COMPARISON OF ESTIMATION METHODS IN FACTOR ANALYSIS

SEPTEMBER 1986

BY

MASAMORI IHARA

SUMMARY

This thesis is concerned with estimation problems in factor analysis. The paper first specified the random factor analysis model and then gave a simpler and more direct proof for Anderson and Rubin's theorem on the identifiability of parameters.

Secondly, the paper reviewed iterative procedures now available and then proposed an algorithm, the partial Gauss-Newton algorithm, which can be dealt with both the least-squares and maximum likelihood (ML) methods. Applied to two real data sets, it was shown to work well. The structure of improper solutions was clarified analytically for both ML and WLS (weighted least-squares) methods.

A Monte Carlo study was carried out to compare the three estimation methods, SLS (simple least-squares), WLS and ML. It was found that SLS performed better than WLS or ML for a small sample size, not exceeding 300. It was proved in a special case that all the three estimation methods tended to underestimate the uniquenesses asymptotically.

Finally, a new consistent estimator of the uniqueness which can be expressed as an explicit function of the sample covariance matrix was proposed. Applied to real data sets, the estimate was shown to be rather close to the ML estimate whichis well known to be asaymptotically best.

CONTENTS

CONTENTS (continued)

CHAPTER 0

INTRODUCTION

Factor analysis aims to reduce the dimensionality of observed multivariate data by explaining the observed inter-relations among the variates involved in terms of simpler relations. The simplification may consist of creating a smaller number of latent (unobservable or hypothetical) variables, or producing a set of classificatory categories. Such an aim is so central to all scientific work that factor analysis has become one of the most popular multivariate statistical techniques in many fields of scientific research as is reviewed in Gnanadesikan and Kettenring [1984] .

As a research instrument, factor analysis was developed originally by psychologists. Spearman, in 1904, being interested to prove his psychological theory that all forms of intellectual performances spring from a single general mental capacity, developed a proof that if a matrix of correlations takes a certain form, then the inter-relationships of all the variables involved could be accounted for by a single underlying general ability factor called the common factor, plus a factor called the unique factor specific to each performance. This mono-factor theory by Spearman [1904] was generalized in the next thirty years, principally by Thurstone [1935] , into principles for multiple factor analysis.

An early theoretical account of the subject was given by Anderson and Rubin [1956] and more recent and comprehensive treatments are provided by Lawley and Maxwell [1971] and Okamoto [1986a] .

The application of factor analysis to the real data starts off with assuming a statistical model specified by parameters to be estimated from the data. The random factor analysis model, which is defined in Chapter 1, is the most important of all the models for factor analysis. The fixed factor analysis model also is important, but it is not dealt with in this thesis.

The next stage is to estimate the unknown parameters in the model under the assumption that the number of common factors is known. Among various estimation methods proposed so far the least-squares (LS) and maximum likelihood (ML) methods are most popular. We need an iterative procedure if we want to use either LS or ML. A great number of computational algorithms have been described by many authors, since Jöreskog [1967] described an algorithm to determine the ML estimates by using the method of Fletcher and Powell [1963] which is an efficient non-linear procedure: see Jennrich and Robinson [1969] , Jöreskog [1977] Lee and Jennrich [1979] , Lee [1980] and Lee and Poon [1985] among others.

In Chapter 2 we first reviewed these algorithms from the points of view of (a) choice of variables, (b) restriction for the rotation and (c) optimization method, and then discussed a new procedure which was described in Okamoto and Ihara [1984] .

When we use an iterative procedure, a solution may be obtained on the boundary of the admissible parameter space. Such an estimate which corresponds to the above solution is called improper solution or Heywood case and various causes of improper solutions have been pointed out by some researchers; see Cliff and Pennell [1967] , Pennell [1968] , van Driel [1978] , Boomsma [1982] and [1985] and Anderson and Gerbing [1984] among others. On the other hand, Jöreskog [1967] clarified the structure of such an estimate in maximum likelihood factor analysis under the assumption that some unique variances are known to be zero. We extended his result to the case where the estimates of some unique variances are zero in Chapter 2.

It is well known that ML leads to asymptotically best estimators for large samples. It needs the normality assumption with respect to the distributions of all the variables in the model, which has been questioned whether to be realistic or not, especially in the field of psychology. On the other hand, there are some researchers like Wold [1982] who favor the least-squares approch which does not necessarily need the normality assumption. Therefore we believe that it is worth while comparing the performance of LS and ML for a small sample. No exact expression to evaluate the sampling error, for instance, variance or mean squared error, has yet been obtained and we believe that only the Monte Carlo approach will be available from now on. Many Monte Carlo studies have been carried out with respect to maximum likelihood factor analysis, but LS seems untried so far, much less experimental comparison of LS and ML methods: again see Cliff and Pennell [1967] , Pennell [1968] , Boomsma [1982] and [1985] and Anderson and Gerbing [1984] . Ihara and Okamoto [1985] carried out a Monte Carlo study to compare the three methods, the simple (SLS) and weighted (WLS) least-squares methods and the ML method, where SLS and WLS are two most important members of a family of least-squares methods and defined in Chapter 2. In Ihara and Okamoto [1985] we used only one numerical model, so that we carried out additional experiments using a new model in order to make the conclusions obtained in Ihara and Okamoto [1985] more reliable.

In Chapter 3 we first discussed experiments and results and then prove from the viewpoint of an asymptotic theory that the estimators of uniquenesses obtained by each method, SLS, WLS or ML, tend to be negatively biased. This tendency for WLS was first indicated by Jöreskog and Goldberger [1972] and after that Boomsma [1982] and [1985] and Anderson and Gerbing [1984] found in their experimental studies that the ML estimators also had such tendency. However, as far as the author is aware, any analytic approach has not been attempted until present.

Both the LS and ML estimates are determined as a solution of simultaneous non-linear differential equations, so that we can not express the estimates as explicit functions of the sample covariance matrix S. In Chapter 4 we proposed an entirely new estimators of uniquenesses which can be expressed as explicit functions of S and hence can be obtained without using any iterative procedures.

Chapter 2 is mainly based on Okamoto and Ihara [1984] and Ihara [1986] and Chapter 3 on Ihara and Okamoto [1985] and Ihara [1985]. The contents of Chapters 1 and 4 consist of Ihara and Kano [1986].

CHAPTER 1

SPECIFICATION OF FACTOR ANALYSIS MODEL

1.1 DEFINITION OF MODELS

In a random factor analysis model an observable random vector x of p components is usually represented in the form

$$x = \mu + \Lambda f + e, \tag{1.1}$$

where μ is a fixed vector of p population means, Λ is a fixed $p \times k$ matrix of factor loadings, f is a random vector of k ($k < p$) common factors and e is a random vector of p unique factors (or unique factors plus specific factors). We assume that $E(f) = 0$, $E(f f') = I$, $E(e) = 0$, $E(f e') = 0$ and $E(e e') = \Psi$, diagonal and non-negative definite matrix. Then the covariance matrix Σ of x is

$$\Sigma = \Lambda \Lambda' + \Psi \tag{1.2}$$

from (1.1) and the above assumptions.

Let us write $\Sigma = (\sigma_{ij})$, $\Lambda = (\lambda_{ir})$ and $\Psi = (\delta_{ij} \psi_i)$, where δ_{ij} stands for Kronecker's delta. Then from (1.2) we have

$$\sigma_{ii} = \sum_{r=1}^{k} \lambda_{ir}^2 + \psi_i \tag{1.3}$$

for each i ($i = 1, . . . , p$). In the terminology of factor analysis, the proportion of the first term in the right-hand side of (1.3) to the left-hand side

$$h_{ii} = \sum_{r=1}^{k} \lambda_{ir}^2 / \sigma_{ii}$$

is called the communality of the variable x_i and the quantity $1 - h_{ii}$ the uniqueness, whereas ψ_i is called a unique variance of x_i.

There are problems about the model (1.1) , such as what covaiance matrix Σ can be represented by (1.2) for a given k and, if there is such a representation, what restrictions shall be put on Λ and Ψ to make them unique. In the way of statistical inference, there is the problem of estimating Λ and Ψ from a set of observations on x ; the principal factor analysis method is the simplest, while the least-squares and maximum likelihood methods are most popular. Another problem is to test whether k is a given number and thus decide what number k is. The above problems with respect to the model are discussed in the following sections, whereas the estimating problem is discussed in the following chapters. However, the problem about the number k is not treated in this thesis.

1.2 IDENTIFICATION PROBLEM

Among various problems with respect to the model, the most important problem may be what covariance matrix Σ can be represented in the form of (1.2) ; given a p x p positive definite matrix Σ, can it be expressed as $\Lambda\Lambda' + \Psi$? However, since our main subject in this thesis is to compare the performance of estimation methods in factor analysis, we shall no more consider the problem.

On the other hand, when we want to construct a numerical model in a Monte Carlo study, it is the most important problem to confirm whether the population covariance matrix Σ can be expressed uniquely as $\Lambda\Lambda' + \Psi$ or not, so that we need consider the problem what restrictions shall be put on Λ and Ψ to make them unique. This is called the identification problem and various conditions have been proposed so far by many authors: see, for example, Albert [1944a] and [1944b], Anderson and Rubin [1956], Tumura and Fukutomi [1968], Tumura and Sato [1980] and [1985], Williams [1981] and Kano [1986]. Here we will present only a condition called Anderson and Rubin's sufficient condition [1956] on the identifiablity, for which we develop the proof by Ihara and Kano [1986] because it is a simpler and more direct proof than that provided by Anderson and Rubin [1956].

THEOREM 1.1 (Anderson and Rubin [1956])

When any row vector of the matrix Λ in (1.2) is deleted, if there remain two disjoint non-singular submatrices of order k, then the matrix Ψ is uniquely determined.

PROOF. It is sufficient only to prove that the ψ_1 can be uniquely determined from the matrix Σ.

Partition the matrices Σ, Λ and Ψ as follows:

$$\Sigma = \begin{array}{cc} \left[\begin{array}{cccc} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \Sigma_{22} & \Sigma_{23} & \Sigma_{24} \\ \sigma_{31} & \Sigma_{32} & \Sigma_{33} & \Sigma_{34} \\ \sigma_{41} & \Sigma_{42} & \Sigma_{43} & \Sigma_{44} \end{array}\right] & \begin{array}{c} 1 \\ k \\ k \\ p-2k-1 \end{array} \\ \begin{array}{cccc} 1 & k & k & p-2k-1 \end{array} & \end{array},$$

$$(1.4)$$

$$\Lambda = \begin{array}{cc} \left[\begin{array}{c} \lambda_1 \\ \Lambda_2 \\ \Lambda_3 \\ \Lambda_4 \end{array}\right] & \begin{array}{c} 1 \\ k \\ k \\ p-2k-1 \end{array} \\ k & \end{array} \text{ and } \Psi = \begin{array}{cc} \left[\begin{array}{cccc} \psi_1 & & & 0 \\ & \Psi_2 & & \\ & & \Psi_3 & \\ 0 & & & \Psi_4 \end{array}\right] & \begin{array}{c} 1 \\ k \\ k \\ p-2k-1 \end{array} \\ \begin{array}{cccc} 1 & k & k & p-2k-1 \end{array} & \end{array}.$$

Substituting the Σ, Λ and Ψ in the form of (1.4) into (1.2), we have

$$\begin{aligned} &\sigma_{11} = \lambda_1 \lambda_1' + \psi_1, \quad \sigma_{12} = \lambda_1 \Lambda_2', \\ &\sigma_{31} = \Lambda_3 \lambda_1' \text{ and } \Sigma_{32} = \Lambda_3 \Lambda_2'. \end{aligned} \qquad (1.5)$$

Under the assumption we can suppose that the two submatrices Λ_2 and Λ_3 are non-singular. Using the non-singulality of the matrix $\Sigma_{32} = \Lambda_3\Lambda_2'$ and relations in (1.5) , we have

$$\sigma_{12}\Sigma_{32}^{-1}\sigma_{31} = \lambda_1\Lambda_2' \ (\Lambda_3\Lambda_2')^{-1}\Lambda_3\lambda_1' = \lambda_1\lambda_1' .$$

Substituting this into $\sigma_{11} = \lambda_1\lambda_1' + \psi_1$ yields

$$\psi_1 = \sigma_{11} - \sigma_{12}\Sigma_{32}^{-1}\sigma_{31}. \qquad (1.6)$$

Now, let $\Sigma = \Lambda\Lambda' + \Psi = AA' + V$, and partition A and V into the same fashion as in (1.4) . Then the non-singularity of the matrix $\Sigma_{32} = A_3A_2'$ leads to that of the two submatrices A_2 and A_3. Thus we can obtain $\psi_1 = v_1$ by going along the same lines as the derivation of (1.6) . This completes the proof. [Q.E.D]

1.3 ROTATION OF FACTORS

Suppose that given a p x p positive definite matrix Σ, we can express it as $\Lambda\Lambda' + \Psi$ and the Ψ is unique. If the number of common factors k is equal to 1, then Λ reduces to a column vector of p components. It is unique, apart from a possible change of sign of all its components, which corresponds merely to changing the sign of the factors. Such sign changes are merely trivial and we shall ignore them through this thesis.

When $k > 1$, there is an infinity of choices for Λ. For equations (1.1) and (1.2) are still satisfied if we replace f and Λ by Tf and $\Lambda T'$, respectively, where T is any orthogonal matrix of order k. In the terminology of factor analysis, this corresponds to a rotation of the factors. To eliminate the indeterminacy of the solution Λ due to the arbitrariness of the rotation, we need to put a restriction on Λ. Among various restrictions proposed so far we shall use the restriction that

> For a given positive definite matrix B of order p
> $\Lambda' B \Lambda$ is a diagonal matrix.

For other restrictions see Anderson and Rubin [1956].

CHAPTER 2

COMPUTATIONAL ALGORITHMS FOR FACTOR ANALYSIS

2.0 GENERAL REMARKS

In this chapter we first review computational algorithms which are available for determining either the LS or ML estimates of the factor loading matrix Λ and the unique variance matrix Ψ, and then discuss a new algorithm which was proposed in Okamoto and Ihara [1984] .

Suppose that a random sample $x_1, \ldots, x_n$ is drawn from a population according to a random factor analysis model and the problem is to estimate the unknown parameters in the model, μ, Λ and Ψ, under the assumption that the number of common factors is known. The population mean μ can be easily estimated by the sample mean so that our problem is only to estimate (Λ, Ψ) by using the sample covariance matrix S. This is usually accomplished by minimizing a suitable function which measures the degree of descrepancy between S and the population covariance matrix Σ; that is, the estimates $(\hat{\Lambda}, \hat{\Psi})$ are determined as a solution which minimizes the function subject to the condition (1.2) . Since the derivatives of the function are usually non-linear with respect to (Λ, Ψ) , we need an iterative procedure in order to obtain the solution $(\hat{\Lambda}, \hat{\Psi})$. Therefore, we first review computational algorithms proposed so far.

The principal factor analysis method is the simplest of all. This method is obtained by applying the notion of principal component analysis to factor analysis. It is well known that in difficult situations the iterative process needs a great number of iterations before attaining the convergence so that a prescribed number gives rise to stop the iteration earlier for other stopping criteria. Such a disadvantage is seen also in the iterative methods described by Lawley [1940] and Hemmerle [1963] to obtain the ML estimates.

Jöreskog [1967] was a breakthrough which presented an efficient algorithm to determine the ML estimates. Three more papers dealing with the subject appeared after it: Jennrich and Robinson [1969], Clarke [1970] and Lee and Jennrich [1979]. On the other hand, after Harman and Jones [1966] described an algorithm called MINRES (minimizing residuals) for a least-squares method, there appeared five more algorithms to deal with a family of least-squares methods: Derflinger [1969] and [1979], Jöreskog and Goldberger [1972], Jöreskog [1977] and Okamoto and Ihara [1983b]. They may be reviewed from the following points of view.

(a) Choice of variables. There are two choices for the variables in which the iteration proceeds: Ψ or (Λ and Ψ). For the former case, only the diagonal elements of Ψ are used in the iteration, whereas for the latter case all elements of Ψ jointed with Λ are used.

(b) Restriction on Λ. As a positive definite weight matrix B in Section 1.3, Jöreskog [1967], Clarke [1970] and Jöreskog and Goldberger [1972] chose $B = \Psi^{-1}$, Derflinger [1969] used $B = I$ and Jennrich and Robinson [1969] adopted

$B = S^{-1}$ for the first time.

(c) Optimization method. As an algolithm to maximize the likelihood function, Jöreskog [1967] used the method of Fletcher-Powell [1963] which is generally called the Davidon-Fletcher-Powell method (in short DFP). Jennrich and Robinson [1969], Derflinger [1969] and [1979], Clarke [1970] and Jöreskog and Goldberger [1972] used the Newton-Raphson method (NR) which needs second order derivatives. Lee and Jennrich [1979] used the Gauss-Newton method (GN) and stated that the performance of GN is better than that of NR. In maximum likelihood context, GN is usually called Fisher's scoring method in the statistical community. On the other hand, the algorithms used by Harman and Jones [1966] and Okamoto and Ihara [1983b] are the Gauss-Seidel method (GS) and the method of Marquardt [1963], respectively.

The algolithm described by Okamoto and Ihara [1984] is characterized by the triplet (Ψ, W, GN) in order to (a), (b) and (c) with $W = S^{-1}$ or $= D_s^{-2}$, where $D_s^{-2} = (D_s^2)^{-1}$ and D_s^2 is a diagonal matrix whose diagonal elements are the sample variances of the observed variables. These considerations are summarized in Table 2.1.

2.1 LEAST-SQUARES AND MAXIMUM LIKELIHOOD METHODS

In the least-squares approach in factor analysis it is required to determine the value of (Λ, Ψ) which minimizes the function

$$L(\Lambda,\Psi) = \frac{1}{2}\,\mathrm{tr}\,[(S - \Sigma)W]^2 \qquad (2.1)$$

subject to the condition (1.2) , where W is a positive definite weight matrix of order p.

Choices of W give rise to a family of least-squares methods and two most important members of the family are the simple (SLS) and weighted (WLS) least-squares methods which correspond to the choices $W = D_s^{-2}$ and $W = S^{-1}$,respectively. On the other hand, the maximum likelihood method (ML) attempts to maximize the likelihood function or equivalently minimize the function

$$M(\Lambda,\Psi) = \mathrm{tr}\,(S^{-1}\Sigma) - \log|S\Sigma^{-1}| - p \qquad (2.2)$$

again subject to the condition (1.2) , where we need the normality assumption with respect to the distribution of the observable variables.

Let us use the letter F in general to mean either the function L or M. Lee and Jennrich [1979] considered to minimize the function $F(\Lambda,\Psi)$ with respect to (Λ,Ψ) directly by using the Gauss-Newton method. Here the minimization of $F(\Lambda,\Psi)$ is done in two steps, as was initiated by Jöreskog [1967] .

First F is minimized with respect to Λ for given Ψ. The minimizer $\Lambda(\Psi)$ is determined by the equation

$$\partial F / \partial \Lambda = 0 . \qquad (2.3)$$

Then the function

$$G(\Psi) = F(\Lambda(\Psi), \Psi) \tag{2.4}$$

is minimized with respect to Ψ.

Since

$$\partial L / \partial \Lambda = -2W(S - \Sigma)W\Lambda,$$

$$\partial M / \partial \Lambda = -2\Sigma^{-1}(S - \Sigma)\Sigma^{-1}\Lambda$$

(see, e.g., equations (8) and (9) in Jöreskog [1977]) ,equation (2.3) is written as

$$(S - \Sigma)W\Lambda = 0 \tag{2.5}$$

with $W = D_s^{-2}$, S^{-1} and Σ^{-1} for SLS, WLS and ML, resecptively.

Now, equation (2.5) for $W = \Sigma^{-1}$ is shown to be equivalent to equation (2.5) with $W = S^{-1}$ (see, e.g., equation (6) in Jennrich and Robinson [1969]) so that we have

$$\begin{aligned} W &= D_s^{-2} \quad \text{for SLS,} \\ &= S^{-1} \quad \text{for WLS and ML} \end{aligned} \tag{2.6}$$

as far as equation (2.5) is concerned.

Substitution of (1.2) into (2.5) yields

$$(S - \Psi) W\Lambda = \Lambda (\Lambda' W\Lambda)$$

or equvalently

$$W^{1/2} (S - \Psi) W^{1/2} (W^{1/2}\Lambda) = W^{1/2}\Lambda (\Lambda' W\Lambda) . \quad (2.7)$$

To eliminate the degree of freedom of the rotation for Λ, we can postulate that

$\Lambda' W\Lambda$ is a diagonal matrix,

which was explained in Section 1.3. Then (2.7) means that diagonal elements of the matrix $\Lambda' W\Lambda$ are eigenvalues of the matrix

$$W^{1/2} (S - \Psi) W^{1/2} \quad (= S^{*}, \text{ say }) ,$$

whereas column vectors of the matrix $W^{1/2} \Lambda$ are eigenvectors of S^{*}. If we denote by Γ_1 the diagonal matrix determined by the k largest eigenvalues of S^{*} and by V_1 the p x k matrix with the associated orthonormal eigenvectors as columns, then

$$\Lambda = W^{-1/2} V_1 \Gamma_1{}^{1/2} . \quad (2.8)$$

2.2 PARTIAL GAUSS-NEWTON METHOD

We shall first develop a general methodology and later discuss the three particular cases, SLS, WLS and ML. Before attempting minimization of the function $G(\Psi)$ in (2.4), let us change the notations. Let λ be the pk x 1 vector consisting of all elements λ_{ir} of Λ $(i = 1, \ldots, p;\ r = 1, \ldots, k)$ and ψ be the p x 1 vector of diagonal elements ψ_i of Ψ $(i = 1, \ldots, p)$. Instead of $F(\Lambda(\Psi), \Psi)$, $\Lambda(\Psi)$ and $G(\Psi)$, we write $F(\lambda(\psi), \psi)$, $\lambda(\psi)$ and $G(\psi)$, respectively.

Assume that we need an interative procedure to minimize $G(\psi)$. For each cycle of interations, let ψ be an initial value and $\Delta\psi$ an increment of ψ minimizing $G(\psi)$ approximately, which will be determined later. Corresponding to $\Delta\psi$, the function $\lambda(\psi)$ increases, up to the first order, by the increment

$$\Delta\lambda = J\Delta\psi, \qquad (2.9)$$

where $J = \Delta\lambda / \Delta\psi = (\partial\lambda_{ir} / \partial\psi_j)$ is a pk x p matrix called Jacobian matrix. By Taylor's expansion up to the second order we have

$$\begin{aligned} G(\psi + \Delta\psi) &= F(\lambda(\psi + \Delta\psi), \psi + \Delta\psi) \\ &= F(\lambda(\psi), \psi) + (\Delta\psi' \frac{\partial F}{\partial\psi} + \Delta\lambda' \frac{\partial F}{\partial\lambda}) \\ &\quad + \frac{1}{2}(\Delta\psi' \frac{\partial^2 F}{\partial\psi\,\partial\psi'}\Delta\psi + \Delta\psi' \frac{\partial^2 F}{\partial\psi\,\partial\lambda'}\Delta\lambda \end{aligned}$$

$$+ \Delta\lambda' \frac{\partial^2 F}{\partial\lambda\,\partial\psi'}\Delta\psi + \Delta\lambda' \frac{\partial^2 F}{\partial\lambda\,\partial\lambda'}\Delta\lambda) .$$

Jöreskog [1977] used the Newton-Raphson method to minimize $G(\psi + \Delta\psi)$, whereas we use the Gauss-Newton method. As was shown in Lee and Jennrich [1979], it is equivalent to approximating the matrix

$$\begin{bmatrix} \dfrac{\partial^2 F}{\partial\psi\,\partial\psi'} & \dfrac{\partial^2 F}{\partial\psi\,\partial\lambda'} \\ \dfrac{\partial^2 F}{\partial\lambda\,\partial\psi'} & \dfrac{\partial^2 F}{\partial\lambda\,\partial\lambda'} \end{bmatrix}$$

by its expectation, which will be denoted by

$$U = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix} .$$

Thus, using (2.3) and (2.9), the last expression of $G(\psi + \Delta\psi)$ can be approximated by

$$G(\psi) + \Delta\psi' \cdot g + \frac{1}{2}\Delta\psi' \cdot H \cdot \Delta\psi , \qquad (2.10)$$

where g and H are the gradient vector of F and pseudo Hessian matrix of G defined by

$$g = \frac{\partial F}{\partial \psi} \text{ and } H = U_{11} + U_{12}J + J'U_{21} + J'U_{22}J,$$

respectively. The expression (2.10) is minimized at

$$\Delta\psi = -H^{-1}g, \qquad (2.11)$$

provided that the matrix H is positive definite. Thus, we need the vector g and two matrices U and J in order to determine the increment $\Delta\psi$ by (2.11). This method was called the partial Gauss-Newton method in Okamoto and Ihara [1984], since the Gauss-Newton method is applied only to the parameter ψ, not to the whole parameters (Λ, Ψ) as in Lee and Jennrich [1979]. In case the iteration should diverge, a suitable constant multiple of the unit matrix might be added to H in (2.11) by the up-and-down Marquardt algorithm described in Okamoto and Ihara [1983a] or [1983b].

Now, we turn to the particular cases, SLS, WLS and ML. From relevant expressions given in Lee and Jennrich [1979], Section 6 and (3.5), elements of g, U_{11}, U_{21} and U_{22} can be written as

$$g_i = [W(\Sigma - S)W]_{ii},$$

$$U(\psi_i, \psi_j) = W_{ij}^2,$$

$$U(\lambda_{ir}, \psi_j) = 2W_{ij}(W\Lambda)_{jr},$$

$$U(\lambda_{ir}, \lambda_{jt}) = 2W_{ij}(\Lambda' W\Lambda)_{rt}$$

$$+ 2(W\Lambda)_{it}(W\Lambda)_{jr},$$

for i, j = 1, . . . ,p; r, t = 1, . . . , k, where the symbol A_{ij} stands for the (i,j) element for any matrix A and $W = D_s{}^{-2}$, S^{-1} and Σ^{-1} for SLS, WLS and ML, respectively. It is noted that the matrix U for ML is twice the Fisher information matrix per sample element.

We have finally to calculate J. Let $\gamma_1 \geqq \cdots \geqq \gamma_p$ be the eigenvalues of the matrix S^* and let V be the orthogonal matrix of order p with the associated orthonormal eigenvectors as columns. Then it will be shown in Appendix that the elements of J are given by

$$\frac{\partial \lambda_{ir}}{\partial \psi_j}$$

$$= \gamma^{1/2} \sum_{\substack{t=1 \\ t \neq r}}^{p} (\gamma_t - \gamma_r)^{-1} (W^{1/2}V)_{it} (W^{1/2}V)_{jt} (W^{1/2}V)_{jr}$$

$$- \frac{1}{2}\gamma_r{}^{-1/2} (W^{-1/2}V)_{ir} [(W^{1/2}V)_{jr}]^2, \qquad (2.12)$$

where $W = D_s{}^{-2}$ for SLS and $W = S^{-1}$ for WLS and ML by (2.6).

2.3 COMPARISON OF PERFORMANCE

We want to see how efficient the new method is in analyzing the real data as compared with the Newton-Raphson method. Let us use the two data sets from Rao [1956, p.110] and Harman [1960, p.373] which were used by Jennrich and Robinson [1969] to evaluate their algorithm based on the Newton-Raphson method.

The iterative processes for these data are shown in Tables 2.2 and 2.3, where the solutions obtained by Rao and Harman were chosen as the starting values in each case. The terms RMS (g) and RMS ($\Delta\psi$) denote the root mean square of the components g_i's and $\Delta\psi_i$'s, respectively. The stopping rule of iterations was that at least one of them is less than 10^{-4}. The Δ_i's are defined to be square roots of the ψ_i's.

The new method is readily seen to be very efficient by comparing these tables with Tables 1 and 3 in Jennrich and Robinson [1969] . It has also another advantage as follows. Starting from Harman's solution $\Delta_8 = 0.451$, they failed to attain the convergence but we arrived at the same solution as before at the seven cycle, only one cycle more than in Table 2.3.

The superiority of the partial Gauss-Newton method to the Newton-Raphson method was anticipated by us, since the direct Gauss-Newton method due to Lee and Jennrich [1979] was found better than the Newton-Raphson method.

2.4 STRUCTURE OF IMPROPER SOLUTION

As we can see in Tables 2.2 or 2.3, sometimes we arrive at a solution ($\hat{\Lambda}$, $\hat{\Psi}$) in which some diagonal elements of $\hat{\Psi}$ are zero. Such a solution which is obtained on the boundary of the admissible parameters space is called improper solution or Heywood case. Many Monte Carlo studies have been carried out in order to investigate causes of improper solutions in maximum likelihood factor analysis. Jöreskog [1967] attempted to clarify the structure of such a solution in maximum likelihood factor analysis and obtained a structure under the assumption that the first m ($\leqq$ k) uniquenesses are equal to zero.

In this section we extend his result to the case where the estimates of the first m ($\leqq$ k) uniquenesses are equal to zero.

THEORM 4.1 (Ihara [1986])

Assume that the first m ($\leqq$ k) diagonal elements of the matrix $\hat{\Psi}$, the ML estimator of Ψ , are equal to zero and partition the matrices S , $\hat{\Lambda}$, the ML estimator of Λ , and $\hat{\Psi}$ as follows:

$$S = \begin{array}{cc} \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} & \begin{array}{c} m \\ p-m \end{array} \\ \begin{array}{cc} m & p-m \end{array} & \end{array} , \quad \hat{\Lambda} = \begin{array}{cc} \begin{bmatrix} \hat{\Lambda}_{11} & \hat{\Lambda}_{12} \\ \hat{\Lambda}_{21} & \hat{\Lambda}_{22} \end{bmatrix} & \begin{array}{c} m \\ p-m \end{array} \\ \begin{array}{cc} m & k-m \end{array} & \end{array}$$

$$\text{and} \quad \hat{\Psi} = \begin{array}{cc} \left[\begin{array}{cc} 0 & 0 \\ & \\ 0 & \hat{\Psi}_2 \end{array}\right] & \begin{array}{l} m \\ \\ p-m \end{array} \\ \begin{array}{cc} m & p-m \end{array} & \end{array} . \tag{2.13}$$

Then the ML estimators obtained under the restriction that $\hat{\Lambda}' S^{-1} \hat{\Lambda}$ is a diagonal matrix are given as follows:

(a) the submatrix $\hat{\Lambda}_{12}$ becomes zero,

(b) the two submatrices $\hat{\Lambda}_{11}$ and $\hat{\Lambda}_{21}$ are determined by

$$\hat{\Lambda}_{11} = S_{11} Q \quad \text{and} \quad \hat{\Lambda}_{21} = S_{21} Q , \tag{2.14}$$

where Q is an $m \times m$ matrix such that $Q Q' = S_{11}^{-1}$, and

(c) the two submatrices $\hat{\Lambda}_{22}$ and $\hat{\Psi}_2$ are determined as a solution which minimizes the function

$$M^{*}(\hat{\Lambda}_{22}, \hat{\Psi}_2) = \mathrm{tr}\,(S_{22\cdot 1} \Sigma_{22}^{-1}) + \log | S_{22\cdot 1}^{-1} \Sigma_{22} | - (p - m)$$

subject to the condition $\Sigma_{22} = \Lambda_{22} \Lambda_{22}' + \Psi_2$, where

$$S_{22\cdot 1} = S_{22} - S_{21}S_{11}^{-1}S_{12}.$$

PROOF. Let $d_1 \leqq \cdots \leqq d_k$ be the k smallest eigenvalues of the matrix $S^{-1/2}\hat{\Psi}S^{-1/2}$ (= N , say) , and denote by V the p x k matrix with the associated orthonormal eigenvectors as columns. Then the $\hat{\Lambda}$ in (2.5) is shown to be given by

$$\hat{\Lambda} = S^{1/2}V(I - D)^{1/2}, \qquad (2.15)$$

where D is a diagonal matrix consisting of the d_i's. See, for example, (60) in Jöreskog [1977] .

Partition the matrices D , V , $S^{-\frac{1}{2}}$ and $S^{\frac{1}{2}}$ as follows:

$$D = \begin{array}{cc} \left[\begin{array}{cc} D_1 & O \\ O & D_2 \end{array}\right] & \begin{array}{c} m \\ k-m \end{array} \\ \begin{array}{cc} m & k-m \end{array} & \end{array}, \quad V = \begin{array}{c} (V_1, V_2), \\ \begin{array}{cc} m & k-m \end{array} \end{array} \qquad (2.16)$$

$$S^{-1/2} = \begin{array}{cc} \left[\begin{array}{cc} A_{11} & A_{12} \\ A_{21} & A_{22} \end{array}\right] & \begin{array}{c} m \\ p-m \end{array} \\ \begin{array}{cc} m & p-m \end{array} & \end{array} \text{ and } S^{1/2} = \begin{array}{cc} \left[\begin{array}{cc} B_{11} & B_{12} \\ B_{21} & B_{22} \end{array}\right] & \begin{array}{c} m \\ p-m \end{array} \\ \begin{array}{cc} m & p-m \end{array} & \end{array}.$$

Then from the assumption we have

$$N V_1 = V_1 D_1 = 0 \qquad (2.17)$$

and

$$N V_2 = V_2 D_2 . \qquad (2.18)$$

Premultiplication of (2.17) by the matrix (B_{21}, B_{22}) and then by $\hat{\Psi}_2^{-1}$ leads to

$$(A_{21}, A_{22}) V_1 = 0 ,$$

which implies that the matrix V_1 is represented as

$$V_1 = (B_{11}', B_{21}')' Q \qquad (2.19)$$

with an m x m matrix Q. Substitution of (2.19) into $V_1' V_1 = I$ and then using $(B_{11}', B_{21}')(B_{11}', B_{21}')' = S_{11}$ yields $Q Q' = S_{11}^{-1}$. Thus we have (b) from (2.15), (2.16) and (2.19).

On the other hand, premultipling (2.18) by (B_{11}, B_{12}) and then using $(B_{11}, B_{12})(A_{12}', A_{22}')' = 0$, (2.15) and (2.16), we obtain (a).

The ML estimator $\hat{\Sigma} = \hat{\Lambda}\hat{\Lambda}' + \hat{\Psi}$ is then given by

$$\hat{\Sigma} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{21} S_{11}^{-1} S_{12} + \hat{\Sigma}_{22} \end{bmatrix} . \qquad (2.20)$$

Partition S^{-1} and $\hat{\Sigma}$ as

$$S^{-1} = \begin{array}{cc} \left[\begin{array}{cc} S^{11} & S^{12} \\ S^{21} & S^{22} \end{array}\right] & \begin{array}{l} m \\ p-m \end{array} \\ \begin{array}{cc} m & p-m \end{array} & \end{array} \quad \text{and} \quad \hat{\Sigma}^{-1} = \begin{array}{cc} \left[\begin{array}{cc} \hat{\Sigma}^{11} & \hat{\Sigma}^{12} \\ \hat{\Sigma}^{21} & \hat{\Sigma}^{22} \end{array}\right] & \begin{array}{l} m \\ p-m \end{array} \\ \begin{array}{cc} m & p-m \end{array} & \end{array} .$$

Noting that S^{22} and $\hat{\Sigma}^{22}$ are equal to $S_{22\cdot 1}{}^{-1}$ and $\hat{\Sigma}_{22}{}^{-1}$, respectively, we can show that

$$S^{-1}\hat{\Sigma} = \left[\begin{array}{cc} I & S_{11}{}^{-1}S_{12}(I - S_{22\cdot 1}{}^{-1}\hat{\Sigma}_{22}) \\ 0 & S_{22\cdot 1}{}^{-1}\hat{\Sigma}_{22} \end{array}\right]$$

and

$$\mathrm{tr}[(S - \hat{\Sigma})\hat{\Sigma}^{-1}] = \mathrm{tr}[(S_{22\cdot 1} - \hat{\Sigma}_{22})\hat{\Sigma}_{22}{}^{-1}] .$$

Using these results at the function M in (2.2) reduces M to the function M* and we obtain (c) . [Q.E.D.]

Now, the generalized least-squares estimators $(\hat{\Lambda}, \hat{\Psi})$ are defined as a solution which minimizes the function

$$GL(\Lambda, \Psi) = \frac{1}{2}\mathrm{tr}[(S - \Sigma)S^{-1}]^2$$

subject to the condition (1.2) and thus the solution $\hat{\Lambda}$ for given $\hat{\Psi}$ is represented as the same form as that of the ML estimator $\hat{\Lambda}$ in (2.15) . Therefore we can prove the following by going along the same lines as the proof for the first theorem.

THEOREM 4.2 (Ihara [1986])

Assume that the first m ($\leqq$ k) diagonal elements of $\hat{\Psi}$ are zero and partition the matrices S , $\hat{\Lambda}$ and $\hat{\Psi}$ similarly as (2.13) . Then the submatrices $\hat{\Lambda}_{11}$, $\hat{\Lambda}_{12}$ and $\hat{\Lambda}_{21}$ are given by

$$\hat{\Lambda}_{11} = S_{11}Q \text{ and } \hat{\Lambda}_{21} = S_{21}Q$$

respectively, Q being an m x m matrix such as $QQ' = S_{11}^{-1}$, and the two submatrices $\hat{\Lambda}_{22}$ and $\hat{\Psi}_2$ are determined as a solution which minimizes the function

$$GL^{*}(\Lambda_{22},\Psi_{2}) = \frac{1}{2}\mathrm{tr}\left[(S_{22\cdot 1} - \Sigma_{22})S_{22\cdot 1}^{-1}\right]^{2}$$

subject to the condition $\Sigma_{22} = \Lambda_{22}\Lambda_{22}' + \Psi_{2}$.

2.5 CONCLUDING REMARKS

It was found in the aspect of computational performance that the Gauss-Newton method, partial or direct, is better than the Newton-Raphson method. Moreover, it will be found at the next chapter that SLS is better than WLS or ML at least for a small sample size, not exceeding 300.

The author believes that these two findings result from the same reason that a simpler method performs better than a more complicated method does. In fact, for the former issue, the Newton-Raphson method utilizes second order derivatives and hence is more complicated than the Gauss-Newton method which does with only first order derivatives. For the latter issue, the weight matrix $W = D_s^{-2}$ used in SLS is far simpler than the weights $W = S^{-1}$ or Σ^{-1} which appear for WLS or ML. The more complicated a quantity is, the more sensitive it seems to be to random fluctuations at earlier and unstable stages in an iterative process.

An application of the up-and down Marquardt algorithm described by Okamoto and Ihara [1983a] or [1983b] to the partial Gauss-Newton method is advised for a practitioner of factor analysis in dealing with only one sample, since it may be more likely to lead to a proper solution with a moderate loss in computing time.

APPENDIX

Calculation of the Jacobian matrix

In order to prove (2.12) , we generalize the argument in Jennrich and Robinson [1969] which dealt with only the maximum likelihood method. Define $K = W^{\frac{1}{2}}$ and $\Gamma = \mathrm{diag}(\gamma_1, \gamma_2, \ldots, \gamma_p)$, the eigenvalue matrix of the matrix S^*.

Then Γ and the matrix V are determined by the simultaneous equations

$$K(S-\Psi)KV = V\Gamma,$$

$$V'V = I$$

as functions of Ψ. Let $\Delta\Gamma$ and ΔV be the increments of Γ and V, respectively, when Ψ increases by $\Delta\Psi$. Then

$$K(S-\Psi-\Delta\Psi)K(V+\Delta V) = (V+\Delta V)\ (\Gamma+\Delta\Gamma)$$

$$(V+\Delta V)'\ (V+\Delta V) = I.$$

Defining $\Delta Q = V' \cdot \Delta V$, we have from the first equation that

$$(\Delta Q)_{tr} = (\gamma_t - \gamma_r)^{-1}\sum_{j=1}^{p}(KV)_{jt}(KV)_{jr}\Delta\psi_j,\ (t\neq r) \quad (A.1)$$

$$\Delta\gamma_r = -\sum_{j=1}^{p}[(KV)_{jr}]^2\Delta\psi_j \quad (r=1,\ .\ .\ .\ ,p) \quad (A.2)$$

and from the second equation that

$$(\Delta Q)_{rr} = 0 \quad (r=1,\ .\ .\ .\ ,p). \quad (A.3)$$

The relation $\Delta V = V \cdot \Delta Q$, which is equivalent to the definition of ΔQ, yields that

$$(\Delta V)_{mr} = \sum_{t=1}^{p} V_{mr} (\Delta Q)_{tr} \quad \text{for any } m \text{ and } r. \qquad (A.4)$$

Expression (2.8) or $\Lambda = K^{-1} V_1 \Gamma_1^{1/2}$ implies that

$$\Delta \Lambda = K^{-1} (\Delta V_1 \cdot \Gamma_1^{1/2} + \frac{1}{2} V_1 \Gamma_1^{1/2} \cdot \Delta \Gamma_1)$$

or elementwise

$$\Delta \lambda_{ir} = \sum_{m=1}^{p} (K^{-1})_{im} \{ (\Delta V)_{mr} \gamma_r + \frac{1}{2} V_{mr} \gamma_r^{1/2} \Delta \gamma_r \} .$$

Using (A.1) through (A.4), we obtain

$$\begin{aligned}
\Delta \lambda_{ir} &= \sum_{m=1}^{p} (K^{-1})_{im} \{ \sum_{t \neq r} V_{mt} \gamma_r^{1/2} (\gamma_t - \gamma_r)^{-1} \\
&\qquad \cdot \sum (K V)_{jt} (K V)_{jr} \Delta \psi_j \\
&\qquad - \frac{1}{2} V_{mr} \gamma_r^{-1/2} \sum [(K V)_{jr}]^2 \Delta \psi_j \} \\
&= \sum \{ \gamma_r^{1/2} \sum (\gamma_t - \gamma_r)^{-1} (K^{-1} V)_{it} (K V)_{jt} (K V)_{jr} \\
&\quad - \frac{1}{2} \gamma_r^{-1/2} (K^{-1} V)_{ir} [(K V)_{jr}]^2 \} \Delta \psi_j .
\end{aligned}$$

This implies (2.12) [Q.E.D.]

CHAPTER 3

EXPERIMENTAL COMPARISON OF LEAST-SQUARES AND MAXIMUM LIKELIHOOD METHODS IN FACTOR ANALYSIS

3.0 GENERAL REMARKS

In this chapter three popular methods to estimate the unknown parameters in the factor analysis model, the simple (SLS) and weighted (WLS) least-squares methods and the maximum likelihood method (ML) , are compared by a Monte Calro study.

The most popular estimation methods in factor analysis may be ML proposed by Lawley [1940] under the normality assumption and the generalized least-squares method (GLS) , being equivalent to WLS,by Jöreskog and Goldberger [1972] . For it is well known that ML leads to asymptotically best estimators for large samples and it is shown by Jöreskog and Goldberger [1972] that under the normality assumption the GLS estimators have the same asymptotic properties as the ML estimators. Here there are two problems. One problem is that the situation where the sample size is infinitely large is only hypothetical so that the actual sample size is always finite and usually less than a few hundreds. Another problem is that the normality assumption has been questioned in analyzing the real data whether to be realistic or not. In such circumstances, the computatinal simplicity may change to the good performance for small samples and thus SLS which is asymptotically less efficient than ML or WLS may perform most favorably for small samples. Therefore the author believes that it is worth while comparing the three estimation methods, SLS, WLS and ML.

We used the following three criteria to compare these three methods. The first criterion is the frequency of occurence of improper solutions or nonconvergent solutions. The second is the number of iterations before attaining the convergence and the third is the estimation error of the solution from the true value of the parameters. It is very difficult to deal with any of these criteria analytically. No exact expression to evaluate the sampling error, for instance, variance or mean squared error, has yet been obtained. Thus we have decided to carry out experimental comparison in terms of these three criteria.

In Ihara and Okamoto [1985] we used only one numerical model based on Emmett's data [1949] so that we carried out an additional experiment to make the conclusions in Ihara and Okamoto [1985] more reliable.

3.1 ALGORITHM

We want to compare the three estimation methods, SLS, WLS and ML, so that as an algorithm for the least-squares and maximum likelihood methods, we adopted the partial Gauss-Newton method presented by Okamoto and Ihara [1984], which replaces the Newton-Raphson method used in Jöreskog [1977] by the Gauss-Newton method which performed well in Lee and Jennrich [1979].

As is shown by Krane and McDonald [1978], the three estimation methods are scale invariant so that there are two situations in estimating the parameters (Λ, Ψ) which satisfies (1.2) according as Σ is a covariance matrix or a correlation matrix. In the former case, the equation (2.5) to determine $\Lambda(\Psi)$ is given by

$$(S - \Sigma) W\Lambda = 0 ,$$

where S is the sample covariance matrix and W equals D_s^{-2}, S^{-1} and Σ^{-1} for SLS, WLS and ML, respectively. In the latter case, $\Lambda(\Psi)$ is determined by

$$(R - \Sigma) W\Lambda = 0 ,$$

where R is the sample correlation matrix and $W = I$, R^{-1} and Σ^{-1} for SLS, WLS and ML, respectively. We applied the partial Gauss-Newton method to the function G in (2.4) , using the sample correlation matrix.

If a cycle in the iterative process assigned a negative value to some component of uniquenesses (Heywood case) , we forced the solution to be proper by shortening the increment of the estimate in the cycle so that the new point would lie on the boundary of proper solutions and then continued the iteration.

As an initial value of Ψ at which iterative computation starts, we adopted SMC (squared multiple correlation) due to Guttman [1956] which is used in many studies including Okamoto and Ihara [1983b] . The stopping rule of the iteration was that either the number of iterations exceed 30 or

$$\min(\mathrm{RMS}(g), \mathrm{RMS}(\Delta\Psi)) < 10^{-4}$$

where RMS stands for the root mean square of the components of a vector, g and $\Delta\Psi$ denoting the gradient vector and the successive difference of the diagonal elements of Ψ, respectively, in the iteration process.

3.2 SPECIFICATION OF THE EXPERIMENTS

There are two major factors which are likely to affect experimental results; the sample size and the uniquenesses or communalities. We treated three levels of the sample size, 100, 300 and 1000 as representatives of small, moderate and large samples and two levels, small and large, for the uniquenesses defined later.

It is an intriguing problem to decide upon the population model to be used in a Monte Carlo study. In Ihara and Okamoto [1985] we thought that in order that a comparison based on a single numerical model would be convincing enough, the model should be familiar to readers. After looking through the literature, we found that Emmett's data [1949] with $p = 9$ is referred to most frequently so that we adopted it as a numerical model, Model 1, which provided the basis of our experiments.

By rounding the maximum likelihood estimate $\hat{\Lambda}$ from Lawley and Maxwell [1971, p.42] , where the number of common factors is assumed known to be 3, to the nearest tenth we set

$$\Lambda_1' = \begin{bmatrix} 0.7 & 0.7 & 0.5 & 0.8 & 0.7 & 0.8 & 0.7 & 0.4 & 0.8 \\ 0.3 & 0.2 & 0.3 & -0.3 & -0.3 & -0.4 & 0.4 & 0.3 & 0.4 \\ 0.1 & -0.2 & -0.2 & -0.1 & -0.2 & 0.1 & -0.1 & 0.5 & 0.0 \end{bmatrix} .$$

in Ihara and Okamoto [1985] . However, the entries at two cells (1,8) and (3,4) differ from the exact rouding by 0.1. This modification was done deliberately to treat these cells differently from almost same values at the cells (1,3) and (3,9) , respectively. It is noted that Λ_1 satisfies Anderson and Rubin's sufficient condition on the identifiability. (See Theorem 1.1 in Chapter 1.) The condition that every population variance is unity leads to the unique variance matrix

$$\Psi_1 = \mathrm{diag}\,(0.41,\ 0.43,\ 0.62,\ 0.26,\ 0.38,\ 0.19,\ 0.34,\ 0.50,\ 0.20)$$

and hence the population correlation matrix becomes

$$\Sigma_1 = \Lambda_1 \Lambda_1' + \Psi_1 \quad .$$

The unique variance matrix varied with two levels, small and large. The smaller level was defined by Ψ_1 mentioned above, whereas the larger level was defined by

$$\Psi_2 = \mathrm{diag}\,(0.54,\ 0.56,\ 0.71,\ 0.41,\ 0.51,\ 0.34,\ 0.47,\ 0.57,\ 0.35)$$

which was obtained from Λ_2 by replacing the first row of the matrix Λ_1' by

$$(0.6,\ 0.6,\ 0.4,\ 0.7,\ 0.6,\ 0.7,\ 0.6,\ 0.3,\ 0.7)$$

which is smaller than the original row by 0.1 componentwise. The corresponding correlation matrix is defined by

$$\Sigma_2 = \Lambda_2\Lambda_2' + \Psi_2$$

similarly as for Σ_1.

As a second model, Model 2, we adopted a loading matrix

$$\Lambda' = \begin{bmatrix} a & a/\sqrt{2} & a/\sqrt{2} & a/2 & & & & & \\ & a/\sqrt{2} & & a/2 & a & a/\sqrt{2} & a/\sqrt{2} & a/2 & \\ & & & & & a/\sqrt{2} & a/2 & a/\sqrt{2} & a/2\sqrt{2} \end{bmatrix}$$

with $(p,k) = (10,3)$ which was obtained by modifying the model with $(p,k) = (15,4)$ due to Cliff and Pennell [1967] principally in reducing the value of k. They chose two values of the parameter a, 0.9 and 0.7, to represent two levels of the unique variance matrix, small and large, respectively, which will be denoted by Ψ_1 and Ψ_2 similarly as for the first model. Instead of the exact value of $a/\sqrt{2}$ for $a = 0.9$ and 0.7, however, their approximations 0.64 and 0.50 were used in our experiments as was the case with Cliff and Pennell's study. The corresponding correlation matrices were defined similarly as before.

For every combination of levels of the three conditions,

$$\text{Model}\begin{bmatrix}\text{Model 1}\\ \\ \text{Model 2}\end{bmatrix} \times \Psi \begin{bmatrix}\text{Small}\\ \\ \text{Large}\end{bmatrix} \times n \begin{bmatrix}100\\ 300\\ 1000\end{bmatrix} ,$$

we generated 200 sample correlation matrices drawn from the Wishart distribution $W_p(\Sigma, n-1)$, using Smith and Hocking's program [1972]. Every data matrix was analyzed by the three methods, SLS, WLS and ML.

In addition to assessing the frequency of improper or non-convergent solutions and the number of iterations we also evaluated the error of estimates of the unique variances and factor loadings. For each combination of two levels of the models and Ψ, denote by ψ_i and $\hat{\psi}_i$ the true value and an estimate, respectively, of the i-th uniqueness for $i = 1, \ldots, p$. Then RMSEU (Root Mean Square Error for Uniqueness) stands for the average of

$$[\sum_{i=1}^{p}(\hat{\psi}_i - \psi_i)^2/p]^{1/2}$$

computed across replications in each of the following two cases:

(a) all replications that lead to proper solutions for a particular method in question, or

(b) all replications that lead to proper solutions for all methods.

Bias for Ψ was computed for each i as the average of $\hat{\psi}_i - \psi_i$ across replications only in the case (a).

On the other hand, following Cliff's method [1966] to deal with the error of the estimated loadings, we computed the estimate $\hat{\lambda}_{ir}$ by fitting the solution obtained by each of the three methods to the true value λ_{ir} by least-squares method and defined RMSEL (Root Mean Square Error for Loadings) by the average of

$$[\sum_{i=1}^{p} \sum_{r=1}^{k} (\hat{\lambda}_{ir} - \lambda_{ir})^2 / (pk)]^{1/2}$$

across replications in each of the two cases (a) and (b).

3.3 RESULTS OF THE EXPERIMENTS

First five tables are concerned with Model 1 for two forms of Emmett's model. Table 3.1 shows the proportion of proper (P), improper (IP) and non-convergent (NC) solutions for every combination of levels of the three conditions, method (3 levels; SLS, WLS and ML), unique variance matrix Ψ (2 levels; small and large) and sample size n (3 levels; 100, 300 and 1000) . In terms of this criterion, performance for the methods depends mainly on the value of n and partly on Ψ . When n = 100, SLS is best and WLS and ML behave similarly. When n = 300 for small Ψ or when n = 1000, there are few differences among the three methods. The table also shows that the proportion of IP or of NC for each method decreases as n increases or Ψ decreases. It was found that most of IP or NC solutions took place at the uniqueness ψ_8, though the result is not tablated.

Table 3.2 gives values of the three criteria, median, mean and standard deviation for the number of iterations before attaining convergence across replications in the case (a) . It is seen from the table that every criterion is smallest for SLS as compared with WLS and ML, irrespective of the values on n and Ψ, whereas the latter two perform similarly. The value of every criterion decreases as n increase or as average Ψ entries decreases.

Table 3.3 shows values of RMSEU for every combination of levels of the three conditions, methods, n and Ψ values in the cases (a) and (b) . Conclusions are similar to those for Table 3.1. When n = 100 or when n = 300 with large Ψ, SLS is best of the three methods and the other two are almost alike, while all three show similar performance when n = 300 with small Ψ or when n = 1000. The value of RMSEU decreases as n increases or Ψ decreases. On the other hand, for a given method, n and Ψ, the value of (b) is smaller than that of (a) .

The results for the criterion RMSEL for estimated loadings are shown in Table 3.4 in the same style as in Table 3.3

and interpretations are similar.

Table 3.5 gives the bias of estimated uniqueness in the case (a) when n = 100 or 300. Even though the number of replications was thus made larger than for the case (b) , it seems still to be too small to make the results stable, partly because the information was diluted by computing the bias componentwise. However, two conclusions can be drawn from the table. First, generally speaking, the bias tends to decrease as n increases or average Ψ value decreases. Second, WLS is negatively biased as was indicated by Jöreskog and Goldberger [1972] , the absolute value of the bias being larger than for either SLS or ML, whereas the latter two behave similarly except when n = 100, where SLS has slightly smaller bias than ML does.

The next four tables are concerned with Model 2 adopted from Cliff and Pennell's model [1967] , corresponding to Tables 3.1, 3.2, 3.3 and 3.4 in this order. Throughout all tables, every criterion decreases as n increases or Ψ decreases so that we shall concentrate on other features. Table 3.6 shows that when n = 100 SLS is better than WLS and ML and the latter two are almost similar. Model 2 looks much easier than Model 1 since IP or NC rarely occurs as soon as n attains 300. Table 3.7 shows that the three methods may be arranged in the order of SLS (best) , WLS and ML (wrost) with respect to the number of iterations. Combining Table 3.8 and 3.9, we find that when n = 100 or n = 300 the order of preference is SLS, ML and WLS for large Ψ (a = 0.7) but ML is best for small Ψ (a = 0.9) . When n = 1000, there is not much difference between the performance of the three methods.

3.4. ASYMPTOTIC BIAS OF ESTIMATORS OF THE UNIQUNESS

Analyzing two sets of real data, Jöreskog and Goldberger [1972] indicated that the weighted least-squares estimate of the uniqueness was systematically smaller than the maximum likelihhod estimate so that WLS tends to be negatively biased. Boomsma [1982] and [1985] reported that ML had similar tendency in her Monte Calro study with respect to maximum likelihood factor analysis. We found in our experimental study that the three methods, SLS, WLS and ML, tended to lead to negatively biased estimate of the uniqueness and moreover the tendency was most remarkable for WLS when the sample size n was not large.

In this section we show that the three methods lead to estimators of uniquenesses with negative biases at least when (p,k) = (3,1) .

Let p = 3 and k = 1 throughout this section and denote by triplet (a,b,c) any permutation of (1,2,3) . Suppose that (n - 1) S , S being the sample covariance matrix, is distributed in the Wishart distribution $W(\Sigma, n-1)$. We can postulate that $\sigma_{ab}\sigma_{bc}\sigma_{ca}$ is positive because it is a necessary and sufficient condition on the identifiability of parameters in (1.2) when (p,k) = (3,1) . See also Theorem 5.5 in Anderson and Rubin [1956] . Since S converges to Σ almost surely, we can assume that $s_{ab}s_{bc}s_{ca}$ is positive with probability one when n becomes sufficiently large. Then we can decompose S into a similar form as that of Σ in (1.4) and hence the estimators $(\hat{\Lambda}, \hat{\Psi})$ by the three methods, SLS, WLS and ML, coincide with each other, which are given by

$$\hat{\psi}_a = \hat{\psi}_a(S) = s_{aa} - s_{ab}s_{ca} / s_{bc}, \tag{3.1}$$

for a = 1, 2, 3. Thus we can prove the following.

THEOREM 3,1 The estimator $\hat{\psi}_a$ (a = 1, 2, 3) in (3.1) has the asymptotic bias

$$\mathrm{AB}(\psi_a) = -[\psi_a + (\lambda_a / \lambda_b\lambda_c)^2\psi_b\psi_c] / (n-1) \tag{3.2}$$

up to the order n^{-1}.

PROOF. It is sufficient to prove the theorem for the case a = 1. Then (b,c) = (2,3) in (3.1) so that $\hat{\psi}_1$ is the function of only the variables s_{11}, s_{12}, s_{13} and s_{23}. Let denote by s and σ the vectors $(s_{11}, s_{12}, s_{13}, s_{23})'$ and $(\sigma_{11}, \sigma_{12}, \sigma_{13}, \sigma_{23})'$, respectively. Then Taylor's expansion of $\hat{\psi}_1$ at $s = \sigma$ up to order of n^{-1} is given by

$$\hat{\psi}_1 = \psi_1 + (s - \sigma)' g + (s - \sigma)' H (s - \sigma) / 2 + o_p(n^{-1}), \tag{3.3}$$

where

$$g = \frac{\partial \hat{\psi}_1}{\partial s}(\sigma) = (1, -\sigma_{13}/\sigma_{23}, -\sigma_{12}/\sigma_{23}, \sigma_{12}\sigma_{13}/\sigma_{23}^2)'$$

and

$$H = \frac{\partial^2 \hat{\psi}_1(\sigma)}{\partial s \, \partial s'} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -1/\sigma_{23} & \sigma_{13}/\sigma_{23}^2 \\ 0 & -1/\sigma_{23} & 0 & \sigma_{12}/\sigma_{23}^2 \\ 0 & \sigma_{13}/\sigma_{23}^2 & \sigma_{12}/\sigma_{23}^2 & -2\sigma_{12}\sigma_{13}/\sigma_{23}^3 \end{bmatrix}$$

Thus we have

$$\mathrm{bias}(\hat{\psi}_1) = \mathrm{tr}[H \cdot E(s - \sigma)(s - \sigma)']/2 + o(n^{-1}). \qquad (3.4)$$

Since $(n-1)S$ is distributed in the Wishart distribution $W(\Sigma, n-1)$, we have $E(s - \sigma) = 0$ and $E(s_{ab} - \sigma_{ab})(s_{cd} - \sigma_{cd}) = (\sigma_{ac}\sigma_{bd} + \sigma_{ad}\sigma_{bc})/(n-1)$ for a, b, c, d = 1, 2, 3. Substituting these results and the relation (1.2) which is here $\sigma_{ab} = \lambda_a\lambda_b + \delta_{ab}\psi_a$ for a, b = 1, 2, 3 into the right-hand side of (3.4) yields (3.2.).

[Q.E.D.]

It should be noted that this theorem suggests that the tendency found in our Monte Calro study exists at least in asymptotic point of view.

The vector x of observations may be standardized because of the scale-invariance. Factor analysis of the sample correlation matrix results in the standardized estimator

$$\hat{\psi}_a^* = \hat{\psi}_a / s_{aa}$$

for a = 1, 2, 3. Then we can prove the following by going along the same lines as the proof for the first theorem.

THEOREM 3.2 Let $\psi_a^* = \psi_a / \sigma_{aa}$ and $\lambda_a^* = \lambda_a / \sqrt{\sigma_{aa}}$ for a = 1, 2, 3. Then the estimator $\hat{\psi}_a^*$ has the asymptotic bias

$$AB(\hat{\psi}_a^*) = -[\psi_a^* + (\lambda_a^* / \lambda_b^* \lambda_c^*)^2 \psi_b^* \psi_c^* \\ - 2(\lambda_a^*)^2 \psi_a^*] / (n-1) \qquad (3.5)$$

up to the order of n^{-1}.

The expression in the bracket in (3.5) can be rewritten as

$$(\lambda_a^* / \lambda_b^* \lambda_c^*)^2 \psi_b^* \psi_c^* - \psi_a^* (1 - 2\psi_a^*) ,$$

so that the asymptotic bias for $\hat{\psi}_a^*$ becomes negative if $\psi_a^* \geqq 0.5$.

Finally, the asymptotic relative biases for $\hat{\psi}_a$ and $\hat{\psi}_a^*$ are defined by $AB(\hat{\psi}_a) / \psi_a$ and $AB(\hat{\psi}_a^*) / \psi_a^*$, respectively, and then we obtain that

$$AB(\hat{\psi}_a) / \psi_a = -[1 + (\lambda_a / \lambda_b \lambda_c)^2 \psi_b \psi_c / \psi_a] / (n-1)$$

and

$$AB(\hat{\psi}_a^*) / \psi_a^* = -[1 + (\lambda_a^* / \lambda_b^* \lambda_c^*)^2 \psi_b^* \psi_c^* / \psi_a^* - 2\lambda_a^{*2}] / (n-1) .$$

This fact implies that the asymptotic relative bias for the standardized estimator is closer to zero than that for non-standardized estimator is.

3.5 CONCLUDING REMARKS

Throughout the experiments in this chapter, the performance of any of SLS, WLS and ML was found to always improve when the sample size increases or the uniquenesses decrese, irrespective of the criterion employed to evaluate it. The first half of this finding is intutively natural and actually agrees with the reports by Pennell [1968] , Boomsm [1982] and [1985] and Anderson and Gerbing [1984] . On the other hand, the last half agrees with Boomsma [1982] and [1985] , Cliff and Pennell [1967] and Pennell [1968] but not necessarily with Anderson and Gerbing [1984] who showed that in some situations the proportion of improper solutions was the largest when the uniquenesses were the smallest of three levels considered there. Another interesting feature, the effect of the number of variables per factor, which was treated by Anderson and Gerbing [1984] was not considered here.

As far the preference among the three methods, SLS, WLS and ML, these results suggest that SLS is most reliable if the problem is difficult in the sense that the sample size is rather small and unique variances are large. A main reason for this finding is that the algorithm of the SLS is the simplest of the three competing methods under investigation and that in general the simplest method would be most efficient for a problem which requires a complicated computation.

A remarkable advantage of the least-squares methods is that it can be applied to any data without assuming any particular probability distribution for the sample, whereas the maximum likelihood method is heavily dependent on the underlying distribution, usually a multivariate normal distribution. Though all samples in this study were generated from normal distributions, the author conjectures that SLS would be more favorable than ML based on the normality assumption and presumable than WLS when the samples are drawn from more general populations. Thus, SLS deserves more attention from statisticians, theoretical or applied, than that paid to at present, though this suggestion is against the current trend in the statistical community which seems to favor ML or WLS among the least-squares family.

Some discussion would be needed on the program-dependency of findings in this study. There is a certain difference between the proportion of IP or NC solutions and the number of iterations on one hand and the error of estimates on the other hand. For a given set of sample correlation matrices, the value of each criterion in the first group maybe heavily dependent on the computer program adopted in the study. However, the present author believes that the result would not change much as far as the comparison of the method is concerned, since the simplicity of SLS would be valid for any algorithm applicable to factor analysis. As for the error of estimates, the estimated value for a particular sample and under a particular method should be the same for any program, provided the computation starts from a reasonably good initial value and the iterative process converges.

CHAPTER 4

NON-ITERATIVE ESTIMATORS IN FACTOR ANALYSIS

4.0 INTRODUCTORY REMARKS

In previous two chapters we have treated the least-squares (LS) and maximum likelihood (ML) methods which are most popular among various methods to estimate the unknown parameters in the factor analysis model. The estimates ($\hat{\Lambda}$, $\hat{\Psi}$) are determined as a solution which minimizes a suitable discrepancy function F (S , Σ) subject to the condition (1.2) , where S and Σ are the sample and population covariance matrices, respectively Since the derivatives of F are non-linear with respect to (Λ , Ψ) , the solution can not be expressed as an explicit function of S . Thus it is usually obtained by means of an iterative procedure as is reviewed in Chapter 2.

Now, among various findings obtained from experimental comparisons in Chapters 2 and 3 the following is the most remarkable: a simpler method performs better than a complicated method does. For this suggests that if we can obtain estimators ($\hat{\Lambda}$, $\hat{\Psi}$) as explicit functions of S , then such estimators may behave better than the LS or ML estimators do for small samples, since we can obtain such estimators without using any iterative procedures. Unfortunately, the population covariance matrix Σ in (1.2) is non-linear with respect to (Λ , Ψ) so that we can not simultaneously express Λ and Ψ as explicit functions of Σ , but we can do only the Ψ as is shown in Ihara and Kano [1986] .

In this chapter we explicate the new estimator of the uniqueness proposed in Ihara and Kano [1986] and apply it to two data sets from Emmett [1949] and Holzinger and Swineford (see Lawley and Maxwell [1971] , p.96) .

4.1 ESTIMATORS OF THE UNIQUENESS

Among various estimators of the uniqueness the estimator SMC (Squared Multiple Correlation) due to Guttman [1956] is most popular. Let us write $S = (s_{ij})$ and $S^{-1} = (s^{ij})$. Then SMC is defined by

$$\hat{\psi}_i = 1 / s^{ii} \qquad (4.1)$$

for i = 1, . . . ,p. Since we have the inequality

$$\Sigma^{-1} < \Psi^{-1} \qquad (4.2)$$

from (1.2) , provided Ψ is positive definite and the number of common factors k is not equal to 0, SMC is a positively biased estimator.

Jöreskog [1967] proposed an initial value of the uniqueness by modifing SMC so as to reduce its bise. Let us denote it by JOR. Then JOR is defined by

$$\hat{\psi}_i = (1 - k / 2p) / s^{ii}$$

for i = 1, . . . ,p.

There remain two more estimators which are widely used as an initiator in analyzing the real data; the highest correlation (HIGH) and ZERO defined by

$$\text{HIGH: } \hat{\psi}_i = s_{ii}\,(1 - \max_{j \neq i} |r_{ir}|)$$

and

$$\text{ZERO: } \hat{\psi}_i = 0 ,$$

respectively. In Okamoto and Ihara [1983b] we carried out an experimental comparison of these four estimators, SMC, JPR, HIGH and ZERO, as the starting point of our iterative computation and then found that SMC was best of all, whereas Okamoto [1986b] reported in his Monte Carlo study that JOR performed better than SMC did.

If we use these estimators as the initiator, they may be superior to our new estimator defined later because our estimator will be time-consuming to compute as compared with them, but our estimator can be shown to be better than them in the sense that it has analytically desirable properties such as consistency, asymptotic normality and scale invariance.

4.2 NEW ESTIMATOR OF THE UNIQUENESS

From the proof for Anderson and Rubin's sufficient condition on the identifiability in Section 1.2 the unique variance matrix Ψ in (1.2) was shown to be expressed as an explicit function of Σ. Thus if the sample covariance matrix S is partitioned in the same fashion as Σ in (1.4) , then we can define an estimator of ψ_i by

$$\hat{\psi}_1 = s_{11} - s_{12} S_{32}^{-1} s_{31}, \qquad (4.3)$$

provided the submatrix S_{32} is non-singular. Note that the expression (4.3) can be rewritten as the reciprocal number of the (1,1) element of the inverse of the submatrix defined by

$$S^{*} = \begin{array}{cc} \left[\begin{array}{cc} s_{11} & s_{12} \\ s_{31} & S_{32} \end{array}\right] & \begin{array}{c} 1 \\ k \end{array} \\ \begin{array}{cc} \;1 & \;\;k \end{array} & \end{array}$$

and thus the expression is similar to that for Guttman's SMC. In the case of SMC we have only to calculate the inverse of S, whereas in the case of the new estimator we need to do the inverse of S^{*} for every index i so that it may be time-consuming to obtain the new estimator.

The estimator $\hat{\psi}_1$ is a continuous function of S and differentiable at $S = \Sigma$, so that by using Theorem (ii) on p.387 of Rao [1973] we can prove the following.

THEOREM 4.1 (Ihara and Kano [1986])

(i) If S is a consistent estimator of Σ, then $\hat{\psi}_1$ is a consistent estimator of ψ_1

and

(ii) if the asymptotic distribution of $n^{1/2}(S - \Sigma)$ is normal, then that of $n^{1/2}(\hat{\psi}_1 - \psi_1)$ is normal.

On the other hand, if S converges to Σ, both SMC and JOR can be shown not to be consistent estimators by noting the inequality (4.2).

If S is transformed into DSD by a diagonal matrix D with positive diagonal elements $d_1, \ldots, d_p$, we will want that the estimator $\hat{\psi}_i$ is transformed into $d_i^2 \hat{\psi}_i$ for each i ($i = 1, \ldots, p$). If an estimator has such a property, we call it scale-invariant estimator, for which we have the following

THEOREM 4.2 (Ihara and Kano [1986])

All the five estimators mentioned above, SMC, JOR, HIGH, ZERO and our estimator, are scale invariant.

Now, different choices of the submatrix S_{32} in (4.3) may yield different values of the estimator $\hat{\psi}_1$. In the next section we will give a procedure for the choice of S_{32} which will be reasonable and simple for computation.

4.3 APPLICATIONS

As described in the last paragraph in the previous section, the value of the estimator depends on the choice of the submatrix S_{32}. Therefore we are required to determine how to choose S_{32}. It will be reasonable to suppose that the stability of S_{32} is an important factor of that of $\hat{\psi}_1$. For each possible choice of S_{32} we calculated the corresponing estimate of ψ_9 and the absolute value (Δ, say) of the determinant of S_{32} in Emmett's data [1949] and Holzinger and Swineford's data (see Lawley and Maxwell [1971] , p.96) with $(p,k) = (9,3)$. For each data the number of estimates for ψ_9 is ${}_8C_3 \cdot {}_5C_3 \,/\, 2 = 280$ and they were grouped into several classes according to the value of Δ. In each class we calculated the mean of the estimates $\hat{\psi}_9$ and the root mean squared error (RMSE) of $\hat{\psi}_9$ to the maximum likelihood estimate (MLE) which is 0.231 for Emmett and 0.421 for Holzinger and Swineford. Table 4.1 shows that the larger the value of Δ becomes, the closer $\hat{\psi}_9$ becomes to the MLE in general in both senses of Mean and RMSE. Thus we suggested the use of S_{32} with the maximum value of Δ in order to hopefully obtain the best estimator. Table 4.2 shows the results when the method was applied to the two data sets mentioned above and it can be seen that our estimate is rather close to the MLE.

In practice, it may be time-consuming to try all possible choices of the matrix S_{32}. Our estimation method would work well by using the maximum value of Δ among randomly chosen S_{32}'s, for instance 10 in number, though the result is not reported.

REFERENCES

[1] Albert, A.A. (1944a) : The matrices of factor analysis. Proc. Nat. Acad. Sci., 30, 90-95.

[2] Albert, A.A. (1944b) : The minimum rank of a correlation matrix. Proc. Nat. Acad. Sci., 30, 144-146.

[3] Anderson, J.C. & Gerbing, D.W. (1984) : The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. Psychometrika, 49, 155-173.

[4] Anderson, T.W. & Rubin, H. (1956) : Statistical inference in factor analysis. Proc. Third Berkley Symp., 5, 111-150.

[5] Boomsma, A. (1982) : The robustness of LISREL against small sample sizes in factor analysis models. Systems under Indirect Observation (K.G. Jöreskog et al., ed.) Part 1, North Holland, Amsterdam, 149-173.

[6] Boomsma, A. (1985) : Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. Psychometrika, 50, 229-242.

[7] Clarke, M.R.B. (1970) : A rapidly convergent method for maximum-likelihood factor analysis. Brit. J. Math. Statist. Psychol., 23, 43-52.

[8] Cliff, N. (1966) : Orthogonal rotation to congruence. Psychometrika, 31, 33-42.

[9] Cliff, N. & Pennell, R. (1967) : The influence of communality, factor strength and loading size on the sampling characteristic of factor loadings. Psychometrika, 32, 309-326.

[10] Derflinger, G. (1969) : Efficient methods for obtaining minres and maximum-likelihood solutions in factor analysis. Metrika, 14, 214-231.

[11] Derflinger, G. (1979) : A general computing algorithm for factor analysis. Biom. J., 21,25-38.

[12] Emmett, W.G. (1944) : Factor analysis by Lawley's method of maximum likelihood. Brit. J. Psychol. Statist., 2, 90-97.

[13] Fletcher, R. & Powell, M.J.D. (1963) : A rapidly convergent method for minimizing a sum of squares of nonlinear functions. Dept. Comp. Sci. Report 71-6 (Yale Univ. New Haven, CT) .

[14] Ganadeskan, R. & Kettenring, J.R. (1984) : A pragmatric review of multivariate methods in applications. Statistics: An Appraisal (H.A. David et al., ed.) Iowa State univ., 309-337.

[15] Guttman, L. (1956) : "Best possible" systematic estimates of communalities. Psychometrika, 21, 273-285.

[16] Harman, H.H. (1960) : Modern Factor Analysis. Univ. of Chicago Press.

[17] Harman, H.H. & Jones, W.H. (1966) : Factor analysis by minimizing residuals (minres) . Psychometrika, 31, 351-368.

[18] Hemmerle, W.J. (1965) : Obtaining maximum-likelihood estimates of factor loadings and communalities using an easily implemented iterative computer procedure. Psychometrika, 30, 291-302.

[19] Ihara, M. (1985) : Asymptotic bias of estimators of the uniqueness in factor analysis. Mathematica Japonica, 30, 885-889.

[20] Ihara, M. (1986) : The structure of improper solutions in maximum likelihood factor analysis. Statist. Prob. Letters, 4, to appear.

[21] Ihara, M. & Okamoto, M. (1985) : Experimental comparison of least-squares and maximum likelihood methods in factor analysis. Statist. Prob. Letters, 3, 287-293.

[22] Ihara, M. & Kano, Y. (1986) : A new estimator of the uniqueness in factor analysis. Psychometrika, to appear.

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

[24] Jöreskog, K.G. (1967) : Some contributions to maximum likelihood factor analysis. Psychometrika, 32, 443-482.

[25] Jöreskog, K.G. (1977) : Factor analysis of least-squares and maximum likelihood methods. Statistical Methods for Digital Computers (K. Enslein et al., ed.) , 3. Wiley, New York, 125-153.

[26] Jöreskog, K.G & Goldberger, A.S. (1972) : Factor analysis by generalized least squares. Psychometrika, 37, 243-260.

[27] Kano, Y. (1986) : Conditions on consistency of estimators in covariance structure model. J. Japan Statist. Soc., 16, 75-80.

[28] Krane, W.R. & McDonald, R.P. (1978) : Scale invariance and the factor analysis of correlation matrices. Brit. J. Math. Psychol., 31, 218-228.

[29] Lawlew, D.N. (1940) : The estimation of factor loadings by the method of maximum likelihood. Proc. Roy. Soc. Edinb., A 60, 64-82.

[30] Lawley, D.N. & Maxwell, A.E. (1971) : Factor Analysis as a Statistical Method. Second ed., Butterworth, London.

[31] Lee, S.Y. (1980) : Estimation of covariance structure models with parameter subject to functional restraints. Psychometrika, 45, 309-324.

[32] Lee, S.Y. & Jennrich, R.I. (1979) : A study of algorithms for covariance structure analysis with specific comparisons using factor analysis. PAsychometrika, 44, 99-113.

[33] Lee, S.Y. & Poon, W.Y. (1985) : Further developments on constrained estimation in analysis of covariance structures. The Statisticain, 34, 305-316.

[34] Marquardt, D.W. (1963) : An algorithm for least-squares estimation of non-linear parameters, SIAM J., 11, 431-441.

[35] Okamoto, M. (1986a) : Inshi-bunseki no Kiso (in Japanese) . Nihon Kagaku-gijutsu Renmei Shuppan, Tokyo, Japan.

[36] Okamoto, M. (1986b) : Early-step estimators in least-squares factor analysis. to be published.

[37] Okamoto, M. & Ihara, M. (1983a) : A simple Marquadt algorithm for the nonlinear least-squares problem. Statist. Prob. Letters, 1, 301-305.

[38] Okamoto, M. & Ihara, M. (1983b) : A new algorithm for least-squares solution in factor analysis. Psychometrika, 48, 597-605.

[39] Okamoto, M. & Ihara, M. (1984) : Partial Gauss-Newton algorithm for least-squares and maximum likelihood methods in factor analysis. J. Japan Statist. Soc., 14, 137-144.

[40] Pennell, R. (1968) : The influence of communality and N on the sampling distributions of factor loadings. Psychometrika, 33, 423-439.

[41] Rao, C.R. (1955) : Estimation and tests of significance in factor analysis. Psychometrika, 20, 93-111.

[42] Smith, W.B. & Hocking, R.R. (1972) : Wishart variate generator. Appl. Statist., 21, 341-345.

[43] Spearman, C. (1904) : "General intelligence", objectively determined and measured. Amer. J. Psychol., 15, 201-293.

[44] Thurstone, L.L. (1935) : The vectors of mind. Univ. of Chicago Press.

[45] Tumura, Y. & Fukutomi, K. (1968) : On the identification in factor analysis. Rep. Statist. Appl. Res. JUSE, 15, 98-103.

[46] Tumura, Y. & Sato, M. (1980) : On the identification in factor analysis. TRU Mathematics, 16, 121-131.

[47] Tumura, Y. & Sato, M. (1985) : Some notes on the identification in factor analysis. (personal communication) .

[48] van Driel, O.P (1978) : On variance causes of improper solutions in maximum likelihood factor analysis. Psychometrika, 43, 225-243.

[49] Williams, J.S. (1981) : A note on the uniqueness of minimum rank solutions in factor analysis. Psychometrika, 46, 109-110.

[50] Wold, H. (1982) : Models for knowledge. The Making of Statisticians (J. Gani, ed.) , Springer, Berlin, 190-212.

ACKNOWLEDGEMENT

I would like to express my sincere thanks to Professor M. Okamoto of Osaka University for guiding my graduate study encouragingly. I am also grateful to Dr. Y. Kano who has been my best worker with the same major, factor analysis. Without efforts made by these two persons, I could not write this thesis.

I wish to express my gratitude to Professor N. Inagaki of Osaka University and Drs. S. Shirahata, T. Isogai, T. Toyooka and S. Eguchi for their constructive suggestions and comments from many points of view.

I am also thankful for the benifits I have received from many other teachers and colleagues; members of Statistics Group of Osaka University and also members of Kansai Multivariate Analysis Research Group which are organized by Professor Y. Nayatani of Osaka Electro-Communication University and Dr. M. Goto of Sionogi Kaiseki Center.

AUTHOR'S ADDRESS
Faculty of Engineering
Osaka Electro-Communication University
18-8 Hatsu-cho, Neyagawa, Osaka 572
JAPAN

Table 2.1

Computational algorithm for factor analysis

Authors	Method	Variable	Constraint	Algorithm
Before 1960	ML	Λ & Ψ	———	PFA·
Harman-Jones (1966)	LS	Λ	———	GS
Jöreskog (1967)	ML	Ψ	Ψ^{-1}	DFP
Jennrich-Robinson (1969)	ML	Ψ	S^{-1}	NR
Derflinger (1969)	ML,LS	Ψ	Ψ^{-1}, I	NR
Clarke (1970)	ML	Ψ	Ψ^{-1}	NR
Jöreskog-Goldberger (1972)	LS	Ψ	Ψ^{-1}	NR
Jöreskog (1977)	ML,LS	Ψ	S^{-1}	NR
Lee-Jennrich (1979)	ML,LS	Λ & Ψ	———	GN
Okamoto-Ihara (1983b)	LS	Λ	———	Marquardt
Okamoto-Ihara (1984)	ML,LS	Ψ	S^{-1}, D_s^{-2}	GN

PFA: Principal Factor Analysis Method

GS: Gauss-Seidel Method

DFP: Davidon-Fletcher-Powell Method

NR: Newton-Raphson Method

GN: Gauss-Newton Method

Marquardt: Marquardt Method

Table 2.2 Convergence of the partial Gauss-Newton ML algorithm applied to Rao's data using Rao's solution for starting values.

Iter.	F	RMSg	RMS$\Delta\psi$	Δ_1	Δ_2	Δ_3	Δ_4	Δ_5	Δ_6	Δ_7	Δ_8	Δ_9
0	.09963	——	——	.4400	.5700	.8800	.9000	.7300	.4300	.5500	.7500	.6300
1	.06974	.1644	.0413	.3437	.5834	.8796	.9060	.7250	.4297	.5425	.7515	.5487
2	.06790	.0175	.0396	.0000	.5920	.8797	.9058	.7286	.4302	.5441	.7495	.5476
3	.06774	.0141	.0027	.0000	.5872	.8792	.9030	.7296	.4318	.5445	.7481	.5468
4	.06774	.0011	.0002	.0000	.5871	.8793	.9029	.7298	.4321	.5445	.7479	.5466
5	.06774	.0001	.0000	.0000	.5871	.8793	.9029	.7298	.4321	.5445	.7479	.5466

Table 2.3 Convergence of the partial Gauss-Newton ML algorithm applied to Harman's data using Harman's solution for starting values with Δ_8 replaced by .700.

Iter.	F	RMSg	RMS$\Delta\psi$	Δ_1	Δ_2	Δ_3	Δ_4	Δ_5	Δ_6	Δ_7	Δ_8
0	.13181	——	——	.3990	.3050	.4110	.4380	.2920	.6040	.6470	.7000
1	.08263	.3947	.0286	.3667	.1760	.4384	.3937	.3037	.5970	.6413	.6989
2	.07641	.1322	.0117	.3649	.0000	.4470	.3906	.3063	.5971	.6410	.7046
3	.07572	.0564	.0038	.3585	.0000	.4403	.3938	.3001	.5994	.6418	.7006
4	.07571	.0035	.0006	.3571	.0000	.4404	.3950	.3008	.5993	.6408	.7008
5	.07571	.0011	.0002	.3571	.0000	.4405	.3954	.3007	.5994	.6408	.7007
6	.07571	.0004	.0001	.3570	.0000	.4405	.3955	.3007	.5994	.6407	.7007

Table 3.1

Proportion of proper, improper and nonconvergent solutions (Model 1)

	Small Ψ			Large Ψ		
n	100	300	1000	100	300	1000
SLS	78.5	93.0	100.0	73.5	90.5	99.0
	21.0	7.0	0.0	24.5	9.5	1.0
	0.5	0.0	0.0	2.0	0.0	0.0
WLS	61.0	93.0	100.0	49.5	85.0	99.5
	34.5	7.0	0.0	44.0	12.5	0.5
	4.5	0.0	0.0	6.5	2.5	0.0
ML	63.0	91.0	100.0	56.5	87.0	99.0
	34.0	8.5	0.0	40.5	11.5	1.0
	3.0	0.5	0.0	3.0	1.5	0.0

The upper value shows the proportion of proper solutions (P), the middle value for improper solutions (IP) and the lower value for non-convergent solutions (NC).

Table 3.2

The number of iterations for proper solutions (Model 1)

		Small Ψ			Large Ψ		
	n	100	300	1000	100	300	1000
SLS	Median	5	4	4	5	5	4
	Mean	5.91	4.50	4.01	6.33	5.12	4.23
	S.D.	3.40	1.39	0.49	2.97	1.80	0.59
WLS	Median	8	5	4	8	7	5
	Mean	9.56	6.95	4.48	9.72	8.27	4.89
	S.D.	5.12	3.04	0.63	4.76	3.95	0.92
ML	Median	9	6	5	10	7	5
	Mean	9.87	7.23	4.94	10.88	7.98	5.27
	S.D.	4.45	2.83	0.84	5.17	3.58	1.02

Table 3.3

Values of RMSEU(multiplied by 10^4) (Model 1)

		Small Ψ			Large Ψ		
	n	100	300	1000	100	300	1000
SLS	(a)	960	552	322	1271	729	422
	(b)	914	541	322	1194	700	422
WLS	(a)	985	569	315	1290	741	428
	(b)	945	550	315	1228	726	421
ML	(a)	966	558	315	1312	755	420
	(b)	948	549	315	1256	734	420

Table 3.4

Values of RMSEL(multiplied by 10^4)(Model 1)

		Small Ψ			Large Ψ		
	n	100	300	1000	100	300	1000
SLS	(a)	784	441	249	1012	564	311
	(b)	764	436	249	969	547	311
WLS	(a)	811	451	245	1011	571	311
	(b)	798	444	245	993	565	311
ML	(a)	813	447	245	1040	579	311
	(b)	799	444	245	1002	568	311

Table 3.5

Bias for uniqueness (x100) for proper solutions (Model 1)

(i)SmallΨ

Variate		ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6	ψ_7	ψ_8	ψ_9
True value		0.41	0.43	0.62	0.26	0.38	0.19	0.34	0.50	0.20
100	SLS	-3		-5	-1	-2	1	-2		
	WLS	-6	-3	-8	-3	-5		-4	-2	-1
	ML	-3		-5	-1	-2		-3	2	
300	SLS								-1	
	WLS	-1	-1	-2	-1	-1		-1	-1	
	ML									

(ii)LargeΨ

Variate		ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6	ψ_7	ψ_8	ψ_9
True value		0.54	0.56	0.71	0.41	0.51	0.34	0.47	0.57	0.35
100	SLS	-5	-1	-8	-2	-4		-3	2	-1
	WLS	-7	-3	-12	-5	-7	-2	-6	-2	-3
	ML	-5		-8	-3	-4		-4	2	-1
300	SLS	-1	-1	-1		-1		-1		
	WLS	-2	-2	-2	-2	-2	-1	-2	-1	-1
	ML	-1	-1			-1	-1	-1		

Table 3.6

Proportion of proper, improper and non-convergent solutions(Model 2)

	a = 0.9			a = 0.7		
n	100	300	1000	100	300	1000
	89.0	100.0	100.0	77.5	100.0	100.0
SLS	11.0	0.0	0.0	19.0	0.0	0.0
	0.0	0.0	0.0	3.5	0.0	0.0
	83.0	100.0	100.0	55.0	99.0	100.0
WLS	16.5	0.0	0.0	35.0	1.0	0.0
	0.5	0.0	0.0	10.0	0.0	0.0
	84.5	99.5	100.0	59.0	99.0	100.0
ML	15.0	0.5	0.0	36.0	1.0	0.0
	0.5	0.0	0.0	5.0	0.0	0.0

The upper value shows the proportion of proper solutions (P), the middle value for improper solutions (IP) and the lower value for non-convergent solutions (NC).

Table 3.7

The number of iterations for proper sulutions (Model 2)

		a = 0.9			a = 0.7		
	n	100	300	1000	100	300	1000
	Median	4.00	4.00	3.00	6.00	5.00	4.00
SLS	Mean	4.10	3.33	3.01	7.54	4.99	3.77
	S.D.	1.12	0.49	0.07	3.93	1.88	0.46
	Median	6.00	5.00	4.00	11.00	6.00	4.00
WLS	Mean	6.42	4.74	4.11	11.84	7.38	4.33
	S.D.	2.48	0.86	0.31	5.27	3.97	0.52
	Median	7.00	5.00	4.00	11.00	7.00	5.00
ML	Mean	7.46	5.24	4.35	12.84	7.84	4.74
	S.D.	2.70	0.82	0.48	5.21	3.12	0.63

Table 3.8

Values of RMSEU (multiplied 10^4) (Model 2)

		a = 0.9			a = 0.7		
n		100	300	1000	100	300	1000
SLS	(a)	939	531	286	1423	799	409
	(b)	942	531	286	1389	784	409
WLS	(a)	1106	549	277	1690	907	418
	(b)	1107	548	277	1649	878	418
ML	(a)	902	505	271	1513	836	409
	(b)	898	504	271	1508	830	409

Table 3.9

Values of RMSEL (multiplied by 10^4) (Model 2)

		a = 0.9			a = 0.7		
n		100	300	1000	100	300	1000
SLS	(a)	797	454	247	1183	670	356
	(b)	795	454	247	1155	660	356
WLS	(a)	781	440	238	1210	694	356
	(b)	780	440	238	1195	680	356
ML	(a)	773	438	238	1206	678	355
	(b)	773	438	238	1191	674	355

Table 4.1

Relatinship between Δ and the closeness to MLE

Interval	Emmett			Holzinger & Swineford		
	Frequency	Mean	RMSE	Frequency	Mean	RMSE
$.000 \leqq \Delta < .010$	159	.135	.578	124	.604	2.637
$.010 \leqq \Delta < .020$	54	.237	.054	74	.444	.306
$.020 \leqq \Delta < .030$	38	.227	.032	21	.470	.176
$.030 \leqq \Delta < .040$	22	.224	.030	11	.515	.171
$.040 \leqq \Delta < .050$	7	.240	.015	10	.527	.156
$.050 \leqq \Delta < .100$	0	——	——	9	.432	.089
$.100 \leqq \Delta < .200$	0	——	——	31	.394	.042
SMC	.348			.504		
MLE	.231			.400		

Talbe 4.2

New Estimates and MLE's for the uniqueness on Emmett's and Holzinger & Swineford's data

Variable		1	2	3	4	5	6	7	8	9
Emmett	NEW	.438	.481	.664	.209	.375	.225	.408	.654	.266
	MLE	.451	.427	.617	.212	.381	.177	.400	.462	.231
Holzinger & Swineford	NEW	.499	.624	.470	.301	.376	.318	.343	.313	.446
	MLE	.491	.622	.443	.289	.370	.324	.325	.268	.402