n)}{\gamma_n-\gamma_m}\right]\omega_{im}\omega_{jm}\omega_{in}\omega_{jn}$$

$$=\sum_{m=k+1}^{p}\sum_{n=k+1}^{m-1}(\gamma_m+\gamma_n)(\gamma_m+\gamma_n-1)\omega_{im}\omega_{jm}\omega_{in}\omega_{jn}$$

$$=\frac{1}{2}\sum_{m=k+1}^{p}\sum_{\substack{n=k+1\\ n\neq m}}^{p}(\gamma_m+\gamma_n)(\gamma_m+\gamma_n-1)\omega_{im}\omega_{jm}\omega_{in}\omega_{jn} \quad ,$$

which after substitution into (29), simplification and use of the relation

$$\sum_{m=k+1}^{p}\omega_{im}\omega_{jm}=\delta_{ij}-\sum_{n=1}^{k}\omega_{in}\omega_{jn} \quad ,$$

gives

$$\text{(30)}\qquad \partial^2 g/\partial\theta_i\partial\theta_j=\left(\sum_{m=k+1}^{p}\gamma_m\omega_{im}\omega_{jm}\right)^2+\delta_{ij}(\partial g/\partial\theta_i)$$

$$+\,2\sum_{m=k+1}^{p}(\gamma_m^2-\gamma_m)\omega_{im}\omega_{jm}\sum_{n=1}^{k}\frac{\gamma_n}{\gamma_m-\gamma_n}\omega_{in}\omega_{jn} \quad .$$

When $\gamma_{k+1},\gamma_{k+2},\ldots,\gamma_p$ are all close to one, this is approximately

$$\text{(31)}\qquad \partial^2 g/\partial\theta_i\partial\theta_j\approx\left(\sum_{m=k+1}^{p}\omega_{im}\omega_{jm}\right)^2 \quad .$$

It should be noted that the function and all derivatives of first and second order may be computed accurately everywhere, even at $\theta_i=-\infty$ ($\psi_i=0$).

Let θ denote a column vector with elements $\theta_1, \theta_2, \cdots, \theta_p$ and let h and H denote the column vector and matrix of corresponding derivatives $\partial g/\partial\theta$ and $\partial^2 g/\partial\theta\partial\theta'$, respectively. Let $\theta^{(s)}$ denote the value of θ in the s^{th} iteration and let $h^{(s)}$ and $H^{(s)}$ be the corresponding vector and matrix of first- and second-order derivatives. The iteration procedure may then be written

$$(32) \qquad H^{(s)}\delta^{(s)} = h^{(s)}$$

$$(33) \qquad \theta^{(s+1)} = \theta^{(s)} - \delta^{(s)} \quad ,$$

where $\delta^{(s)}$ is a column vector of corrections determined by (32). The Newton-Raphson procedure is therefore easy to apply, the main computations in each iteration being the computation of the roots and vectors of $\psi S^{-1}\psi$ and the solution of the symmetric system (32). It has been found that the Newton-Raphson procedure is very efficient, generally requiring only a few iterations for convergence. The convergence criterion is that the largest absolute correction be less than a prescribed small number ϵ. The minimizing θ may be determined very accurately, if desired, by choosing ϵ very small.

In detail the numerical method is as follows. The starting point $\theta^{(1)}$ is chosen as [see e.g., Jöreskog, 1963, eqs. 6.20 and 7.10 or Jöreskog, 1967, eq. 26],

$$(34) \qquad \theta_i^{(1)} = \log[(1 - k/2p)(1/s^{ii})]$$

where s^{ii} is the i^{th} diagonal element of S^{-1}. The exact matrix H of second order derivatives given by (30) may not be positive definite in the beginning. Therefore, the approximation E given by (31), which is always Gramian, is used in the first iteration and for as long as the maximum absolute correction is greater than 0.1. After that, H is used if it is positive definite. It has been found empirically that E gives good reductions in function values in the early iterations but is comparatively ineffective near the minimum, whereas H near the minimum is very effective.

To compute the characteristic roots and vectors of $\psi S^{-1}\psi$ in each iteration, we use the Householder transformation to tridiagonal form, the QR method for the roots of the tridiagonal matrix and inverse iteration for the vectors. This is probably the most efficient method available [see Wilkinson, 1965]. The system of equations (32) are solved by the square root factorization $H = TT'$, where T is lower triangular. This shows at an early stage whether H is positive definite or not.

In Heywood cases, when one or more of the $\theta_i \to -\infty$, i.e., $\psi_i \to 0$, a slight modification of the Newton-Raphson procedure is necessary to achieve fast convergence. This is due to the fact that the search for the minimum is then along a "valley" and not in a quadratic region. When $\theta_i \to -\infty$, $\partial g/\partial\theta_i \to 0$ and $\partial^2 g/\partial\theta_i\partial\theta_j \to 0$, $j = 1,2,\ldots,p$, so that when θ_i is small the i^{th} element of h and the i^{th} row and column of H and E are also small. This tends to produce a "bad" correction vector δ and the function may increase instead of decrease. A simple and effective way to deal with this problem is to delete the i^{th} equation in the system (32) and compute the corrections for all the other

θ's from the reduced system. One then computes the correction for θ_i as

$$(35) \qquad \delta_i = (\partial g/\partial\theta_i)/(\partial^2 g/\partial\theta_i^2) \quad .$$

This procedure will decrease θ_i slowly in the beginning but faster the more evident it is that θ_i is a Heywood variable. When θ_i has become less than -10 it is not necessary to change θ_i any more unless $\partial f/\partial\theta_i$ is negative. Thus, the procedure corrects itself quickly if a variable is incorrectly taken as a Heywood variable.

5. Asymptotic Distribution Theory

In this section we show that the GLS estimates and the ML estimates have the same asymptotic properties. In particular we shall evaluate the common asymptotic variance-covariance matrix of the estimates of $\psi_1, \psi_2, \ldots, \psi_p$.

It is assumed that S converges stochastically to Σ of the form (2), and that the elements of $\sqrt{n}$ $(S - \Sigma)$ have an asymptotic multinormal distribution with variances and covariances given by

$$(36) \qquad n\mathcal{E}[(s_{\alpha\beta} - \sigma_{\alpha\beta})(s_{\mu\nu} - \sigma_{\mu\nu})] = \sigma_{\alpha\mu}\sigma_{\beta\nu} + \sigma_{\alpha\nu}\sigma_{\beta\mu} \quad ,$$

which are the elements of $2(\Sigma \otimes \Sigma)$. In particular, this is true when the observations on y are drawn from a multinormal distribution with variance-covariance matrix Σ . The matrices Σ , Λ and ψ now denote the true population values as distinguished from the mathematical variables Λ and ψ used in the previous sections. It is furthermore assumed that $\psi_i \neq 0$, $i = 1,2,\ldots,p$, i.e., that the population is not a Heywood case.

Let $A = \psi\Sigma^{-1}\psi$ and let $\gamma_1 \leq \gamma_2 \leq \cdots \leq \gamma_p$ be the characteristic roots of A with $\omega_1, \omega_2, \cdots, \omega_p$ an orthonormal set of corresponding characteristic vectors. Let $\Gamma_1 = \mathrm{diag}(\gamma_1, \gamma_2, \cdots, \gamma_k)$, $\Gamma_2 = \mathrm{diag}(\gamma_{k+1}, \gamma_{k+2}, \cdots, \gamma_p)$, $\Omega_1 = [\omega_1, \omega_2, \cdots, \omega_k]$ and $\Omega_2 = [\omega_{k+1}, \omega_{k+2}, \cdots, \omega_p]$. We assume that the roots in Γ_1 are all distinct. Then

(37) $$A = \Omega_1\Gamma_1\Omega_1' + \Omega_2\Gamma_2\Omega_2'$$

and

(38) $$A^{-1} = \Omega_1\Gamma_1^{-1}\Omega_1' + \Omega_2\Gamma_2^{-1}\Omega_2' \quad .$$

However, since

(39) $$A^{-1} = \psi^{-1}\Sigma\psi^{-1} = \psi^{-1}\Lambda\Lambda'\psi^{-1} + I \quad ,$$

we have that $\gamma_{k+1} = \gamma_{k+2} = \cdots = \gamma_p = 1$, or

(40) $$\Gamma_2 = I \quad .$$

Defining

(41) $$\Xi = \Omega_2\Omega_2' \quad ,$$

it follows from (37), (38), (40) , $\Omega_1'\Omega_2 = 0$ and $\Omega_2'\Omega_2 = I$ that Ξ has the properties

(42) $$A\Xi = A^{-1}\Xi = \Xi A = \Xi A^{-1} = \Xi^2 = \Xi \quad .$$

Corresponding to the population quantities in (37) and (38) we have the corresponding sample quantities

$$(43) \qquad \hat{A} = \hat{\Omega}_1\hat{\Gamma}_1\hat{\Omega}_1' + \hat{\Omega}_2\hat{\Gamma}_2\hat{\Omega}_2'$$

and

$$(44) \qquad \hat{A}^{-1} = \hat{\Omega}_1\hat{\Gamma}_1^{-1}\hat{\Omega}_1' + \hat{\Omega}_2\hat{\Gamma}_2^{-1}\hat{\Omega}_2' \quad ,$$

where $\hat{\Gamma}_1 = \mathrm{diag}(\hat{\gamma}_1,\hat{\gamma}_2,\ldots,\hat{\gamma}_k)$ and $\hat{\Gamma}_2 = \mathrm{diag}(\hat{\gamma}_{k+1},\hat{\gamma}_{k+2},\ldots,\hat{\gamma}_p)$ are diagonal matrices of the characteristic roots $\hat{\gamma}_1 \leq \hat{\gamma}_2 \leq \cdots \leq \hat{\gamma}_p$ of $\hat{A} = \hat{\psi}S^{-1}\hat{\psi}$ and $\hat{\Omega}_1$ of order $p \times k$ and $\hat{\Omega}_2$ of order $p\mathrm{x}(p - k)$ are matrices of corresponding orthonormal characteristic vectors. These are the quantities obtained at the minimum of $g(\psi)$.

We shall show that $\hat{\psi}$ converges stochastically to ψ . The function $g(\psi)$ in (22) is also a function of S and will now be denoted $g(S,\psi^*)$. The estimate $\hat{\psi}$ is defined as the value of ψ^* that minimizes $g(S,\psi^*)$ for a given S . But $g(S,\psi^*)$ converges stochastically to $g(\Sigma,\psi^*)$ which has a unique minimum at $\psi^* = \psi$. Since the functions are continuous, $\hat{\psi}$ must converge stochastically to ψ .

In deriving various asymptotic results we shall make repeated use of the following well-known lemma [see e.g., Wilks, 1962, p. 103]: If $C = (c_{ij})$ is a matrix whose elements are continuous functions of random variables $x_1,x_2,\ldots,x_m$ and if $\mathrm{plim}\ x_k = \xi_k$ exists and is finite for all k , then $\mathrm{plim}\ C(x) = C(\xi)$.

From this it follows immediately that

$$(45) \qquad \mathrm{plim}\ \hat{A} = \mathrm{plim}\ \hat{\psi}S^{-1}\hat{\psi} = \psi\Sigma^{-1}\psi = A \quad ,$$

and that $\mathrm{plim}\ \hat{\gamma}_m = \gamma_m$, $\mathrm{plim}\ \hat{\omega}_m = \omega_m$.

Hence, from (30) and (40) we have that

$$\text{plim } \partial^2 g/\partial\theta_i\partial\theta_j = \Big(\sum_{m=k+1}^{p} \omega_{im}\omega_{jm}\Big)^2 \quad ,$$

and from (26) that [cf. Anderson & Rubin, 1956, eq. 12.24; Lawley, 1967, eq. 7 and Jöreskog, 1967, eq. 101]

$$(46) \qquad \text{plim } \partial^2 g/\partial\psi_i\partial\psi_j = (4/\psi_i\psi_j)\Big(\sum_{m=k+1}^{p} \omega_{im}\omega_{jm}\Big)^2 \quad .$$

The asymptotic variance-covariance matrix of the ML estimates of the ψ's is given by $(2/n)E^{-1}$, where E is the matrix whose ij^{th} element is given by the right-hand side of (46). We proceed to show that $(2/n)E^{-1}$ is also the asymptotic variance covariance matrix of the GLS estimates of the ψ's.

The GLS estimates $\hat{\psi}_1, \hat{\psi}_2, \ldots, \hat{\psi}_p$ are defined implicitly by the following equations

$$\partial g/\partial\psi_i = 0 \quad , \qquad i = 1,2,\ldots,p \quad ,$$

which by (25) may be written

$$(47) \qquad \text{diag}[\hat{\Omega}_2(\hat{\Gamma}_2^2 - \hat{\Gamma}_2)\hat{\Omega}_2'] = 0 \quad .$$

We shall write (47) linearly in statistical differentials. The symbol δ is used to denote deviations of sample from population values. All such deviations are of order $n^{-1/2}$ in probability and since we assume that n is large, we shall neglect in what follows terms of second and higher degrees in the δ's. Let $\delta\Gamma_2 = \hat{\Gamma}_2 - \Gamma_2 = \hat{\Gamma}_2 - I$, $\delta\Omega_2 = \hat{\Omega}_2 - \Omega_2$ and $\delta A = \hat{A} - A$. Then we have to the order of approximation indicated $\hat{\Gamma}_2^2 = I + 2\delta\Gamma_2$, $\hat{\Gamma}_2^2 - \hat{\Gamma}_2 = \delta\Gamma_2$ and $\hat{\Omega}_2(\hat{\Gamma}_2^2 - \hat{\Gamma}_2)\hat{\Omega}_2' = \Omega_2\delta\Gamma_2\Omega_2'$. But $\delta\Gamma_2 = \Omega_2'\delta A\Omega_2$ which may be verified from $\hat{A}\hat{\Omega}_2 = \hat{\Omega}_2\hat{\Gamma}_2$ and $\hat{\Omega}_2'\hat{\Omega}_2 = I$. Hence, (47) is asymptotically equivalent to

$$(48) \qquad \mathrm{diag}(\Xi\,\delta A\,\Xi) = 0 \quad .$$

Furthermore, with $\delta\psi = \hat{\psi} - \psi$ and $\delta\Sigma = S - \Sigma$ we have to the same order of approximation

$$S^{-1} = (\Sigma + \delta\Sigma)^{-1} = \Sigma^{-1} - \Sigma^{-1}\delta\Sigma\Sigma^{-1}$$

and

$$\begin{aligned} \delta A &= (\psi + \delta\psi)(\Sigma^{-1} - \Sigma^{-1}\delta\Sigma\Sigma^{-1})(\psi + \delta\psi) - A \\ &= \delta\psi\Sigma^{-1}\psi + \psi\Sigma^{-1}\delta\psi - \psi\Sigma^{-1}\delta\Sigma\Sigma^{-1}\psi \\ &= \delta\psi\psi^{-1}A + A\psi^{-1}\delta\psi - A\psi^{-1}\delta\Sigma\psi^{-1}A \quad , \end{aligned}$$

which after substitution into (48) and use of (42) shows that (48) is asymptotically equivalent to

$$(49) \qquad 2\mathrm{diag}(\Xi\,\delta\psi\psi^{-1}\Xi) = \mathrm{diag}(\Xi\,\psi^{-1}\delta\Sigma\psi^{-1}\,\Xi) \quad .$$

From (37) it follows that the elements of $T = \psi^{-1}\delta\Sigma\psi^{-1}$ have a limiting multinormal distribution with variances and covariances given by

$$(50) \qquad n\mathcal{E}(t_{\alpha\beta}t_{\mu\nu}) = a^{\alpha\mu}a^{\beta\nu} + a^{\alpha\nu}a^{\beta\mu} \quad ,$$

where a^{ij} denotes the ij^{th} element of A^{-1} .

Equation (49) is linear in $\delta\psi_1, \delta\psi_2, \ldots, \delta\psi_p$ and may be written in scalar form as

$$(51) \qquad 2 \sum_{x=1}^{p} \xi_{ix}^2 (\delta\psi_x/\psi_x) = \sum_{\alpha=1}^{p} \sum_{\beta=1}^{p} \xi_{i\alpha}\xi_{i\beta}t_{\alpha\beta} \quad , \qquad i = 1,2,\ldots,p \quad ,$$

which may be solved for $\delta\psi_g/\psi_g$ if the matrix Φ with elements $\phi_{ij} = \xi_{ij}^2 = (\sum_{m=k+1}^{p} \omega_{im}\omega_{jm})^2$ is nonsingular. The solution is

$$(52) \qquad \delta\psi_g/\psi_g = \frac{1}{2} \sum_{i=1}^{p} \sum_{\alpha=1}^{p} \sum_{\beta=1}^{p} \phi^{gi}\xi_{i\alpha}\xi_{i\beta}t_{\alpha\beta} \quad , \qquad g = 1,2,\ldots,p \quad .$$

Equation (52) shows that $\delta\psi_1, \delta\psi_2, \ldots, \delta\psi_p$ are asymptotically linear in the elements of T and hence will have a limiting multinormal distribution. To obtain the asymptotic variance-covariance matrix of $\hat{\psi}_1, \hat{\psi}_2, \ldots, \hat{\psi}_p$ we write equation (52) with indices h , j , μ and ν instead of g , i , α and β respectively, multiply these equations and use (50) and (42). This gives

$$\begin{aligned}
(n/\psi_g\psi_h)\mathcal{E}(\delta\psi_g\delta\psi_h) &= (1/4) \sum_i \sum_j \sum_\alpha \sum_\beta \sum_\mu \sum_\nu \phi^{gi}\phi^{hj}\xi_{i\alpha}\xi_{i\beta}\xi_{j\mu}\xi_{j\nu}(a^{\alpha\mu}a^{\beta\nu} + a^{\alpha\nu}a^{\beta\mu}) \\
&= (1/2) \sum_i \sum_j \phi^{gi}\phi^{hj}\xi_{ij}^2 \\
&= (1/2) \sum_i \sum_j \phi^{gi}\phi_{ij}\phi^{jh} \\
&= (1/2)\phi^{gh} \quad .
\end{aligned}$$

Hence,

(53) $\mathcal{E}(\delta\psi_g \delta\psi_h) = (\psi_g \psi_h / 2n)\phi^{gh}$,

which is the gh^{th} element of $(2/n)E^{-1}$. This, therefore, shows that the asymptotic variance-covariance matrix $(2/n)E^{-1}$ is the same for both the ML estimates and the GLS estimates.

Lawley [1967] obtained the unconditional asymptotic distribution of the ML estimate $\hat{\Lambda}$ from the conditional asymptotic distribution of $\tilde{\Lambda}$ for given ψ [Lawley, 1953]. Since the conditional estimate $\tilde{\Lambda}$ is the same for both ML and GLS, it follows that also the GLS estimate $\hat{\Lambda}$ has the same asymptotic distribution.

Another well-known result for the ML method is that n times the minimum value F_{min} of F in (5) is asymptotically distributed as χ^2 with $d = \frac{1}{2}[(p - k)^2 - (p + k)]$ degrees of freedom. The same statement is true also for the GLS method. To prove this we show that both minima are asymptotically equivalent.

Let $\tilde{\psi}$ denote the maximum likelihood estimates of ψ . Since $\tilde{\psi}$ is asymptotically equivalent to $\hat{\psi}$, the characteristic roots $\tilde{\gamma}_1, \tilde{\gamma}_2, \ldots, \tilde{\gamma}_m$ of $\tilde{\psi}S^{-1}\tilde{\psi}$ are asymptotically equivalent to the corresponding roots $\hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_p$ of $\hat{\psi}S^{-1}\hat{\psi}$. The minimum of F is [see e.g., Jöreskog, 1967, eq. 18]

$$\sum_{m=k+1}^{p} (\log \tilde{\gamma}_m + 1/\tilde{\gamma}_m - 1)$$

which is asymptotically equivalent to

$$\sum_{m=k+1}^{p} (\log \hat{\gamma}_m + 1/\hat{\gamma}_m - 1) = \sum_{m=k+1}^{p} [\log(1 + \delta\gamma_m) + \frac{1}{1 + \delta\gamma_m} - 1]$$

$$= \sum_{m=k+1}^{p} (\delta\gamma_m - \frac{1}{2}\,\delta\gamma_m^2 + 1 - \delta\gamma_m + \delta\gamma_m^2 - 1)$$

$$= \frac{1}{2} \sum_{m=k+1}^{p} \delta\gamma_m^2$$

$$= \frac{1}{2} \sum_{m=k+1}^{p} (\hat{\gamma}_m - 1)^2$$

$$= g_{min} \quad .$$

6. Results and Comparisons on Numerical Data

The algorithm described in section 4 has been implemented in a FORTRAN program and run on several matrices. It is interesting to compare the results of GLS and ML on the same two correlation matrices, Data 1 and Data 2, as Jöreskog [1967] and Clarke [1970] analyzed with the ML method. The correlation matrices are given in both of these papers.

Data 1 is a correlation matrix of order 9×9 and is analyzed with three factors. The course of the minimization is shown in Table 1. It is seen that the convergence is quite rapid and that the solution can be determined very accurately, to about five decimals in the θ's. This corresponds to an accuracy of about seven decimals in ψ^2 . The solution is given in Table 2 along with the ML solution. It is seen that the two solutions are very close, so close that interpretations of the data will be the same. The value of χ^2 with 12 degrees of freedom and based on $n = 210$, is 6.98 with GLS and 7.35 with ML. These are very close in this case when the fit is very good.

Data 2 is a correlation matrix of order 10×10 and is analyzed with four factors. The maximum likelihood solution for this data is a Heywood case with $\psi = 0$ for variable 8. The behavior under the GLS minimization is shown in Table 3. In this case it takes nine iterations to achieve convergence. This is because θ_8 goes very slowly to -10 and reaches -10 at iteration 5. After that, convergence is quadratic. The GLS and ML solutions are given in Table 4. Also in this case the two solutions are very close. The corresponding χ^2 values, 19.40 with GLS and 18.45 with ML based on 11 degrees of freedom and $n = 809$, are somewhat more apart, despite the fact that n is large. However, the fit of the factor model is not as good as in the Data 1 example.

It should be noted that for the GLS estimates it does not hold that $\hat{\psi}^2 = \mathrm{diag}(S - \hat{\Lambda}\hat{\Lambda}')$ which holds for ML estimates. In the examples, communalities and uniquenesses do not add up to unity. Also it can be seen in both Table 2 and Table 4 that the GLS estimates of ψ^2 are generally smaller than the ML estimates. This suggests that the GLS estimates may be systematically biased.

References

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

Anderson, T. W. Some scaling models and estimation procedures in the latent class model. In U. Grenander (Ed.), Probability and statistics: The Harald Cramér volume. New York: Wiley, 1959. Pp. 9-38.

Anderson, T. W. & Rubin, H. Statistical inference in factor analysis. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1956. Pp. 111-149.

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

Clarke, M. R. B. A rapidly convergent method for maximum-likelihood factor analysis. The British Journal of Mathematical and Statistical Psychology, 1970, 23, 43-52.

Ferguson, T. S. A method of generating best asymptotically normal estimates with application to the estimation of bacterial densities. Annals of Mathematical Statistics, 1958, 29, 1049-1062.

Harman, H. H. Modern factor analysis. (2nd ed.) Chicago: University of Chicago Press, 1967.

Jennrich, R. I. & Robinson, S. M. A Newton-Raphson algorithm for maximum likelihood factor analysis. Psychometrika, 1969, 34, 111-123.

Jöreskog, K. G. Statistical estimation in factor analysis. Stockholm: Almqvist & Wiksell, 1963.

Jöreskog, K. G. Some contributions to maximum likelihood factor analysis. Psychometrika, 1967, 32, 443-482.

Lawley, D. N. A modified method of estimation in factor analysis and some large sample results. Proceedings of the Uppsala Symposium on Psychological Factor Analysis, March 17-19, 1953. Nordisk Psykologi's Monograph Series No. 3. Stockholm: Almqvist & Wiksell, 1953. Pp. 35-42.

Lawley, D. N. Some new results in maximum likelihood factor analysis. Proceedings of the Royal Society of Edinburgh, Section A, 1967, 67, 256-264.

Malinvaud, E. Statistical methods of econometrics. Chicago: Rand-McNally, 1966.

Neyman, J. Contribution to the theory of χ^2 test. Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1949. Pp. 239-273.

Rothenberg, T. Structural restrictions and estimation efficiency in linear econometric models. Cowles Foundation Discussion Paper No. 213. New Haven, Conn.: Yale University, 1966.

Taylor, W. F. Distance functions and regular best asymptotically normal estimates. Annals of Mathematical Statistics, 1953, 24, 85-92.

Wilkinson, J. H. The algebraic eigenvalue problem. Oxford: Oxford University Press, 1965.

Zellner, A. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. Journal of the American Statistical Association, 1962, 57, 348-368.

TABLE 1

Details of the GLS Minimization for Data 1

Iteration	Type	Function	Max. correction	Max. gradient
0	--	0.1170246	--	2.45×10^{-1}
1	E	0.04017278	6.64×10^{-1}	5.04×10^{-2}
2	E	0.03341929	1.71×10^{-1}	1.06×10^{-2}
3	E	0.03321625	2.73×10^{-2}	7.36×10^{-4}
4	H	0.03321503	2.91×10^{-3}	3.64×10^{-6}
5	H	0.03321503	2.67×10^{-5}	2.37×10^{-10}

TABLE 2

Solutions for Data 1

	GLS				ML			
i	λ_{i1}	λ_{i2}	λ_{i3}	ψ_i^2	λ_{i1}	λ_{i2}	λ_{i3}	ψ_i^2
1	.662	.325	-.082	.445	.664	.321	-.073	.450
2	.688	.255	.191	.416	.689	.247	.193	.427
3	.491	.310	.225	.600	.493	.302	.222	.617
4	.839	-.286	.041	.208	.837	-.292	.035	.212
5	.708	-.309	.162	.370	.705	-.315	.153	.381
6	.823	-.376	-.106	.168	.819	-.377	-.105	.177
7	.660	.404	.073	.387	.662	.396	.078	.400
8	.454	.290	-.484	.473	.458	.296	-.491	.462
9	.763	.434	.001	.227	.766	.427	.012	.231

TABLE 3

Details of the GLS Minimization for Data 2

Iteration	Type	Function	Max. correction	Max. gradient
0	--	0.08765774	9.31×10^{-1}	9.89×10^{-2}
1	E	0.03525694	4.09×10^{-1}	2.23×10^{-2}
2	E	0.02892803	4.65×10^{-1}	2.17×10^{-2}
3	E	0.02574564	8.74×10^{-1}	1.53×10^{-2}
4	E	0.02418829	3.63×10^{0}	7.71×10^{-3}
5	E	0.02401558	1.97×10^{4}	1.10×10^{-3}
6	H	0.02398666	1.00×10^{0}	3.55×10^{-4}
7	H	0.02398479	1.37×10^{-2}	2.99×10^{-5}
8	H	0.02398478	1.49×10^{-3}	4.81×10^{-7}
9	H	0.02398478	1.52×10^{-5}	4.81×10^{-7}

TABLE 4

Solutions for Data 2

	GLS					ML				
i	λ_{i1}	λ_{i2}	λ_{i3}	λ_{i4}	ψ_i^2	λ_{i1}	λ_{i2}	λ_{i3}	λ_{i4}	ψ_i^2
1	-.188	-.756	.034	-.100	.377	-.188	.753	-.035	-.108	.385
2	-.120	-.468	.095	.382	.604	-.120	.468	-.103	.365	.623
3	-.186	-.763	.157	.221	.309	-.186	.767	-.167	.217	.301
4	-.173	-.527	.198	.135	.629	-.173	.526	-.200	.124	.638
5	-.129	-.678	.258	-.345	.336	-.129	.672	-.251	-.349	.347
6	.359	.259	.157	-.047	.767	.359	-.259	-.154	-.048	.778
7	.448	.501	.504	.059	.289	.448	-.504	-.507	.052	.286
8	1.000	-.000	-.000	.000	.000	1.000	.000	.000	.000	.000
9	.429	.282	.212	-.051	.680	.429	-.282	-.209	-.053	.690
10	.316	.232	.505	-.020	.580	.316	-.232	-.496	-.029	.600

A. Appendix

A1. Matrix Derivative of Function $G(\Lambda, \psi)$

To obtain the matrix derivatives we use matrix differentials. In general, $dX = (dx_{ij})$ will denote a matrix of differentials. If F is a function of X and $dF = tr(CdX')$ then $\partial F/\partial X = C$. Since $d\ tr(A) = tr(dA)$, we have for a fixed ψ, with G defined by (6) and Σ by (2),

$$dG = \frac{1}{2} d\ tr(S^{-1}\Sigma - I)^2$$

$$= \frac{1}{2} tr[d(S^{-1}\Sigma - I)^2]$$

$$= tr[(S^{-1}\Sigma - I)d(S^{-1}\Sigma - I)]$$

$$= tr[(S^{-1}\Sigma - I)S^{-1}d\Sigma]$$

$$= tr[(S^{-1}\Sigma - I)S^{-1}(\Lambda d\Lambda' + d\Lambda\Lambda')]$$

$$= 2tr(S^{-1}\Sigma - I)S^{-1}\Lambda d\Lambda'$$

$$= 2tr[S^{-1}(\Sigma - S)S^{-1}\Lambda d\Lambda'] \quad .$$

Hence, the derivative $\partial G/\partial \Lambda$ is that given by (13).

A2. Matrix Derivatives of Characteristic Roots and Vectors

The characteristic roots γ_m and vectors ω_m, $m = 1,2,\ldots,p$, of A are defined by

$$\left.\begin{array}{ll}(A1) & A\omega_m = \gamma_m\omega_m \\ (A2) & \omega_m'\omega_m = 1 \\ (A3) & \omega_m'\omega_n = 0 \quad , \quad n \neq m\end{array}\right\} \quad , \qquad m = 1,2,\ldots,p \quad .$$

Differentiation of these equations gives

$$(A4) \qquad dA\omega_m + Ad\omega_m = d\gamma_m\omega_m + \gamma_m d\omega_m$$

$$(A5) \qquad \omega_m' d\omega_m = 0$$

$$(A6) \qquad \omega_m' d\omega_n + d\omega_m'\omega_n = 0 \quad , \qquad n \neq m \quad .$$

Premultiplication of (A4) by ω_m' and use of (A1) and (A5) gives

$$(A7) \qquad d\gamma_m = \omega_m' dA\omega_m \quad .$$

Let $\epsilon_{mn} = \omega_m' dA\omega_n = \epsilon_{nm}$ for $m,n = 1,2,\ldots,p$. Then premultiplication of (A4) by ω_n' for $n \neq m$ and use of (A1) and (A3) gives

$$\begin{aligned}\epsilon_{mn} &= \gamma_m\omega_n' d\omega_m - \omega_n' Ad\omega_m \\ &= \gamma_m\omega_n' d\omega_m - \gamma_n\omega_n' d\omega_m \\ &= (\gamma_m - \gamma_n)\omega_n' d\omega_m\end{aligned}$$

or

$$(A8) \qquad \omega_n' d\omega_m = \frac{\epsilon_{mn}}{\gamma_m - \gamma_n} \quad , \qquad n \neq m \quad .$$

Multiplying this equation by ω_n , summing over $n \neq m$, using (A5) and remembering that

$$\sum_{n \neq m} \omega_n \omega_n' = I - \omega_m \omega_m'$$

gives $d\omega_m$ as

$$\text{(A9)} \qquad d\omega_m = \sum_{n \neq m} \frac{\epsilon_{mn}}{\gamma_m - \gamma_n} \omega_n \quad .$$

The merits of (A7) and (A9) are that they express the differentials of γ_m and ω_m in terms of the differentials of A .

In our problem we have $A = \psi S^{-1} \psi$ as a function of ψ so that

$$dA = d\psi S^{-1} \psi + \psi S^{-1} d\psi$$

$$= d\psi \psi^{-1} A + A \psi^{-1} d\psi \quad .$$

Substitution of this into the definition of ϵ_{mn} gives

$$\epsilon_{mn} = \omega_m' dA \omega_n$$

$$= \omega_m' d\psi \psi^{-1} A \omega_n + \omega_m' A \psi^{-1} d\psi \omega_n$$

$$= (\gamma_m + \gamma_n) \omega_m' d\psi \psi^{-1} \omega_n$$

$$= (\gamma_m + \gamma_n) \operatorname{tr}(\omega_n \omega_m' \psi^{-1} d\psi) \quad .$$

With this result we have

$$\text{(A10)} \qquad d\gamma_m = 2\gamma_m \operatorname{tr}(\omega_m \omega_m' \psi^{-1} d\psi)$$

and

$$\text{(A11)} \qquad d\omega_m = \sum_{n \neq m} \frac{\gamma_m + \gamma_n}{\gamma_m - \gamma_n} \, tr(\omega_n \omega_m' \psi^{-1} d\psi)\omega_n \quad .$$

Hence the derivatives of γ_m and ω_{im} with respect to ψ_j are

$$\text{(A12)} \qquad \partial\gamma_m / \partial\psi_j = (2\gamma_m/\psi_j)\omega_{jm}^2$$

and

$$\text{(A13)} \qquad \partial\omega_{im}/\partial\psi_j = (1/\psi_j)\, \omega_{jm} \sum_{n \neq m} \frac{\gamma_m + \gamma_n}{\gamma_m - \gamma_n} \omega_{in}\omega_{jn}$$

which are the results used in section 4.