The Annals of Statistics
1984, Vol. 12, No. 1, 1–45

THE 1982 WALD MEMORIAL LECTURES

ESTIMATING LINEAR STATISTICAL RELATIONSHIPS[1]

BY T. W. ANDERSON

Stanford University

This paper on estimating linear statistical relationships includes three lectures on linear functional and structural relationships, factor analysis, and simultaneous equations models. The emphasis is on relating the several models by a general approach and on the similarity of maximum likelihood estimators (under normality) in the different models. In the first two lectures the observable vector is decomposed into a "systematic part" and a random error; the systematic part satisfies the linear relationships. Estimators are derived for several cases and some of their properties given. Estimation of the coefficients of a single equation in a simultaneous equations model is shown to be equivalent to estimation of linear functional relationships.

Received December 1982; revised October 1983.

[1] This work was supported by National Science Foundation Grants SES79-13976 and MCS82-19748 at the Institute for Mathematical Studies in the Social Sciences, Stanford University. The exposition benefited from discussion at a workshop held at the University of Dundee in August 1983 funded by the UK Science and Engineering Research Council. The paper was completed while the author was Wesley C. Mitchell Research Professor of Economics at Columbia University.

AMS 1980 *subject classifications*. Primary, 62H12, secondary, 62H25.

Key words and phrases. Linear functional relations, errors in variables, linear structural relationships, factor analysis, canonical variables, discriminant analysis, structural equations, simultaneous equation models, multivariate analysis of variance (MANOVA), principal components.

1. Introduction. There are many ways to consider the relations between statistical variables. Several of them are based on assumed probabilistic models and have arisen in different fields of application, such as psychology and economics. The terminology, notation, and emphasis vary from area to area. (The list of key words shows some of the diversity.) The purpose of this set of lectures is to bring together various models and appropriate statistical techniques, show the connections among them, and unify the treatment in terms of a general approach.

The titles of the three lectures are I. Linear functional and structural relationships (Section 2–7), II. Factor analysis (Sections 8–16), and III. Simultaneous equations models (Sections 17–26). The first lecture includes several "errors-in-variables" models. The second lecture treats a version of the linear functional and structural relationship that has been used by psychologists to analyze test scores. The third lecture shows that a certain approach to simultaneous equations models of econometrics leads to the linear functional relationship model and hence that the properties of the methods used in one field apply to the other.

This paper is concerned with a vector variable of $p (p \geq 2)$ components. These components may be physical measurements, psychological test scores, or observations on macro-economic quantities. (We shall use lower case light face letters to denote scalar variables, lower case bold face letters to denote vector variables, and capital bold face letters to denote matrix variables; we shall not distinguish in notation among random, running and observed variables.) The observable vector $\mathbf{x}_\alpha$ is considered as decomposed according to

$$\mathbf{x}_\alpha = \mathbf{z}_\alpha + \mathbf{u}_\alpha, \quad \alpha = 1, \cdots, n, \tag{1.1}$$

where the unobservable $\mathbf{z}_\alpha$ can be considered as the "systematic" or "true" part and the unobservable $\mathbf{u}_\alpha$ can be considered as the random "error". It is the systematic part that is of real interest; the investigator would like to know it free of error. In mathematical terms what distinguishes the systematic part is that it varies in a lower-dimensional linear space, of dimension m, say $(m < p)$. The $\mathbf{u}_\alpha$ are random with the properties

$$\mathcal{E}\mathbf{u}_\alpha = \mathbf{0}, \quad \mathcal{E}\mathbf{u}_\alpha\mathbf{u}_\alpha' = \mathbf{\Psi}. \tag{1.2}$$

Furthermore, $\mathbf{u}_1, \cdots, \mathbf{u}_n$ are mutually independent and are independent of $\mathbf{z}_1, \cdots, \mathbf{z}_n$.

Having defined the model, we specify the cases that are examined in this paper. First, there are alternative ways of treating the systematic parts.

I. The unobservable vectors $\mathbf{z}_1, \cdots, \mathbf{z}_n$ can be treated as parameters, that is, fixed or nonstochastic. They are the expected values of $\mathbf{x}_1, \cdots, \mathbf{x}_n$.

II. Alternatively $\mathbf{z}_1, \cdots, \mathbf{z}_n$ can be treated as random or stochastic.

Kendall and Stuart (1979), Chapter 29, have termed I as "functional" and II as "structural." This distinction corresponds to the distinction made in the analysis of variance between Model I, fixed effects, and Model II, components of variance.

In the functional case $\mathbf{z}_1, \cdots, \mathbf{z}_n$ are incidental parameters pertaining to the observed units; inference concerns those units. In the structural case $\mathbf{z}_1, \cdots, \mathbf{z}_n$ constitutes a sample from a population; inference pertains to the population. The distinction affects the mathematical treatment. For example, asymptotic theory in the second case is straightforward because one fixed distribution is involved, but in the first case asymptotic theory depends on the properties of the incidental parameters as $n \to \infty$.

In the simplest case of $p = 2$ and $m = 1$ (treated in detail in Section 2) the systematic parts lie on a line, that is, satisfy one linear equation, say $b_1 z_{1\alpha} + b_2 z_{2\alpha} = b_0$. An observable pair of variables $(x_{1_\alpha}, x_{2_\alpha})$ is a point in the plane that differs from a point $(z_{1\alpha}, z_{2\alpha})$ on the line by a pair of random variables $(u_{1\alpha}, u_{2\alpha})$. The line on which the systematic points lie can optionally be written in parametric form: $z_{1\alpha} = \lambda_1 f_\alpha + \mu_1$, $z_{2\alpha} = \lambda_2 f_\alpha + \mu_2$. We shall consider statistical problems of estimating b_1, b_2, and b_0 or λ_1, λ_2, μ_1, and μ_2.

In general, the m-dimensional space in which the $\mathbf{z}_\alpha$'s vary can be defined in terms of $q(=p - m)$ linear equations

$$\mathbf{B}\mathbf{z}_\alpha = \mathbf{b}, \tag{1.3}$$

where $\mathbf{B}$ is a $q \times p$ matrix of parameters of rank $q < p$ and $\mathbf{b}$ is a column vector of parameters of q components. Each row of (1.3) defines a hyperplane; the m-space is the intersection of these q hyperplanes. An alternative description is the "parametric" form

$$\mathbf{z}_\alpha = \Lambda \mathbf{f}_\alpha + \mu, \tag{1.4}$$

where Λ is a $p \times m$ matrix of rank $m < p$, $\mathbf{f}_\alpha$ is a column vector of m components, and μ is a column vector of p components. The p-component vector $\mathbf{z}_\alpha$ varies over the m-space as the m components of $\mathbf{f}_\alpha$(coordinates in the m-space) vary. We center $\mathbf{f}_\alpha$ by requiring $\sum_{\alpha=1}^{n} \mathbf{f}_\alpha = \mathbf{0}$ in the functional case and $\mathscr{E}\mathbf{f}_\alpha = \mathbf{0}$ in the structural case. Expression (1.4) indicates that $\mathbf{z}_\alpha$ is a linear combination of columns of Λ plus the location vector μ.

What is the connection between these two representations? Substitution of (1.4) into (1.3) yields

$$\mathbf{B}\mathbf{z}_\alpha = \mathbf{B}\Lambda\mathbf{f}_\alpha + \mathbf{B}\mu = \mathbf{b}. \tag{1.5}$$

For (1.5) to be an identity in $\mathbf{f}_\alpha$ we want

$$\mathbf{B}\Lambda = \mathbf{0}, \quad \mathbf{B}\mu = \mathbf{b}. \tag{1.6}$$

In each representation there is an indeterminancy. In the first case the vector equation (1.3) can be multiplied on the left by an arbitrary nonsingular matrix

$\mathbf{A}$, which corresponds to the transformation

$$\mathbf{B} \to \mathbf{AB}, \quad \mathbf{b} \to \mathbf{Ab}. \tag{1.7}$$

In the second case the matrix Λ in (1.4) can be multiplied on the right by an arbitrary nonsingular matrix $\mathbf{C}$; if $\mathbf{f}_\alpha$ is multiplied on the left by $\mathbf{C}^{-1}$

$$\Lambda \to \Lambda\mathbf{C}, \quad \mathbf{f}_\alpha \to \mathbf{C}^{-1}\mathbf{f}_\alpha, \tag{1.8}$$

the product $\Lambda\mathbf{f}_\alpha$ is unchanged. In order to determine uniquely $(\mathbf{B}, \mathbf{b})$ or $(\Lambda, \mathbf{f}_\alpha)$, further restrictions are needed. For example, suppose we partition $\mathbf{B}$ as $(\mathbf{B}_1\ \mathbf{B}_2)$, where $\mathbf{B}_1$ is square and nonsingular; since $\mathbf{B}$ is of rank q we can number the columns of $\mathbf{B}$ (that is, number the coordinates of $\mathbf{z}_\alpha$) so that the first square submatrix is of rank q, which is its order. If we multiply $\mathbf{B}$ on the left by $\mathbf{B}_1^{-1}$, we obtain

$$\mathbf{B}^* = (\mathbf{I}_q\ \mathbf{B}_1^{-1}\mathbf{B}_2) = (\mathbf{I}_q\ \mathbf{B}_2^*). \tag{1.9}$$

The matrix of linear restrictions is uniquely determined if the first square submatrix is required to be the identity.

In the case of the "parametric" form in (1.4) Λ can be partitioned into $p - m = q$ rows and m rows, $\Lambda = (\Lambda_1'\ \Lambda_2')'$, so that the square matrix Λ_2 is nonsingular (renumbering rows if necessary). If we multiply Λ on the right by Λ_2^{-1}, we obtain

$$\Lambda^* = \begin{pmatrix} \Lambda_1\Lambda_2^{-1} \\ \mathbf{I}_m \end{pmatrix} = \begin{pmatrix} \Lambda_1^* \\ \mathbf{I}_m \end{pmatrix}. \tag{1.10}$$

The matrix Λ is uniquely determined if the last square submatrix is required to be $\mathbf{I}_m$. When $\mathbf{B}$ and Λ meet these requirements, (1.6) is

$$\mathbf{0} = \mathbf{B}^*\Lambda^* = (\mathbf{I}_q\ \mathbf{B}_2^*)\begin{pmatrix} \Lambda_1^* \\ \mathbf{I}_m \end{pmatrix} = \Lambda_1^* + \mathbf{B}_2^*; \tag{1.11}$$

that is,

$$\mathbf{B}_2^* = -\Lambda_1^*. \tag{1.12}$$

Let $\mathbf{z}_\alpha' = (\mathbf{z}_\alpha^{(1)\prime}\ \mathbf{z}_\alpha^{(2)\prime})$, where $\mathbf{z}_\alpha^{(1)}$ and $\mathbf{z}_\alpha^{(2)}$ consist of q and m components, respectively, and let $\mathbf{b}^* = \mathbf{B}_1^{-1}\mathbf{b}$. Then (1.3) can be written

$$\mathbf{z}_\alpha^{(1)} = \mathbf{b}^* - \mathbf{B}_2^*\mathbf{z}_\alpha^{(2)}. \tag{1.13}$$

When this is inserted in (1.1), the result is

$$\mathbf{x}_\alpha = \begin{pmatrix} -\mathbf{B}_2^* \\ \mathbf{I}_m \end{pmatrix}\mathbf{z}_\alpha^{(2)} + \begin{pmatrix} \mathbf{b}^* \\ \mathbf{0} \end{pmatrix} + \mathbf{u}_\alpha. \tag{1.14}$$

The "parametric" form based on (1.4) is translated to

$$\mathbf{x}_\alpha = \begin{pmatrix} \Lambda_1^* \\ \mathbf{I}_m \end{pmatrix}\mathbf{f}_\alpha^* + \begin{pmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{pmatrix} + \mathbf{u}_\alpha, \tag{1.15}$$

where $\mu' = (\mu^{(1)'}\mu^{(2)'})$. Substitution of (1.12) into (1.14) yields

$$\mathbf{x}_\alpha = \begin{pmatrix} \Lambda_1^* \\ \mathbf{I}_m \end{pmatrix}(\mathbf{z}_\alpha^{(2)} - \bar{\mathbf{z}}^{(2)}) + \begin{pmatrix} \mathbf{b}^* + \Lambda_1^*\bar{\mathbf{z}}^{(2)} \\ \bar{\mathbf{z}}^{(2)} \end{pmatrix} + \mathbf{u}_\alpha, \tag{1.16}$$

where $\bar{\mathbf{z}}^{(2)} = (1/n)\sum_{\alpha=1}^{n} \mathbf{z}_\alpha^{(2)}$ when the $\mathbf{z}_\alpha^{(2)}$'s are nonstochastic and

$$\mathbf{x}_\alpha = \begin{pmatrix} \Lambda_1^* \\ \mathbf{I}_m \end{pmatrix}(\mathbf{z}_\alpha^{(2)} - \mathscr{E}\mathbf{z}^{(2)}) + \begin{pmatrix} \mathbf{b}^* + \Lambda_1^* \mathscr{E}\mathbf{z}^{(2)} \\ \mathscr{E}\mathbf{z}^{(2)} \end{pmatrix} + \mathbf{u}_\alpha \tag{1.17}$$

when the $\mathbf{z}_\alpha^{(2)}$'s are random. Thus the form using the linear equations (1.3) is equivalent to the form using (1.4). The model (1.1) and (1.3) is often called the "errors in variables" model.

The indeterminacy indicated by (1.7) and (1.8) can be removed by imposition of other restrictions on $\mathbf{B}$ or Λ or $\mathbf{f}_\alpha$; we shall consider alternatives later. The model, including such restrictions and properties of the errors, describes the joint distribution of $\mathbf{x}_1, \cdots, \mathbf{x}_n$. If this distribution uniquely determines the value of a parameter, that parameter is said to be *identified*; if all the parameters are identified, the model is identified.

Error structure. We shall treat three cases of the error covariance matrix Ψ defined in (1.2).

1. $\Psi = \sigma^2\mathbf{I}_p$, where σ^2 is not specified. Here the component errors are uncorrelated and have equal variances; that implies that all measurements are in the same units (for example, centimeters). (The specification $\Psi = \sigma^2\mathbf{I}_p$ states that the components of $\mathbf{u}_\alpha$ are uncorrelated and that the variances or standard deviations of the components of $\mathbf{u}_\alpha$ are equal. The standard deviation of a continuous variable is expressed in the units of measurement of that variable; the equality of standard deviations must be expressed in the relevant units. For example, the variability of measurements of head lengths and breadth is the same if the two standard deviation are three centimeters; they are not the same if the standard deviation of one is three centimeters and the other is three inches (3 cm $\neq$ 3 inches).) This form of the covariance matrix may be appropriate when measurements are made independently with the same instrument (such as a micrometer). If Ψ is specified to be $\sigma^2\Psi_0$, where Ψ_0 is known, the covariance matrix can be transformed into $\sigma^2\mathbf{I}_p$. Let $\Psi_0 = \mathbf{T}\mathbf{T}'$ and $\mathbf{u}^* = \mathbf{T}^{-1}\mathbf{u}$; then $\mathscr{E}\mathbf{u}^*\mathbf{u}^{*\prime} = \sigma^2\mathbf{I}_p$. We shall see that the estimators of the parameters in the functional case are the same for $\Psi = \mathbf{I}_p$ as for $\Psi = \sigma^2\mathbf{I}_p$ (except for the estimation of σ^2).

2. Ψ diagonal. Here the component errors are uncorrelated, but not necessarily with the same variance; the measurements do not need to be in the same units.

3. Ψ unrestricted. In this case replicated observations are needed in order to estimate Ψ. Otherwise the parameters are not identified.

In the standard multiple regression model the m variables constituting $\mathbf{x}_\alpha^{(2)}$, called independent, are measured exactly; that is, the systematic part $\mathbf{z}_\alpha^{(2)}$, is observed directly. (Those components of $\mathbf{u}_\alpha$ are 0). The other q variables constituting $\mathbf{x}_\alpha^{(1)}$, called dependent, differ from their systematic parts, $\mathbf{z}_\alpha^{(1)}$ which are

modeled as a linear transformation of $\mathbf{z}_\alpha^{(2)}$. This linear transformation is usually considered to describe how the independent variables generate the dependent variables. The random terms added to these linear combinations (the first q components of $\mathbf{u}_\alpha$) can be interpreted as representing disturbances in the relations (due to ignored independent variables), errors of measurement in the dependent variables, or both. The assumption that the random terms are independent of the independent variables leads the latter to be considered as explanatory. In the multiple linear functional relationships the random terms are independent of all systematic parts and may be correlated with any or all observed variables. The model itself does not distinguish any subset of variables as "dependent" or as "independent" (or "explanatory"); the variables are treated symmetrically. In this sense setting some coefficients equal to 1 and 0 effects identification but does not necessarily affect interpretation.

To obtain maximum likelihood estimators we shall assume that the errors are normally distributed and also that any random systematic parts are normally distributed. We concentrate then on the first and second order moments. The estimators obtained by maximum likelihood and by generalized least squares can also be used when the random variables are not normally distributed. (There are procedures based on the systematic parts not being normal. See Chapter 29 of Kendall and Stuart (1979).)

Part I. Linear functional and structural relationships.

2. A linear functional relationship: One relationship between two variables. In this case of $p = 2$ and $m = q = 1$ and normal errors, the model (1.1) yields the 2-vector $(x_{1\alpha}, x_{2\alpha})'$ distributed as

$$N\left[\begin{pmatrix} z_{1\alpha} \\ z_{2\alpha} \end{pmatrix}, \sigma^2\mathbf{I}_2\right], \quad \alpha = 1, \cdots, n, \tag{2.1}$$

where the systematic parts satisfy the linear equation

$$z_{2\alpha} = \gamma + \beta z_{1\alpha}, \quad \alpha = 1, \cdots, n, \tag{2.2}$$

where γ and β are to be determined. The model is illustrated in Figure 1. The points on the line are the means $(z_{1\alpha}, z_{2\alpha})$, $\alpha = 1, \cdots, n$, and the circles indicate the circular normal distributions centered at these points. The parameters can be taken as $z_{11}, \cdots, z_{1n}, \gamma, \beta$, and σ^2. An alternative for the slope β is the angle θ that the line makes with the μ- or x-axis, as defined by

$$\tan\theta = \beta. \tag{2.3}$$

The likelihood function based on the observations $(x_{1\alpha}, x_{2\alpha})$, $\alpha = 1, \cdots, n$, is

$$\frac{1}{(2\pi)^n\sigma^{2n}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{\alpha=1}^n [(x_{1\alpha} - z_{1\alpha})^2 + (x_{2\alpha} - \gamma - \beta z_{1\alpha})^2]\right\}. \tag{2.4}$$

Maximization of the likelihood with respect to $z_{11}, \cdots, z_{1n}, \gamma$, and β is equivalent to minimizing the sum of squares in the exponent with respect to these quantities.

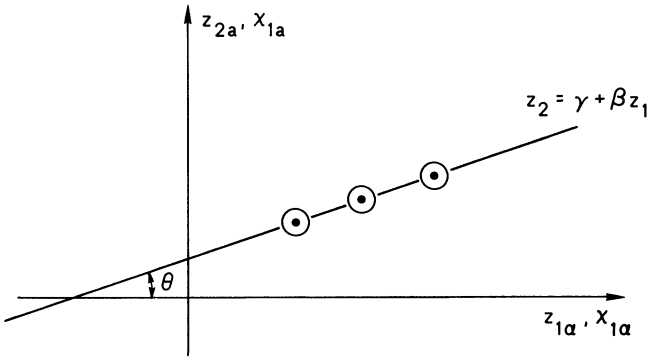

FIG. 1. *Linear functional model for* $p = 2$.

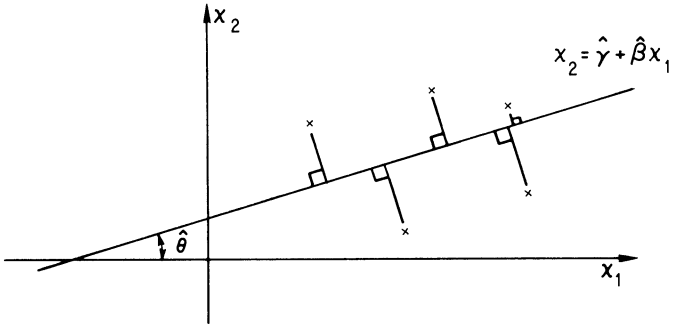

FIG. 2. *Estimation of the linear functional relationship.*

Hence, the maximum likelihood estimates of these parameters are also least squares estimates. (The maximum likelihood estimates for σ^2 known are the same as those for σ^2 unknown.) These estimates minimize the sum of squared distances of the observed points from the fitted line; they minimize

$$\frac{\sum_{\alpha=1}^{n} (x_{2\alpha} - \gamma - \beta x_{1\alpha})^2}{1 + \beta^2}. \tag{2.5}$$

In this sense the estimates are generalized least squares estimates.

Figure 2 illustrates the estimation. The ×'s denote the observed points. The short line segments are orthogonal to the fitted line; it is the sum of squares of their lengths that is minimized. The estimates of the expected values $(\hat{z}_{1\alpha}, \hat{z}_{2\alpha})$ are the projections of the observed points on the line. This estimation procedure is sometimes known as "orthogonal regression." Another way of describing the method is that the direction of the line is the direction of the first principal component; that is, the direction maximizes the scatter; the projections on the line have the maximum variance along the line.

This problem has a long history. In fact, we are just past the hundredth anniversary of the solution. Some of the contributors are Adcock (1878), Kummel (1879), Pearson (1901), Gini (1921), Van Uven (1930), Dent (1935) and Koop-

mans (1937). The method of orthogonal regression was discovered and rediscovered many times, often independently.

In multivariate analysis it is enlightening to interpret the problem geometrically in a space of dimensionality equal to the number of observations, here n. The sets of means $(z_{11}, \cdots, z_{1n})$ and $(z_{21}, \cdots, z_{2n})$ are represented by two points in n-space

$$\mathbf{z}_1^* = (z_{11} - \bar{z}_1, \cdots, z_{1n} - \bar{z}_1), \quad \mathbf{z}_2^* = (z_{21} - \bar{z}_2, \cdots, z_{2n} - \bar{z}_2). \tag{2.6}$$

The points are points in $(n-1)$-dimensional planes $[\sum_{\alpha=1}^n (z_{1\alpha} - \bar{z}_1) = 0 = \sum_{\alpha=1}^n (z_{2\alpha} - \bar{z}_2)]$ perpendicular to the equiangular line. As indicated in Figure 3 and implied by (2.2), the two points are collinear

$$\mathbf{z}_2^* = \beta \mathbf{z}_1^*. \tag{2.7}$$

The random vectors

$$\mathbf{x}_1^* = (x_{11} - \bar{x}_1, \cdots, x_{1n} - \bar{x}_1), \quad \mathbf{x}_2^* = (x_{21} - \bar{x}_1, \cdots, x_{2n} - \bar{x}_2) \tag{2.8}$$

have independent spherical normal distributions with centers at $\mathbf{z}_1^*$ and $\mathbf{z}_2^*$, respectively. (The points $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$ are in the $(n-1)$-dimensional space orthogonal to the equiangular line and can be expressed in terms of $n-1$ coordinates.)

The maximum likelihood estimation procedure can be described as follows. The observed $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$ are projected orthogonally into points $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ on a line; see Figure 4. The line is chosen to minimize the sum of squared distances of $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$ to the line. The ratio of lengths of the two projections is the estimate of β. In formal terms one minimizes $\| \mathbf{x}_1^* - \boldsymbol{\xi}_1 \|^2 + \| \mathbf{x}_2^* - \boldsymbol{\xi}_2 \|^2$ subject to $\boldsymbol{\xi}_2 = \beta \boldsymbol{\xi}_1$ (Anderson and Sawa, 1982).

It seems appropriate in a Wald lecture to include a decision-theoretic result. In the two-dimensional space (Figures 1 and 2) the error structure ($\mathscr{E}\mathbf{u}\mathbf{u}' = \sigma^2 \mathbf{I}_2$) is invariant with respect to orthogonal transformations. The difference between the maximum likelihood estimate of the angle and the angle itself is also invariant. In the n-dimensional space (Figures 3 and 4) the error structure

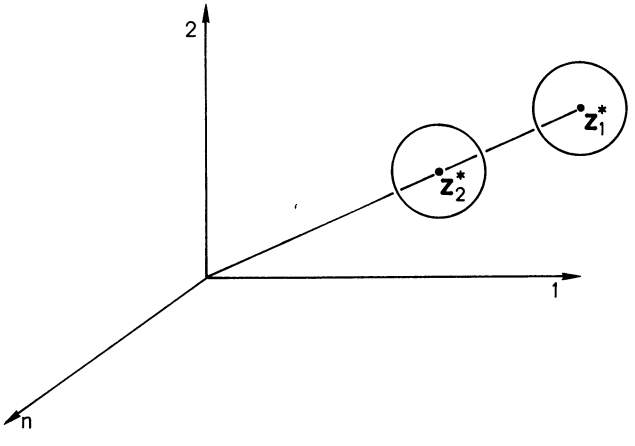

FIG. 3.

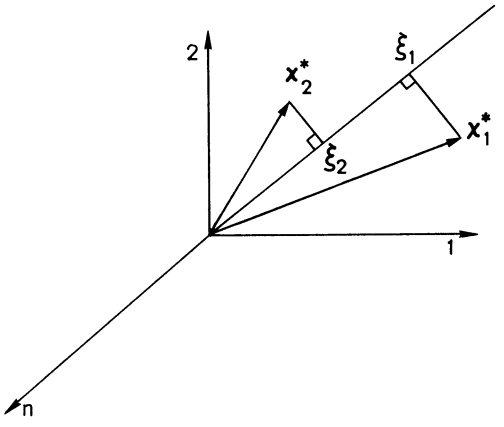

FIG. 4.

is invariant with respect to orthogonal transformations that leave the equiangular line invariant. As loss function we use

$$L(\theta, \hat{\theta}) = \sin^2(\hat{\theta} - \theta), \tag{2.9}$$

which is invariant with respect to the two groups of orthogonal transformations. The loss function takes account of the fact that θ is defined only to within a multiple of π. Then the maximum likelihood estimator of θ is the best invariant estimator. Consequently, it is admissible among all estimators. In fact, for fixed $\| \mathbf{z}_1^* \|^2 + \| \mathbf{z}_2^* \|^2$ and σ^2 the maximum likelihood estimator is admissible (Anderson, Stein and Zaman, 1982). These results are true for the means on a line ($m = 1$) for any dimensionality of observation (p).

3. A linear structural relationship: One relationship between two variables. In this case the systematic parts are random. We write the model as

$$x_1 = \mu_1 + v + u_1, \quad x_2 = \mu_2 + \beta v + u_3, \tag{3.1}$$

where μ_1 and u_2 are constants and

$$\mathscr{E} u_1 = \mathscr{E} u_2 = \mathscr{E} v = 0, \tag{3.2}$$

$$\mathscr{E} u_1^2 = \mathscr{E} u_2^2 = \sigma^2, \quad \mathscr{E} v^2 = \sigma_v^2. \tag{3.3}$$

If we assume u_1, u_2 and v are independently normally distributed, the model is determined by these five parameters μ_1, μ_2, σ^2, σ_v^2 and β. Then x_1 and x_2 have a bivariate normal distribution described by the five observable parameters $\mathscr{E} x_1$, $\mathscr{E} x_2$, Var x_1, Var x_2 and cov(x_1, x_2). These can be expressed in terms of the five model parameters. When the usual estimates of the observable parameters are inserted into these equations, we obtain maximum likelihood estimates of the model parameters. The resulting estimate of β is the same as in the linear functional relationship.

4. Linear functional relationships: Independent errors with equal variances. We now turn from the special case of $p = 2$ and $m = q = 1$ to the model defined by (1.1), (1.2), and (1.3), where p, $m(<p)$, and $q(<p)$ are arbitrary and subject to $m + q = p$ and the $\mathbf{z}_\alpha$'s are nonstochastic. With $\bar{\mathbf{z}} = (1/n)\sum_{\alpha=1}^{n} \mathbf{z}_\alpha$ we define $\boldsymbol{\nu}_\alpha = \mathbf{z}_\alpha - \bar{\mathbf{z}}$ and $\boldsymbol{\mu} = \bar{z}$. Then the model is written as

$$\mathbf{x}_\alpha = \boldsymbol{\nu}_\alpha + \boldsymbol{\mu} + \mathbf{u}_\alpha, \quad \alpha = 1, \cdots, n, \tag{4.1}$$

where $\mathbf{v}_\alpha$ satisfies (1.2),

$$\sum_{\alpha=1}^{n} \boldsymbol{\nu}_\alpha = \mathbf{0}, \tag{4.2}$$

and $n > p$. The relationship (1.3) is equivalent to

$$\mathbf{B}\boldsymbol{\nu}_\alpha = \mathbf{0}, \quad \alpha = 1, \cdots, n, \tag{4.3}$$

where $\mathbf{B}$ is a $q \times p$ matrix. The model can now be described in the terminology of the multivariate analysis of variance with a one-way classification; $\boldsymbol{\nu}_\alpha$ is the effect vector in the αth class.

When the errors are uncorrelated and have equal variances,

$$\boldsymbol{\Psi} = \sigma^2 \mathbf{I}_p. \tag{4.4}$$

To remove some of the indeterminacy in $\mathbf{B}$ we may require

$$\mathbf{B}\mathbf{B}' = \mathbf{I}_q \tag{4.5}$$

since the rank of $\mathbf{B}$ was assumed to be q. There remains the indeterminacy of multiplication of $\mathbf{B}$ on the left by an orthogonal matrix. We assume $q \geq 1$. (If $q = 0$, the model is unidentified; there are $np + 1$ parameters and only np observable data.)

Under the assumption of normality the logarithm of the likelihood function is

$$\begin{aligned} \log L = &-\frac{1}{2} np \log 2\pi - \frac{1}{2} np \log \sigma^2 \\ &- \frac{1}{2\sigma^2} \sum_{\alpha=1}^{n} (\mathbf{x}_\alpha - \boldsymbol{\mu} - \boldsymbol{\nu}_\alpha)'(\mathbf{x}_\alpha - \boldsymbol{\mu} - \boldsymbol{\nu}_\alpha). \end{aligned} \tag{4.6}$$

The maximum likelihood estimators are the values of $\boldsymbol{\mu}, \boldsymbol{\nu}_1, \cdots, \boldsymbol{\nu}_n, \sigma^2, \mathbf{B}$ that maximize (4.7) and satisfy (4.2), (4.3), and (4.5). It is evident from the form of (4.7) that the values of $\boldsymbol{\mu}, \boldsymbol{\nu}_1, \cdots, \boldsymbol{\nu}_n, \mathbf{B}$ maximizing the likelihood are the values minimizing the sum of squares in (4.7). Thus, maximum likelihood estimation is equivalent to (orthogonal) least squares estimation as in the case of $p = 2$. It should also be noticed that the values maximizing the likelihood are the same if σ^2 is known.

Let $\bar{\mathbf{x}} = (1/n)\sum_{\alpha=1}^{n} \mathbf{x}_\alpha$ and

$$\mathbf{C} = (1/n)\sum_{\alpha=1}^{n} (\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})'. \tag{4.7}$$

Let the characteristic roots of $\mathbf{C}$, that is, the roots of

$$|\mathbf{C} - t\mathbf{I}_p| = 0, \tag{4.8}$$

be $t_1 > \cdots > t_p$ (the roots being distinct with probability 1), and let the corresponding characteristic vectors be $\mathbf{w}_1, \cdots, \mathbf{w}_p$ normalized by $\mathbf{w}'\mathbf{w} = 1$, that is,

$$\mathbf{C}\mathbf{w}_i = t_i\mathbf{w}_i, \quad i = 1, \cdots, p. \tag{4.9}$$

Let

$$\mathbf{W}_1 = (\mathbf{w}_1, \cdots, \mathbf{w}_m), \quad \mathbf{W}_2 = (\mathbf{w}_{m+1}, \cdots, \mathbf{w}_p), \quad \mathbf{W} = (\mathbf{W}_1\ \mathbf{W}_2),$$

$$\mathbf{T}_1 = \begin{pmatrix} t_1 & 0 & \cdots & 0 \\ 0 & t_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & t_m \end{pmatrix}, \quad \mathbf{T}_2 = \begin{pmatrix} t_{m+1} & 0 & \cdots & 0 \\ 0 & t_{m+2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & t_p \end{pmatrix}, \tag{4.10}$$

$$\mathbf{T} = \begin{pmatrix} \mathbf{T}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_2 \end{pmatrix}.$$

Then $\mathbf{W}'\mathbf{W} = \mathbf{I}_p$, $\mathbf{C} = \mathbf{W}\mathbf{T}\mathbf{W}' = \mathbf{W}_1\mathbf{T}_1\mathbf{W}_1' + \mathbf{W}_2\mathbf{T}_2\mathbf{W}_2'$, and $\mathbf{I}_p = \mathbf{W}\mathbf{W}' = \mathbf{W}_1\mathbf{W}_1' + \mathbf{W}_2\mathbf{W}_2'$.

Maximum likelihood estimators of $\mathbf{B}$, $\nu_1, \cdots, \nu_n$, σ^2, and μ are

$$\hat{\mathbf{B}} = \mathbf{W}_2', \tag{4.11}$$

$$\hat{\nu}_\alpha = \mathbf{W}_1\mathbf{W}_1'(\mathbf{x}_\alpha - \bar{\mathbf{x}}), \quad \alpha = 1, \cdots, n, \tag{4.12}$$

$$\hat{\sigma}_F^2 = \frac{1}{p}\sum_{i=m+1}^{p} t_i, \tag{4.13}$$

and $\hat{\mu} = \bar{\mathbf{x}}$. (When the maximum likelihood estimators of a parameter are different in the functional and structural models, they will be distinguished by subscripts F and S, respectively.) When $\mathbf{W}_2'$ is multiplied on the left by an orthogonal matrix, another maximum likelihood estimator of $\mathbf{B}$ is obtained.

The directions indicated by the columns of $\mathbf{W}_2$ are the directions of minimum scatter; that is, $\mathbf{w}_i$ minimizes

$$\sum_{\alpha=1}^{n} [\mathbf{w}'(\mathbf{x}_\alpha - \bar{\mathbf{x}})]^2 = n\mathbf{w}'\mathbf{C}\mathbf{w} \tag{4.14}$$

subject to $\mathbf{w}'\mathbf{w}_j = 0$ (or equivalently $\mathbf{w}'\mathbf{C}\mathbf{w}_j = 0$), $j = i + 1, \cdots, p$. See Anderson (1958), Chapter 11. Similarly the directions defined by the columns of $\mathbf{W}_1$ maximize the scatter. The estimator of ν_α is the projection of $(\mathbf{x}_\alpha - \bar{\mathbf{x}})$ on the m-space defined by $\hat{\mathbf{B}}\mathbf{z} = \mathbf{0}$. It should be noted that in order to estimate σ^2 the dimension m must be less than p; there must be at least one linear relation on the expected values of the $\mathbf{x}_\alpha$'s. In effect the variability described by $t_1, \cdots, t_m$ is assigned to the variation in the means and the remainder $\sum_{i=m+1}^{p} t_i$ is assigned

to σ^2. The estimator of σ^2 is too small; for example, $\operatorname{plim}_{n\to\infty}\hat{\sigma}^2 = (q/p)\sigma^2$ when $(1/n)\sum_{\alpha=1}^{n}\boldsymbol{\nu}_\alpha\boldsymbol{\nu}_\alpha'$ converges to a matrix of rank m. The mean sum of squares and cross products of estimated factors is

$$\frac{1}{n}\sum_{\alpha=1}^{n}\hat{\boldsymbol{\nu}}_\alpha\hat{\boldsymbol{\nu}}_\alpha' = \mathbf{W}_1\mathbf{T}_1\mathbf{W}_1'. \tag{4.15}$$

The maximum likelihood estimator of the mean sum of squares and cross products of the observations is

$$\begin{aligned}\frac{1}{n}\sum_{\alpha=1}^{n}\hat{\boldsymbol{\nu}}_\alpha\hat{\boldsymbol{\nu}}_\alpha' + \hat{\sigma}_F^2\mathbf{I}_p &= \mathbf{W}_1\mathbf{T}_1\mathbf{W}_1' + \hat{\sigma}_F^2\mathbf{I}_p \\ &= \mathbf{W}_1(\mathbf{T}_1 + \hat{\sigma}_F^2\mathbf{I}_m)\mathbf{W}_1' + \hat{\sigma}_F^2\mathbf{W}_2\mathbf{W}_2'.\end{aligned} \tag{4.16}$$

Tintner (1945) obtained the maximum likelihood estimators of $\mathbf{B}$ and $\boldsymbol{\nu}_1, \cdots \boldsymbol{\nu}_n$, when $\boldsymbol{\Psi} = \boldsymbol{\Psi}_0$ is known completely and gave examples of their use (Tintner, 1946, 1952). The estimators are the same as when $\boldsymbol{\Psi} = \sigma^2\boldsymbol{\Psi}_0$ and σ^2 is unknown. (For σ^2 known Tintner proposed a test of the hypothesis $m = m_0$ against the alternative $m > m_0$ based on $\sum_{i=m_0+1}^{p} t_i$. When σ^2 is unknown, a test is not possible because the estimate of σ^2 is vacuous for $m = p$.) Geary (1948) discussed them further. Malinvaud (1964) derived the same estimators. Sprent (1966) obtained the estimators by generalizd least squares in a more general setting. Gleser and Watson (1973) found the maximum likelihood estimators in the special case $m = q$; Bhargava (1979) noted that their argument holds for $m \leq q$. (In fact it holds for any m and q.) Theobald (1975) used a purely algebraic method. Gleser (1981) showed that the estimators minimize any orthogonally invariant norm of $(\mathbf{B}\mathbf{B}')^{-1/2}\mathbf{B}(\mathbf{x}_\alpha - \boldsymbol{\mu})$, $\alpha = 1, \cdots, n$. See also Eckart and Young (1936), Rao (1973), Section 8c.6, Höschel and Chan (1980).

5. Linear structural relationships: Independent errors with equal variances. In the structural case with a random effect it will be convenient to formulate (1.1) in a manner similar to (4.1). If $\mathscr{E}\mathbf{z}_\alpha = \boldsymbol{\mu}$ and $\mathbf{v}_\alpha = \mathbf{z}_\alpha - \boldsymbol{\mu}$, (1.1) is

$$\mathbf{x}_\alpha = \mathbf{v}_\alpha + \boldsymbol{\mu} + \mathbf{u}_\alpha, \tag{5.1}$$

where $\mathbf{u}_\alpha$ satisfies (1.2) and $\mathscr{E}\mathbf{v}_\alpha = \mathbf{0}$ (analogous to (4.2)). The random (unobservable) $\mathbf{v}_\alpha$ is assumed to satisfy the linear relationship

$$\mathbf{B}\mathbf{v}_\alpha = \mathbf{0} \tag{5.2}$$

with probability 1 (analogous to (4.3)). Then the covariance matrix of $\mathbf{v}_\alpha$ $\mathscr{E}\mathbf{v}_\alpha\mathbf{v}_\alpha' = \boldsymbol{\Theta}$ satisfies this linear relationship, that is, $\mathbf{B}\boldsymbol{\Theta} = \mathbf{0}$. Sometimes it is convenient to write the covariance matrix $\boldsymbol{\Theta}$ as $\boldsymbol{\Lambda}\boldsymbol{\Lambda}'$, where $\boldsymbol{\Lambda}$ is a $p \times m$ matrix of rank m. (In terms of the "parametric" form (1.4) we have taken $\mathbf{f}$ to have moments $\mathscr{E}\mathbf{f} = 0$ and $\mathscr{E}\mathbf{f}\mathbf{f}' = \mathbf{I}_m$.) The covariance matrix of $\mathbf{x}_\alpha$ is

$$\boldsymbol{\Sigma} = \boldsymbol{\Psi} + \boldsymbol{\Theta} = \boldsymbol{\Psi} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}'. \tag{5.3}$$

We again treat the case $\boldsymbol{\Psi} = \sigma^2\mathbf{I}_p$ and require $\mathbf{B}\mathbf{B}' = \mathbf{I}_q$.

As in the case of linear functional relationships, a maximum likelihood estimator of $\mathbf{B}$ is again $\hat{\mathbf{B}} = \mathbf{W}_2'$; other maximum likelihood estimators are

obtained by multiplying $\mathbf{W}_2'$ on the left by orthogonal matrices. (The maximum likelihood estimators of $\mathbf{B}$, $\mathbf{\Theta}$, and $\mathbf{\Sigma}$ are not the same for σ^2 known as for σ^2 unknown in this model.) The maximum likelihood estimator of σ^2 is

$$\hat{\sigma}_S^2 = \frac{1}{q}\sum_{i=m+1}^{p} t_i, \tag{5.4}$$

the average of the q smallest characteristic roots of $\mathbf{C}$. The maximum likelihood estimator of $\mathbf{\Theta}$ is

$$\hat{\mathbf{\Theta}} = \mathbf{W}_1\mathbf{T}_1\mathbf{W}_1' - \hat{\sigma}_S^2\mathbf{W}_1\mathbf{W}_1' = \mathbf{W}_1(\mathbf{T}_1 - \hat{\sigma}_S^2\mathbf{I}_p)\mathbf{W}_1', \tag{5.5}$$

and the estimator of the covariance matrix of $\mathbf{x}_\alpha$ is

$$\hat{\mathbf{\Sigma}}_S = \hat{\sigma}_S^2\mathbf{I}_p + \hat{\mathbf{\Theta}} = \mathbf{W}_1\mathbf{T}_1\mathbf{W}_1' + \hat{\sigma}_S^2\mathbf{W}_2\mathbf{W}_2'. \tag{5.6}$$

Note that here more of the total variability in terms of $t_1, \cdots, t_p$ is assigned to the directions represented by the columns of $\mathbf{W}_2$ than in the case of the linear functional relationship. The estimator of the mean is $\hat{\mu} = \bar{\mathbf{x}}$. These estimators were obtained by Lawley (1953) and by Theobald (1975); their asymptotic distributions were treated by Anderson (1963).

If the rank of $\mathscr{E}\mathbf{v}_\alpha\mathbf{v}_\alpha' = \mathbf{\Theta}$ is m, then $\mathbf{\Sigma} = \mathbf{\Theta} + \sigma^2\mathbf{I}_p$ has $q = p - m$ characteristic roots equal to σ^2. The null hypothesis that the rank of the covariance matrix of the (unobservable) random systematic part is m_0 (that is, that there are $p - m_0$ linear constraints) is the hypothesis that the m_0 smallest characteristic roots of $\mathbf{\Sigma}$ are equal. The alternative that the rank of $\mathbf{\Theta}$ is greater than m_0 is equivalent to $\mathbf{\Sigma}$ being an arbitrary positive definite matrix. The likelihood ratio criterion λ for testing the hypothesis $m = m_0$ against the alternative $m > m_0$ is defined by

$$\lambda^{2/n} = \frac{|\hat{\mathbf{\Sigma}}_\Omega|}{|\hat{\mathbf{\Sigma}}_\omega|} = \frac{\prod_{i=m_0+1}^{p} t_i}{[(1/q_0)\sum_{i=m_0+1}^{p} t_i]^{q_0}}, \tag{5.7}$$

where $q_0 = p - m_0$. The criterion is the $\frac{1}{2}nq_0$th power of the ratio of the geometric mean of the q_0 smallest roots to the arithmetic mean. As $n \to \infty$, the limiting distribution of $-2\log\lambda$ is χ^2 with $\frac{1}{2}q_0(q_0+1) - 1$ degrees of freedom when the hypothesis is true.

What is the connection between the analysis of linear functional and structural relationships and *principal component analysis* introduced by Hotelling (1933)? The sample principal components of a vector $\mathbf{x}$ are the components of $\mathbf{W}'\mathbf{x}$. They can be considered as forming a new coordinate system with intrinsic statistical meaning. The sample variance of $\mathbf{w}_i'\mathbf{x}$ is t_i and the principal components are uncorrelated in the sample. A primary use of principal components is to reduce the number of variables considered by discarding the principal components with small variances. This amounts to approximating a random vector $\mathbf{x}$ by its first few principal components; more precisely $\mathbf{W}'\mathbf{x}$ is approximated by $\mathbf{W}_1'\mathbf{x}$. It is anticipated that statistical analysis of $\mathbf{W}_1'\mathbf{x}$ will give nearly all the useful information about $\mathbf{x}$. In the case of the linear functional or structural analysis $\mathbf{W}_1'\mathbf{x}$ is considered as the estimate of the systematic part of $\mathbf{x}$ and $\mathbf{x} - \mathbf{W}_1\mathbf{W}_1'\mathbf{x} = \mathbf{W}_2\mathbf{W}_2'\mathbf{x}$ can be considered as the estimate of the error in the q-space orthogonal to the systematic part. (In the m-space of $\mathbf{W}_1'\mathbf{x}$ the systematic part

and error cannot be distinguished.) The error of approximation of $\mathbf{x}$ by $\mathbf{W}_1'\mathbf{x}$ is not necessarily small, but should have the attributes of $\mathbf{u}$, namely, mean $\mathbf{0}$ and covariance matrix corresponding to equal roots.

In principal component analysis $\mathbf{W}_1'\mathbf{x}$ may be a good approximation to $\mathbf{W}'\mathbf{x}$ if $t_{m+1}, \cdots, t_p$ are small. In the linear structural analysis $\mathbf{W}_1'\mathbf{x}$ is a good estimate of the systematic part if $t_{m+1}, \cdots, t_p$ are nearly equal (not necessarily small!).

6. Linear functional relationships: Arbitrary error covariance matrix. When the error covariance matrix is completely unspecified, replicated observations are needed in order to estimate it. We shall suppose that there are $k \geq 2$ observations made on the vector with expected value $\nu_\alpha + \mu$. For ease in exposition we have taken the number of observations on all the vectors to be the same, though that is not necessary. We write

$$\mathbf{x}_{\alpha j} = \nu_\alpha + \mu + \mathbf{u}_{\alpha j}, \quad \alpha = 1, \cdots, n, \quad j = 1, \cdots, k, \tag{6.1}$$

where $\mathbf{u}_{\alpha j}$ and ν_α satisfy (1.2), (4.2), and (4.3). This is equivalent to the multivariate analysis of variance model with fixed factors for the one-way classification.

Let $\bar{\mathbf{x}}_\alpha = (1/k)\sum_{j=1}^k \mathbf{x}_{\alpha j}$ and $\bar{\mathbf{x}} = (1/n)\sum_{\alpha=1}^n \bar{\mathbf{x}}_\alpha$. The usual analysis of variance table is

Source	Sum of Squares	d.f.
Effect	$\mathbf{H} = k\sum_{\alpha=1}^n (\bar{\mathbf{x}}_\alpha - \bar{\mathbf{x}})(\bar{\mathbf{x}}_\alpha - \bar{\mathbf{x}})'$	$n-1$
Error	$\mathbf{G} = \sum_{\alpha=1}^n \sum_{j=1}^k (\mathbf{x}_{\alpha j} - \bar{\mathbf{x}}_\alpha)(\mathbf{x}_{\alpha j} - \bar{\mathbf{x}}_\alpha)'$	$n(k-1)$
Total	$\sum_{\alpha=1}^n \sum_{j=1}^k (\mathbf{x}_{\alpha j} - \bar{\mathbf{x}})(\mathbf{x}_{\alpha j} - \bar{\mathbf{x}})'$	$nk-1$

Let

$$\tilde{\mathbf{H}} = \frac{1}{n}\mathbf{H}, \quad \tilde{\mathbf{G}} = \frac{1}{n(k-1)}\mathbf{G}. \tag{6.2}$$

Then $\tilde{\mathbf{G}}$ is an unbiased estimator of Ψ, and

$$\mathscr{E}\tilde{\mathbf{H}} = \frac{n-1}{n}\Psi + \frac{k}{n}\sum_{\alpha=1}^n \nu_\alpha\nu_\alpha'. \tag{6.3}$$

If there were no linear relationship, $\mathbf{B}$ would be $\mathbf{0}$ and the maximum likelihood estimators of Ψ and ν_α would be $[n(k-1)/(nk)]\tilde{\mathbf{G}}$ and $\bar{\mathbf{x}}_\alpha - \bar{\mathbf{x}}$, respectively; if there were p linear relationships, the estimator of $\mathbf{B}$ could be any nonsingular matrix, ν_α would be $\mathbf{0}$ and the maximum likelihood estimator of Ψ would be a weighted average of $\tilde{\mathbf{H}}$ and $\tilde{\mathbf{G}}$, namely $[1/k][(k-1)\tilde{\mathbf{G}} + \tilde{\mathbf{H}}]$. When there are q $(0 < q < p)$ linear relationships, the estimators are compromises between the above extremes.

Let the roots of

$$|\tilde{\mathbf{H}} - d\tilde{\mathbf{G}}| = 0 \tag{6.4}$$

be $d_1 > \cdots > d_p$ (the roots being distinct with probability 1), and let $\mathbf{y}_1, \cdots, \mathbf{y}_p$

be the vectors satisfying

$$\tilde{\mathbf{H}}y_i = d_i\tilde{\mathbf{G}}y_i, \quad i = 1, \cdots, p, \tag{6.5}$$

and $\mathbf{y}_i'\tilde{\mathbf{G}}\mathbf{y}_i = 1$. These are the characteristic roots and vectors of $\tilde{\mathbf{G}}^{-1}\tilde{\mathbf{H}}$. Let

$$\mathbf{Y}_1 = (\mathbf{y}_1, \cdots, \mathbf{y}_m), \quad \mathbf{Y}_2 = (\mathbf{y}_{m+1}, \cdots, \mathbf{y}_p), \quad \mathbf{Y} = (\mathbf{Y}_1\ \mathbf{Y}_2),$$

$$\mathbf{D}_1 = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & d_m \end{pmatrix}, \quad \mathbf{D}_2 = \begin{pmatrix} d_{m+1} & 0 & \cdots & 0 \\ 0 & d_{m+2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & d_p \end{pmatrix}, \tag{6.6}$$

$$\mathbf{D} = \begin{pmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2 \end{pmatrix}.$$

Then $\mathbf{Y}'\tilde{\mathbf{G}}\mathbf{Y} = \mathbf{I}_p$ and $\mathbf{Y}'\tilde{\mathbf{H}}\mathbf{Y} = \mathbf{D}$. It will be convenient to define $(\mathbf{Y}')^{-1} = \mathbf{Z} = (\mathbf{Z}_1\ \mathbf{Z}_2)$; then $\tilde{\mathbf{G}} = \mathbf{Z}\mathbf{Z}'$ and $\tilde{\mathbf{H}} = \mathbf{Z}\mathbf{D}\mathbf{Z}' = \mathbf{Z}_1\mathbf{D}_1\mathbf{Z}_1' + \mathbf{Z}_2\mathbf{D}_2\mathbf{Z}_2'$. ($\mathbf{Y}$ and $\mathbf{Z}$ are uniquely defined except that any column can be multiplied by -1). Maximum likelihood estimators of $\boldsymbol{\mu}$, $\mathbf{B}$, and $\boldsymbol{\nu}_1, \cdots, \boldsymbol{\nu}_n$ are $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$,

$$\hat{\mathbf{B}} = \mathbf{Y}_2', \tag{6.7}$$

$$\hat{\boldsymbol{\nu}}_\alpha = \mathbf{G}\mathbf{Y}_1\mathbf{Y}_1'(\bar{\mathbf{x}}_\alpha - \bar{\mathbf{x}}) = \mathbf{Z}_1\mathbf{Y}_1'(\bar{\mathbf{x}}_\alpha - \bar{\mathbf{x}}), \quad \alpha = 1, \cdots, n. \tag{6.8}$$

Thus $\hat{\mathbf{B}}\hat{\boldsymbol{\nu}}_\alpha = \mathbf{0}$; the rows of $\hat{\mathbf{B}}$ (columns of $\mathbf{Y}_2$) are orthogonal to $\hat{\boldsymbol{\nu}}_\alpha$, $\alpha = 1, \cdots, n$. Note that

$$\frac{k}{n}\sum_{\alpha=1}^{n}\hat{\boldsymbol{\nu}}_\alpha\hat{\boldsymbol{\nu}}_\alpha' = \mathbf{Z}_1\mathbf{D}_1\mathbf{Z}_1'; \tag{6.9}$$

this is the part of $\tilde{\mathbf{H}} = \mathbf{Z}\mathbf{D}\mathbf{Z}'$ accounted for by the fixed effects. The maximum likelihood estimator of the error covariance matrix is

$$\hat{\boldsymbol{\Psi}}_F = \frac{1}{k}[(k-1)\tilde{\mathbf{G}} + \mathbf{Z}_2\mathbf{D}_2\mathbf{Z}_2']; \tag{6.10}$$

this is a weighted average of $\tilde{\mathbf{G}}$, the unbiased estimator of $\boldsymbol{\Psi}$, and $\mathbf{Z}_2\mathbf{D}_2\mathbf{Z}_2'$, which is the part of $\tilde{\mathbf{H}}$ not assigned to the fixed effects. Note that regardless of the assumed rank

$$\hat{\boldsymbol{\Psi}} + \frac{1}{n}\sum_{\alpha=1}^{n}\hat{\boldsymbol{\nu}}_\alpha\hat{\boldsymbol{\nu}}_\alpha' = \frac{1}{k}[(k-1)\tilde{\mathbf{G}} + \tilde{\mathbf{H}}]. \tag{6.11}$$

The vector $\mathbf{y}_i$ satisfying (6.5) minimizes

$$\frac{k\sum_{\alpha=1}^{n}[\mathbf{y}'(\bar{\mathbf{x}}_\alpha - \bar{\mathbf{x}})]^2}{\sum_{\alpha=1}^{n}\sum_{j=1}^{k}[\mathbf{y}'(\mathbf{x}_{\alpha j} - \bar{\mathbf{x}}_\alpha)]^2} = \frac{\mathbf{y}'\tilde{\mathbf{H}}\mathbf{y}}{\mathbf{y}'\tilde{\mathbf{G}}\mathbf{y}} \tag{6.12}$$

among vectors satisfying $\mathbf{y}'\tilde{\mathbf{G}}\mathbf{y}_j = 0, j = m+1, \cdots, p$. The q rows of $\hat{\mathbf{B}}$ (that is, the q columns of $\mathbf{Y}_2$) are in the directions of minimum scatter.

These results were obtained by Anderson (1951a). (In their surveys Madansky

(1959) and Moran (1971) overlooked this paper as well as Anderson and Rubin (1956).) The results were rediscovered and extended by Villegas (1961), Hannan (1967), Dolby (1976), Nussbaum (1976), and Healy (1980).

Fisher (1938) had considered a related problem in discriminant analysis: how to discriminate among n populations when the means $(\mu + \nu_\alpha)$ are linearly related. See also Bartlett (1947), (1948).

The likelihood ratio criterion λ for testing the hypothesis that $m = m_0$ against the alternative $m > m_0$ is defined by

$$\begin{aligned}\lambda^{2/n} &= (k-1)^{k(p-m_0)} \left| (k-1)\mathbf{I}_{p-m_0} + \mathbf{D}_2 \right|^{-k} \\ &= \prod_{i=m_0+1}^{p} \left(\frac{k-1}{k-1+d_i} \right)^k. \end{aligned} \tag{6.13}$$

When the null hypothesis is true, $-2 \log \lambda$ has a limiting χ^2-distribution with $(p - m_0)(n - 1 - m_0)$ degrees of freedom as $k \to \infty$ and n is fixed. See Anderson (1951b).

7. Linear structural relationships: Arbitrary error covariance matrix. We write the model as

$$\mathbf{x}_{\alpha j} = \mathbf{v}_\alpha + \mu + \mathbf{u}_{\alpha j}, \quad \alpha = 1, \cdots, n, \quad j = 1, \cdots, k, \tag{7.1}$$

where $\mathbf{u}_{\alpha j}$ satisfies (1.2), $\mathscr{E}\mathbf{v}_\alpha = \mathbf{0}$, and $\mathscr{E}\mathbf{v}_\alpha\mathbf{v}_\alpha' = \mathbf{\Theta} = \mathbf{\Lambda}\mathbf{\Lambda}'$, as in Section 5. (Dolby (1976) has combined the linear functional and structural relationships into a more general model, called the ultrastructural relationship, where $\mathbf{z}_\alpha$ in (1.1) has the distribution $N(\nu_\alpha, \mathbf{\Theta})$.) The covariance matrix of $\mathbf{x}_{\alpha j}$ is (5.3). We have

$$\mathscr{E}\tilde{\mathbf{H}} = \frac{n-1}{n}(\mathbf{\Psi} + k\mathbf{\Theta}), \quad \mathscr{E}\tilde{\mathbf{G}} = \mathbf{\Psi}. \tag{7.2}$$

If there were no linear relationship, then $\mathbf{B}$ would be $\mathbf{0}$ and the maximum likelihood estimators of $\mathbf{\Psi}$ and $\mathbf{\Theta}$ would be $\tilde{\mathbf{G}}$ and $(1/k)(\tilde{\mathbf{H}} - \tilde{\mathbf{G}})$ provided the latter is positive semidefinite; if there were p linear relationships, $\mathbf{B}$ could be any nonsingular matrix, $\mathbf{\Theta}$ would be $\mathbf{0}$, and the maximum likelihood estimator of $\mathbf{\Psi}$ would be the same weighted average of $\tilde{\mathbf{G}}$ and $\tilde{\mathbf{H}}$ as in the case of p linear functional relationships. It should be noted that

$$\frac{1}{k}(\tilde{\mathbf{H}} - \tilde{\mathbf{G}}) \doteq \frac{1}{k}\mathbf{Z}(\mathbf{D} - \mathbf{I}_p)\mathbf{Z}' \tag{7.3}$$

is almost an unbiased estimator of $\mathbf{\Theta}$, but if it is not positive semidefinite it cannot be the maximum likelihood estimator. In that case the maximum likelihood estimator has the form of (7.3), but the diagonal elements of $\mathbf{D}$ that are less than 1 are replaced by 1. The rank of $\hat{\mathbf{\Theta}}$ then is the number of $d_1, \cdots, d_p$ that exceed 1. (The probability of any root being 1 is 0.)

Let p^* be the number of roots greater than 1, and let $m^* = \min(m, p^*)$. Let

$$\mathbf{D}_1^* = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & d_{m^*} \end{pmatrix}, \quad \mathbf{D}_2^* = \begin{pmatrix} d_{m^*+1} & 0 & \cdots & 0 \\ 0 & d_{m^*+2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & d_p \end{pmatrix}, \tag{7.4}$$

$\mathbf{Y}_1^*$ consist of the first m^* columns of $\mathbf{Y}$ and $\mathbf{Y}_2^*$ consist of the remaining columns, $\mathbf{Z}_1^*$ the first m^* columns of $\mathbf{Z}$ and $\mathbf{Z}_2^*$ the remaining columns. Our model now is that there are *at least* $p - m = q$ linear relationships; that is, $\mathbf{B}$ has at least q rows. Then a maximum likelihood estimator of $\mathbf{B}$ is $\mathbf{Y}_2^{*\prime}$. The maximum likelihood estimator of $\boldsymbol{\Theta}$ is

$$\hat{\boldsymbol{\Theta}} = \frac{1}{k}\mathbf{Z}_1^*(\mathbf{D}_1^* - \mathbf{I}_{m^*})\mathbf{Z}_1^{*\prime}; \tag{7.5}$$

this estimator is positive semidefinite of rank m^*. If all the roots are less than 1, (7.5) is vacuous and $\hat{\boldsymbol{\Theta}} = \mathbf{0}$. Note here that $\hat{\mathbf{B}}\hat{\boldsymbol{\Theta}} = \mathbf{0}$ as required. The estimator of $\boldsymbol{\Theta}$ is based on the part of $\tilde{\mathbf{H}}$ that is associated with the m^* columns of $\mathbf{Z}$ with a correction based on $\tilde{\mathbf{G}}$ because $\tilde{\mathbf{H}}$ estimates $k\boldsymbol{\Theta} + \boldsymbol{\Psi}$. The maximum likelihood estimator of the covariance matrix $\boldsymbol{\Psi}$ is

$$\hat{\boldsymbol{\Psi}}_S = \frac{1}{k}[(k-1)\tilde{\mathbf{G}} + (\mathbf{Z}_2^*\mathbf{D}_2^*\mathbf{Z}_2^{*\prime} + \mathbf{Z}_1^*\mathbf{Z}_1^{*\prime})]; \tag{7.6}$$

this is a weighted average of $\tilde{\mathbf{G}}$, the unbiased estimator of $\boldsymbol{\Psi}$, and the sum of $\mathbf{Z}_2^*\mathbf{D}_2^*\mathbf{Z}_2^{*\prime}$, which is the part of $\tilde{\mathbf{H}}$ not assigned to the class effects, and $\mathbf{Z}_1^*\mathbf{Z}_1^{*\prime}$, which was not included in $\hat{\boldsymbol{\Theta}}$. If more than q roots are less than 1, $\mathbf{D}_2^*$ has fewer elements than $\mathbf{D}_2$ and $\mathbf{Z}_2\mathbf{D}_2\mathbf{Z}_2' - \mathbf{Z}_2^*\mathbf{D}_2^*\mathbf{Z}_2^{*\prime}$ is positive semidefinite. Regardless of the assumed rank

$$\hat{\boldsymbol{\Psi}}_S + \hat{\boldsymbol{\Theta}} = \frac{1}{k}[(k-1)\tilde{\mathbf{G}} + \tilde{\mathbf{H}}]. \tag{7.7}$$

The likelihood ratio criterion λ for testing the hypothesis that $m \leq m_0$ against the alternative $m > m_0$ is defined by

$$\lambda^{2/n} = \prod_{i=m_0+1}^{p^*} \frac{k^k d_i}{(k - 1 + d_i)^k}; \tag{7.8}$$

if $p^* \leq m_0$, the criterion is 1. The hypothesis is rejected if the criterion is too small. The usual asymptotic theory as $n \to \infty$ does not hold here; the limiting distribution of $-2 \log \lambda$ is not a χ^2-distribution.

Anderson (1946b) found maximum likelihood estimators and likelihood ratio criteria in terms of the components of variance model as did Morris and Olkin (1964) independently. Tukey (1951) discussed an example of the statistical analysis. Klotz and Putter (1969) published the maximum likelihood estimators for the case of no linear relationships assumed ($m = p$). Amemiya and Fuller

(1984) obtained maximum likelihood estimators when the rank of $\mathbf{\Theta}$ is taken to be exactly m; if $p^* < m$, the maximum likelihood estimators in this model do not exist. They also found the likelihood ratio criterion to test the null hypothesis that $m = m_0$ against the alternative that $\mathbf{\Psi} + k\mathbf{\Theta}$ is an arbitrary positive definite matrix.

Part II. Factor analysis.

8. The factor analysis model. We write our model in the "parametric" form

$$\mathbf{x}_\alpha = \mathbf{\Lambda}\mathbf{f}_\alpha + \boldsymbol{\mu} + \mathbf{u}_\alpha, \tag{8.1}$$

where $\mathscr{E}\mathbf{u}_\alpha = \mathbf{0}$ and $\mathscr{E}\mathbf{u}_\alpha\mathbf{u}_\alpha' = \mathbf{\Psi}$. The $p \times m$ matrix $\mathbf{\Lambda}$ consists of "factor loadings," the vector $\mathbf{f}_\alpha$ consists of m "common" factor scores, and $\mathbf{u}_\alpha$ represents the errors. In psychological terms a component of $\mathbf{u}_\alpha$ may include a random factor unique to that test as well as an error. An essential assumption is that the error covariance matrix $\mathbf{\Psi}$ is diagonal. As indicated in the introduction, (8.1) is one way of representing the linear functional or structural relationship.

I. In the case of fixed factors we assume

$$\textstyle\sum_{\alpha=1}^{n} \mathbf{f}_\alpha = \mathbf{0}, \quad \frac{1}{n}\sum_{\alpha=1}^{n} \mathbf{f}_\alpha\mathbf{f}_\alpha' = \mathbf{\Phi}, \tag{8.2}$$

where $\mathbf{\Phi}$ is unknown. The $\mathbf{f}_\alpha$, $\alpha = 1, \cdots, n$, are unknown "incidental" parameters.

II. In the case of random factors we have

$$\mathscr{E}\mathbf{f}_\alpha = \mathbf{0}, \quad \mathscr{E}\mathbf{f}_\alpha\mathbf{f}_\alpha' = \mathbf{\Phi}. \tag{8.3}$$

Furthermore, $\mathbf{f}_1, \cdots, \mathbf{f}_n$ are independent of $\mathbf{u}_1, \cdots, \mathbf{u}_n$. Then $\mathscr{E}\mathbf{x}_\alpha = \boldsymbol{\mu}$ and

$$\operatorname{cov}(\mathbf{x}_\alpha) = \mathscr{E}(\mathbf{x}_\alpha - \boldsymbol{\mu})(\mathbf{x}_\alpha - \boldsymbol{\mu})' = \mathbf{\Sigma} = \mathbf{\Psi} + \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}'. \tag{8.4}$$

We can distinguish two cases. In the case of "orthogonal" factors

$$\mathbf{\Phi} = \mathbf{I}_m. \tag{8.5}$$

This implies with random factors that

$$\mathbf{\Sigma} = \mathbf{\Psi} + \mathbf{\Lambda}\mathbf{\Lambda}'. \tag{8.6}$$

The observed covariance matrix is considered to be the sum of a positive definite diagonal matrix and a positive semidefinite matrix of rank m. In the case of "oblique" factors the covariance matrix of the factors is arbitrary. We shall discuss the interpretation and use of oblique factors later.

One can take two views of this model. One view is that the model represents some reality. In particular, the components of $\mathbf{x}_\alpha$ can be scores on psychological tests and the components of $\mathbf{f}_\alpha$ can be values of intelligence factors or primary abilities. The latter are more fundamental or intrinsic. Spearman (1904) proposed

a general intelligence factor to explain the positive correlations observed between scores on intelligence tests—the single-factor model. This idea was generalized into the multiple-factor model as presented here (Garnett, 1919); the foremost developer was Thurstone (1935), (1947).

Another point of view is that this model is a way of describing mutual dependence. As we shall see later, the location and scale descriptions (depending on means and variances) can be distinguished from interdependence descriptions (depending on the correlations).

As noted before, (8.1) is another way of writing the linear functional and structural relationships. We have $\mathbf{B}\Lambda = \mathbf{0}$, where $\mathbf{B}$ is a $q \times p$ matrix and $q = p - m$.

9. Identification and "rotation". As observed earlier, in (8.1) Λ can be replaced by $\Lambda\mathbf{C}$ and $\mathbf{f}_\alpha$ by $\mathbf{C}^{-1}\mathbf{f}_\alpha$ without changing $\Lambda\mathbf{f}_\alpha$. There are several ways of resolving this indeterminacy.

1. One way is to require $\Phi = \mathbf{I}$, the case of orthogonal factors. There remains the indeterminacy of multiplication by an orthogonal $\mathbf{C}$, a "rotation." A convenient way to eliminating this indeterminacy is to require

$$\Lambda'\Psi^{-1}\Lambda = \Gamma = \text{Diagonal.} \tag{9.1}$$

The diagonal elements of Γ are the nonzero roots of

$$|\Sigma - \Psi - \gamma\Psi| = 0. \tag{9.2}$$

Condition (9.1) determines Λ uniquely when the roots are different and the diagonal elements of Γ are ordered.

2. Another method of obtaining uniqueness is to require

$$\Lambda = \begin{pmatrix} \mathbf{I}_m \\ \Lambda_2 \end{pmatrix}. \tag{9.3}$$

This leads to a relatively simple covariance matrix of the asymptotic distribution of the maximum likelihood estimators.

3. A more general way, which includes the above, is to require 0's in specified positions and a normalization of each column of Λ. For example, we might specify

$$\mathbf{x} = \mu + \begin{bmatrix} 1 & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & \lambda_{32} \\ 0 & \lambda_{42} \\ 0 & 1 \end{bmatrix} \begin{pmatrix} v \\ a \end{pmatrix} + \mathbf{u} = \mu + \begin{bmatrix} v \\ \lambda_{21}v \\ \lambda_{31}v + \lambda_{32}a \\ \lambda_{42}a \\ a \end{bmatrix} + \mathbf{u}. \tag{9.4}$$

The components of $\mathbf{x}$ may be the scores on 5 tests, and v and a measures of verbal and arithmetic abilities. The first two tests depend only on verbal ability and the last two tests depend only on arithmetic ability. There are two specified 0's in each column. In addition, the scales of v and a are determined by

the first element of the first column and the last element of the second column. In general, there must be at least $m - 1$ specified 0's in each column. The scales of the factors may alternatively be specified by requiring their variances to be 1 ($\phi_{ii} = 1$).

The number of observable variances and covariances is $\frac{1}{2}p(p + 1)$. The number of parameters in Ψ and Λ is $p + pm$. To eliminate the indeterminacy of a rotation (when $\Phi = \mathbf{I}_m$) there are $\frac{1}{2}m(m - 1)$ restrictions, such as (9.1). If

$$\frac{1}{2}p(p + 1) + \frac{1}{2}m(m - 1) - (p + pm) = \frac{1}{2}[(p - m)^2 - (p + m)] \tag{9.5}$$

is positive, Σ and the restrictions will usually uniquely determine Ψ and Λ; if (9.5) is negative, Ψ and Λ are not uniquely determined; and if (9.5) is 0, there are usually a finite number of solutions to (8.4). Anderson and Rubin (1956) give more precise statements.

Rotation. A requirement such as (9.1) is a convenient method of insuring uniqueness, but the resulting Λ does not have intrinsic meaning. The factor analyst may apply a transformation (orthogonal in the case of $\Phi = \mathbf{I}$ or nonsingular in any case) to obtain a matrix of factor loadings Λ that can be interpreted in terms of the subject matter. The attempt is to make many factor loadings close to 0. Then the interpretation is that certain factors are absent on certain tests; that helps identify the factors. A matrix with many 0's is said to have *simple structure.* Thurstone (1935), (1947) gave some rules about the pattern of 0's. See also Reiersøl (1950).

10. Estimation of the factor structure in the case of fixed factors: Nonexistence of maximum likelihood estimators. Let $\mathbf{x}_\alpha = (x_{i\alpha}, \cdots, x_{p\alpha})'$ be an observation on the variable defined by (8.1) with $\mathbf{f}_\alpha$ being a nonstochastic vector (an incidental parameter), $\alpha = 1, \cdots, n$, satisfying (8.2) for $\Phi = \mathbf{I}_m$. The likelihood function is

$$L = \frac{1}{((2\pi)^p \prod_{i=1}^p \psi_{ii})^{n/2}} \exp\left\{-\frac{1}{2}\sum_{\alpha=1}^n \sum_{i=1}^p \frac{(x_{i\alpha} - \mu_i - \sum_{j=1}^m \lambda_{ij} f_{j\alpha})^2}{\psi_{ii}}\right\}. \tag{10.1}$$

The likelihood function does not have a maximum. To show this fact, let $\mu_1 = 0$, $\lambda_{11} = 1$, $\lambda_{1j} = 0$, $j \neq 1$, $f_{1\alpha} = x_{1\alpha}$, $\alpha = 1, \cdots, n$. Then

$$\lambda_{11} f_{1\alpha} + \mu_1 = x_{1\alpha}, \quad \alpha = 1, \cdots, n. \tag{10.2}$$

The numerator in the exponent for $i = 1$ is 0 and ψ_{11} appears only in the constant. As $\psi_{11} \to 0$, $L \to \infty$. Consequently, the likelihood function does not have a maximum and the maximum likelihood estimators do not exist (Anderson and Rubin, 1956). Lawley (1941) set the partial derivatives of the likelihood equal to 0, but Solari (1969) showed that the solution yields only a stationary value, but not a maximum. Lawley's experience with numerical examples was that estimates of ψ_{ii}'s converged toward 0. Unaware of the nonexistence of maximum likelihood estimators, Lindley and El-Sayyad (1968) recommended a Bayesian approach for $p = 2$; see also Lindley (1947). Copas (1972) studied the likelihood function when rounding errors were included.

11. Estimation of the factor structure in the case of random factors. Let $\mathbf{x}_\alpha$ be an observation on the random variable given by (8.1) satisfying (8.3) for $\Phi = \mathbf{I}_m$. If $\mathbf{f}_\alpha$ and $\mathbf{u}_\alpha$ are independently normally distributed, then $\mathbf{x}_\alpha$ is an observation from $N(\mu, \Psi + \Lambda\Lambda')$. The sufficient statistics are $\bar{\mathbf{x}}$ and

$$\mathbf{C} = (1/n) \sum_{\alpha=1}^{n} (\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})'. \tag{11.1}$$

We require

$$\Lambda'\Psi^{-1}\Lambda = \Gamma = \text{Diagonal}. \tag{11.2}$$

Then the likelihood equations reduce to this constraint and the pair of equations

$$\text{Diagonal } \Psi = \text{Diagonal } (\mathbf{C} - \Lambda\Lambda'), \tag{11.3}$$

$$\mathbf{C}\Psi^{-1}\Lambda = \Lambda(\mathbf{I}_m + \Gamma). \tag{11.4}$$

Equation (11.4) can be written as

$$(\mathbf{C} - \Psi)\Psi^{-1}\Lambda = \Lambda\Gamma \tag{11.5}$$

or as

$$(\mathbf{C} - \Psi)(\Psi^{-1}\Lambda) = \Psi(\Psi^{-1}\Lambda)\Gamma. \tag{11.6}$$

The columns of $\hat{\Lambda}$ are characteristic vectors of $\mathbf{C}\Psi^{-1}$ or of $(\mathbf{C} - \Psi)\Psi^{-1}$. They correspond to the m largest characteristic roots.

It is of interest to compare this solution with those of Sections 5 ($\Psi = \sigma^2\mathbf{I}_p$) and 7 ($\Psi$ unknown). In the case of $\sigma^2\mathbf{I}_p$ we used the characteristic vectors and roots of $\mathbf{C}$. The matrix $\mathbf{W}_1$ satisfied

$$\mathbf{C}\mathbf{W}_1 = \mathbf{W}_1\mathbf{T}_1, \quad \mathbf{W}_1'\mathbf{W}_1 = \mathbf{I}_m. \tag{11.7}$$

The first equation of (11.7) can be written

$$\mathbf{C}(\sigma^2\mathbf{I}_p)^{-1}(\sigma\mathbf{W}_1\Gamma^{*1/2}) = (\sigma\mathbf{W}_1\Gamma^{*1/2})(\mathbf{I}_m + \Gamma^*), \tag{11.8}$$

where the diagonal matrix Γ^* is

$$\Gamma^* = (1/\sigma^2)\mathbf{T}_1 - \mathbf{I}_m. \tag{11.9}$$

The equation (11.8) is analogous to (11.4) with Ψ replaced by $\sigma^2\mathbf{I}_p$, Γ by Γ^*, and Λ by

$$\Lambda^* = \sigma\, \mathbf{W}_1\Gamma^{*1/2}. \tag{11.10}$$

Then (11.2) is equivalent to

$$\Lambda^{*\prime}(\sigma^2\mathbf{I}_p)^{-1}\Lambda^* = \Gamma^*. \tag{11.11}$$

The analog of (11.3) is

$$\text{tr } \Psi = \text{tr}(\mathbf{C} - \Lambda^*\Lambda^{*\prime}) \tag{11.12}$$

or

$$p\sigma^2 = \sum_{i=1}^{p} t_i - \text{tr } \mathbf{W}_1\sigma^2\Gamma^*\mathbf{W}_1' = \sum_{i=1}^{p} t_i - \sum_{i=1}^{m} t_i + m\sigma^2, \tag{11.13}$$

which yields $q\sigma^2 = \sum_{i=m+1}^{p} t_i$.

In the case of Ψ unknown studied in Section 7 the derivative equations can be manipulated to obtain

$$\tilde{\mathbf{H}}\Psi^{-1}\Lambda^{+} = \Lambda^{+}(\mathbf{I}_m + \Gamma^{+}), \tag{11.14}$$

$$\Psi = [1/(nk)](\mathbf{H} + \mathbf{G}) - \Lambda^{+}\Lambda^{+\prime}, \tag{11.15}$$

$$\Gamma^{+} = k\Lambda^{+\prime}\Psi^{-1}\Lambda^{+}, \tag{11.16}$$

where $\Theta = \Lambda^{+}\Lambda^{+\prime}$. These are analogous to (11.2), (11.3), and (11.4) with $\tilde{\mathbf{H}}$ as the analog of $\mathbf{C}$ in (11.4) and $[1/(nk)](\mathbf{H} + \mathbf{G})$ as the analog of $\mathbf{C}$ in (11.15). Then

$$\hat{\Lambda}^{+} = [1/(n\sqrt{k})](\mathbf{G} + \mathbf{H})\mathbf{Y}_1(\mathbf{D}_1 - \mathbf{I}_m)^{1/2}[\mathbf{D}_1 + (k-1)\mathbf{I}_m]^{-1} \tag{11.17}$$

if $\mathbf{D}_1 - \mathbf{I}_m$ is positive semidefinite. Ψ is then obtained from (11.15), which is equivalent to (7.7).

The solution of the likelihood equations (11.2), (11.3), and (11.14) is not straightforward because (11.3) and (11.4) have to be solved simultaneously. Lawley (1940) suggested using an initial estimate of Ψ and then solving (11.5) and (11.2) for Λ and Γ. In turn (11.3) can be solved for a new value of Ψ. Unfortunately, this method does not necessarily converge. Other computational methods must be used, such as the Fletcher-Powell algorithm. See Lawley and Maxwell (1971) for example. The methods maximize the likelihood function numerically.

A possible computational device is the EM (expectation-maximization) algorithm (Dempster, Laird, and Rubin (1977) and Rubin and Thayer (1982)). The idea is to treat the unobservable $\mathbf{f}_\alpha$'s as missing data. Under the assumption that the $\mathbf{f}_\alpha$'s and $\mathbf{u}_\alpha$'s have a joint normal distribution the sufficient statistics are the means and covariances of the $\mathbf{x}_\alpha$'s and $\mathbf{f}_\alpha$'s. The E step of the algorithm is to obtain the expectation of the covariances on the basis of trial values of the parameters. The M step is to maximize the likelihood function on the basis of these covariances; this step provides updated values of the parameters. The steps alternate, and the procedure may converge to the maximum likelihood estimators. Wu (1983) has discussed some questions of convergence of the EM algorithm.

The modern computer has revolutionized the estimation in factor anaysis. At the time of the summary by Anderson and Rubin (1956) the *centroid* method was the only method in use, but it has now disappeared. This crude procedure, based on addition and subtraction operations, can be considered as a rough approximation to the modification of principal component analysis (described below).

There are at least two serious problems with obtaining numerically the maximum likelihood estimators. One is that there might be several relative maxima. In that case a convergent computational method may lead to different solutions depending on the starting values. A protection against converging to a relative maximum that is not an absolute maximum is to use a grid of starting values to explore the likelihood function. However, if p is not small, the calculation at a grid of values may be prohibitively expensive.

Another problem is that a maximum may occur at a point at which one or more of the diagonal elements of Ψ are negative, the so-called Heywood case. To

avoid such a solution the program may require that each diagonal element be greater than some small constant or greater than some small multiple of the corresponding diagonal element of $\mathbf{C}$. If a computational solution yields a negative value for ψ_{ii} or a value equal to the prescribed minimum, the investigator may conclude that ψ_{ii} is 0. That implies that the ith test score is exactly a linear combination of factor scores. Then the factor scores can be transformed so that this linear combination is a new factor coordinate; thus, a test score is exactly a factor score. The interpretation is uncomfortable; a factor (a primary ability) is measured exactly by one test. A way to avoid negative estimates of the error variance ψ_{ii} is to write the likelihood in terms of the error standard deviations; estimates of 0 may still result.

Another approach to estimation is generalized least squares. If $\mathbf{\Sigma}_0$ were the true covariance matrix of the normally distributed $\mathbf{x}_\alpha$'s, then generalized least squares estimators would be the values of $\mathbf{\Psi}$ and $\mathbf{\Lambda}$ that minimize

$$\tfrac{1}{2}\operatorname{tr}\{\mathbf{\Sigma}_0^{-1}[\mathbf{C}-(\mathbf{\Psi}+\mathbf{\Lambda}\mathbf{\Lambda}')]\}^2. \tag{11.18}$$

This expression can be represented as the quadratic form

$$[\mathbf{c}-\sigma(\mathbf{\Psi},\mathbf{\Lambda})]'[\operatorname{cov}\mathbf{c}]^{-1}[\mathbf{c}-\sigma(\mathbf{\Psi},\mathbf{\Lambda})], \tag{11.19}$$

where $\mathbf{c}$ represents the elements of $\mathbf{C}$ arranged in a vector, $\sigma(\mathbf{\Psi},\mathbf{\Lambda})$ is $\mathbf{\Psi}+\mathbf{\Lambda}\mathbf{\Lambda}'$ arranged in a corresponding vector, and cov c is the covariance matrix of c under normality (Anderson, 1973). Jöreskog and Goldberger (1972) have proposed substituting $\mathbf{C}$ for $\mathbf{\Sigma}_0$ in (11.18). The estimators minimize

$$\tfrac{1}{2}\operatorname{tr}\{\mathbf{C}^{-1}[\mathbf{C}-(\mathbf{\Psi}+\mathbf{\Lambda}\mathbf{\Lambda}')]\}^2. \tag{11.20}$$

An alternative is to minimize

$$\tfrac{1}{2}\operatorname{tr}\{(\mathbf{\Psi}+\mathbf{\Lambda}\mathbf{\Lambda}')^{-1}[\mathbf{C}-(\mathbf{\Psi}+\mathbf{\Lambda}\mathbf{\Lambda}')]\}^2. \tag{11.21}$$

In either case the minimizing value of $\mathbf{\Lambda}$ for given $\mathbf{\Psi}$ is given by (11.4), the likelihood equation. Browne (1974) has shown that the GLS estimator of $\mathbf{\Psi}$ has the same asymptotic distribution as the maximum likelihood estimator. Dahm and Fuller (1981) have shown that if cov $\mathbf{c}$ in (11.17) is replaced by a matrix converging to cov $\mathbf{c}$, and $\mathbf{\Psi}$, $\mathbf{\Lambda}$ and $\mathbf{\Phi}$ depend on some parameters, then the asymptotic distributions of the resulting estimators are the same as for maximum likelihood estimation.

Let

$$\mathbf{\Theta}=\mathbf{\Psi}-\mathbf{\Lambda}(\mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{\Lambda})^{-1}\mathbf{\Lambda}', \tag{11.22}$$

which is of rank $p-m$. A necessary and sufficient condition that the transformation from $\mathbf{\Sigma}$ defined by (8.6) to $\mathbf{\Psi}$ and $\mathbf{\Lambda}$ (satisfying (9.1) with $\gamma_{11} > \cdots > \gamma_{mm}$) be single-valued and continuous in a neighborhood of $\mathbf{\Psi}$ and $\mathbf{\Lambda}$ is that the matrix with elements of θ_{ij}^2 be nonsingular. Then if $\sqrt{n}(\mathbf{C}-\mathbf{\Sigma})$ has a limiting normal distribution (for example, if the $\mathbf{x}_\alpha$'s are identically distributed and have second-order moments), the maximum likelihood estimators have an asymptotic normal distribution (Anderson and Rubin, 1956). Lawley (1953) found formulas

for the asymptotic covariances of the estimators of $\mathbf{\Lambda}$ when $\mathbf{\Psi}$ is known; he modified them (Lawley, 1967) for $\mathbf{\Psi}$ unknown (and (8.5) and (9.1) are satisfied), and these formulas were corrected by Jennrich and Thayer (1973). The results are too complicated to give here. The asymptotic covariances are simpler when $\mathbf{\Lambda}$ is identified by requiring a submatrix to be $\mathbf{I}_m$ (Fuller, Pantula and Amemiya, 1982).

The likelihood ratio criterion for testing the null hypothesis that $m = m_0$ against the alternative that $m > m_0$ is

$$\lambda = \Pi_{i=m_0+1}^{p} (1 + \hat{\gamma}_i)^{(1/2)n}, \tag{11.23}$$

where $\hat{\gamma}_{m+1}, \cdots, \hat{\gamma}_p$ are the q smallest roots of

$$|\mathbf{C} - \hat{\mathbf{\Psi}} - \gamma\hat{\mathbf{\Psi}}| = 0. \tag{11.24}$$

Under the null hypothesis and the condition that (θ_{ij}^2) is nonsingular, $-2 \log \lambda$ has a limiting χ^2-distribution with $\frac{1}{2}[(p - m)^2 - (p + m)]$ degrees of freedom. (Since (11.3) implies $0 = \text{tr}(\mathbf{C} - \hat{\mathbf{\Psi}})\hat{\mathbf{\Psi}}^{-1} = \text{tr}\, \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}'\hat{\mathbf{\Psi}}^{-1} = \text{tr}\, \hat{\mathbf{\Gamma}} = \sum_{i=1}^{m} \hat{\gamma}_i$, $-2 \log \lambda$ is approximately $(n/2) \sum_{i=m_0+1}^{p} \hat{\gamma}_i^2$).

12. Estimation of the factor structure in the case of fixed factors. Since maximum likelihood estimators do not exist in the case of fixed factors, what estimation methods can be used? One possibility is to use the maximum likelihood method appropriate for random factors. The theorem of Anderson and Rubin (1956) on asymptotic normality of the estimators holds here. Fuller, Pantula and Amemiya (1982) verified this theorem and found the asymptotic covariances when $\mathbf{\Lambda}$ is identified by a specified submatrix $\mathbf{I}_m$.

The sample covariance matrix under normality has the noncentral Wishart distribution (Anderson, 1946a) depending on $\mathbf{\Psi}$, $\mathbf{\Lambda\Phi\Lambda}'$, and $n - 1$. Anderson and Rubin (1956) proposed maximizing this likelihood function. However, one of the equations is difficult to solve. Again, the estimators are asymptotically equivalent to the maximum likelihood estimators for the random factor case. (Whittle, 1952, has discussed some special cases.)

13. Change of units of measurement. The factor analysis model can be set up for arbitrary units of measurement of the observable components. If units are changed, the model changes accordingly. For a diagonal matrix $\mathbf{D}$ with positive diagonal elements let

$$\mathbf{x}^* = \mathbf{Dx}, \quad \mathbf{\Lambda}^* = \mathbf{D\Lambda}, \quad \mu^* = \mathbf{D}\mu, \quad \mathbf{u}^* = \mathbf{Du}, \quad \mathbf{\Psi}^* = \mathbf{D\Psi D}. \tag{13.1}$$

Then the basic model becomes

$$\mathbf{x}^* = \mathbf{\Lambda}^*\mathbf{f} + \mu^* + \mathbf{u}^*, \tag{13.2}$$

where $\text{cov}(\mathbf{u}^*) = \mathbf{\Psi}^*$. The sample second-order moment matrix is transformed to $\mathbf{C}^* = \mathbf{DCD}$. The logarithm of the likelihood function is a constant plus a constant

times

$$-\log|\Psi^* + \Lambda^*\Lambda^{*\prime}| - \operatorname{tr} \mathbf{C}^*(\Psi^* + \Lambda^*\Lambda^{*\prime})^{-1} = -\log|\Psi + \Lambda\Lambda'| - \operatorname{tr} \mathbf{C}(\Psi + \Lambda\Lambda') - 2\log|\mathbf{D}|. \tag{13.3}$$

The maximum likelihood estimators of Λ^* and Ψ^* are

$$\hat{\Lambda}^* = \mathbf{D}\hat{\Lambda}, \quad \hat{\Psi}^* = \mathbf{D}\hat{\Psi}\mathbf{D}. \tag{13.4}$$

That is, the estimated factor loadings and error variances are merely changed by the units of measurement.

The fact that the factors do not depend on the location and scale factors is one reason for considering factor analysis as an analysis of interdependence. Another point of view (Howe, 1955) is that the partial correlations among the test scores are 0 given the factor scores; the estimation procedure based on minimizing the determinant of the sample partial correlations is equivalent to maximum likelihood for random factors. Rao (1955) has interpreted the relation between the test scores and factor scores in terms of canonical correlations. Anderson (1959) has set factor analysis in a more general model of conditional independence.

If $d_{ii} = 1/\sqrt{c_{ii}}$, $i = 1, \cdots, p$, then $\mathbf{C}^*$ is the sample correlation matrix. Computations can be done in these terms.

14. Relation to principal component analysis. What is the relation to principal component analysis proposed by Hotelling (1933)? As explained in Section 5, the vector of sample principal components is the vector $\mathbf{W}'\mathbf{x}$, where the columns of $\mathbf{W}$ are the characteristic vectors of $\mathbf{C}$ normalized by $\mathbf{W}'\mathbf{W} = \mathbf{I}_p$. Then

$$\mathbf{C} = \mathbf{W}\mathbf{T}\mathbf{W}' = \sum_{i=1}^{p} \mathbf{w}_i t_i \mathbf{w}_i'. \tag{14.1}$$

If $t_{m+1}, \cdots, t_p$ are small, $\mathbf{C}$ can be approximated by

$$\mathbf{W}_1\mathbf{T}_1\mathbf{W}_1' = \sum_{i=1}^{m} \mathbf{w}_i t_i \mathbf{w}_i', \tag{14.2}$$

and $\mathbf{x}$ is approximated by

$$\mathbf{W}_1\mathbf{W}_1'\mathbf{x} = \sum_{i=1}^{m} (\mathbf{w}_i'\mathbf{x})\mathbf{w}_i. \tag{14.3}$$

Then the sample covariance of the difference between $\mathbf{x}$ and the approximation (14.3) is the sample covariance of

$$\mathbf{x} - \mathbf{W}_1\mathbf{W}_1'\mathbf{x} = \mathbf{W}_2\mathbf{W}_2'\mathbf{x}, \tag{14.4}$$

which is $\mathbf{W}_2\mathbf{T}_2\mathbf{W}_2' = \sum_{i=m+1}^{p} \mathbf{w}_i t_i \mathbf{w}_i'$, and the sum of the variances of the components is $\sum_{i=m+1}^{p} t_i$.

This analysis is in terms of some common unit of measurement. The first m components "explain" a large proportion of the "variance" of $\mathbf{C}$. When the units of measurement are not the same (for example, when the units are arbitrary), it is customary to standardize each measurement to (sample) variance 1. However,

then the principal components do not have the above interpretation in terms of variance.

Another difference between principal component analysis and factor analysis is that the former does not separate the error from the systematic part. This fault is easily remedied, however. Thomson (1934) proposed the following estimation procedure for the factor analysis model. A diagonal matrix Ψ is subtracted from $\mathbf{C}$ and the principal component analysis is carried out on $\mathbf{C} - \Psi$. However, Ψ is determined so that $\mathbf{C} - \Psi$ is close to rank m. The equations are

$$(\mathbf{C} - \Psi)\Lambda = \Lambda\mathbf{L}, \tag{14.5}$$

$$\text{Diagonal}(\Psi + \Lambda\Lambda') = \text{Diagonal } \mathbf{C}, \tag{14.6}$$

$$\Lambda'\Lambda = \mathbf{L} = \text{Diagonal}. \tag{14.7}$$

The last equation is a normalization and takes out the indeterminacy in Λ. This method allows for the error terms, but still depends on the units of measurement. The estimators are consistent but not (asymptotically) efficient in the usual factor analysis model.

15. Identification by specified zeros. Simple structure as recommended by Thurstone is obtained by a linear transformation of a factor loading matrix, selected to make many loadings close to 0 (and satisfying several other conditions). We now consider the specification of 0 loadings in advance of the statistical analysis. The objectives are the identification of the parameters and the interpretation of the factors.

First we suppose that the covariance matrix of the factors, $\mathscr{E}\mathbf{f}_\alpha\mathbf{f}_\alpha' = \Phi$ is unrestricted. We shall now derive sufficient conditions for identification by preassigned 0's. Suppose at least $m - 1$ 0's are specified in the first column. The tests can be numbered so that the factor loading matrix has the form

$$\Lambda = \begin{pmatrix} \mathbf{0} & \Lambda^{+} \\ \lambda & \Lambda^{++} \end{pmatrix}, \tag{15.1}$$

where $\mathbf{0}$ in the first column has at least $m - 1$ elements. Furthermore, the matrix Λ^{+} must be of rank $m - 1$. Then multiplication of (15.1) on the right by a nonsingular matrix $\mathbf{A}$ yields

$$\Lambda\mathbf{A} = \begin{pmatrix} \mathbf{0} & \Lambda^{+} \\ \lambda & \Lambda^{++} \end{pmatrix}\begin{pmatrix} a_{11} & \mathbf{a}_{12} \\ \mathbf{a}_{21} & \mathbf{A}_{22} \end{pmatrix} = \begin{pmatrix} \Lambda^{+}\mathbf{a}_{21} & — \\ \lambda a_{11} + \Lambda^{++}\mathbf{a}_{21} & — \end{pmatrix}. \tag{15.2}$$

Since $\Lambda\mathbf{A}$ must satisfy the same zero restrictions as Λ, $\Lambda^{+}\mathbf{a}_{21} = \mathbf{0}$; if the rank of Λ^{+} is $m - 1$, then $\mathbf{a}_{21} = \mathbf{0}$ and the rest of the first column is λa_{11}, which is proportional to the original λ. Thus, a sufficient condition for identification by 0's is that each column has at least $m - 1$ specified 0's and the matrix composed of the rows of Λ corresponding to the 0's in a certain column has rank $m - 1$. Koopmans and Reiersøl (1950), Reiersøl (1950), and Anderson and Rubin (1956) adapted these conditions from similar identification conditions for simultaneous equation models in econometrics (Section 18). Howe (1955) and Lawley (1958)

treated analysis based on models with specified 0 restrictions; Jöreskog (1969) termed such analysis "confirmatory factor analysis."

Each component factor score needs to have its unit or scale determined. That can be done by requiring an element in each column of Λ to be 1. (Requiring the submatrix of Λ consisting of the first m rows to be $\mathbf{I}_m$ is one way of imposing the conditions.) An alternative scaling is to require $\phi_{ii} = 1$, $i = 1, \cdots, m$; this makes Φ a correlation matrix.

The derivatives of the likelihood function set equal to 0 are

$$\text{Diagonal } \Sigma^{-1}[\mathbf{C} - (\Psi + \Lambda\Phi\Lambda')]\,\Sigma^{-1} = \text{Diagonal } \mathbf{0}, \tag{15.3}$$

$$\Lambda'\Sigma^{-1}[\mathbf{C} - (\Psi + \Lambda\Phi\Lambda')]\Sigma^{-1}\Lambda = \mathbf{0} \tag{15.4}$$

for positions in Φ that are not specified and

$$\Sigma^{-1}[\mathbf{C} - (\Psi + \Lambda\Phi\Lambda')]\Sigma^{-1}\Lambda\Phi = \mathbf{0} \tag{15.5}$$

for positions in Λ not specified, where

$$\Sigma = \Psi + \Lambda\Phi\Lambda'. \tag{15.6}$$

These equations cannot be simplified as in Section 11 because (15.5) holds only for the unspecified positions in Λ, and hence one cannot multiply by Σ on the left. These equations are not useful for computation, but the likelihood function can be maximized numerically. Howe (1955) and Anderson and Rubin (1956) found alternative expressions for the likelihood equations. The latter gave very general conditions for the asymptotic normality of the estimators.

Now we turn to the case of orthogonal factors. If $\mathscr{E}\mathbf{ff}' = \Phi = \mathbf{I}_m$, we need at least $j - 1$ 0's specified in the jth column of Λ, $j = 2, \cdots, m$ (after suitably renumbering the factors). For example, if $m = 3$, we could have the pattern

$$\Lambda = \begin{pmatrix} \lambda_{11} & 0 & 0 \\ \lambda_{21} & \lambda_{22} & 0 \\ \lambda_{31} & \lambda_{32} & \lambda_{33} \\ \vdots & \vdots & \vdots \\ \lambda_{p1} & \lambda_{p_2} & \lambda_{p3} \end{pmatrix}. \tag{15.7}$$

Multiplication of (15.1) on the right by any nontrivial orthogonal matrix will change the pattern. Anderson and Rubin (1956) also proposed similar restrictions on $\mathbf{A}\Lambda$ for some preassigned $\mathbf{A}$.

16. Orthogonal vs. oblique factors. If Φ is specified to be diagonal, the factors are said to be *orthogonal.* The components are uncorrelated in the population or sample according to whether the factors are considered as random or as fixed. If Φ is not necessarily diagonal, the factors are said to be *oblique.* The idea of uncorrelated factor scores has appeal.

Some psychologists claim that the orthogonality of the factor scores is essential if one is to consider the factor scores as more basic than the test scores. Considerable debate has gone on among psychologists concerning this point. On

the other side, Thurstone (1947), page vii, has said "it seems just as unnecessary to require that mental traits shall be uncorrelated in the general population as to require that height and weight be uncorrelated in the general population."

As we have seen, given a pair of matrices Λ, Φ, equivalent pairs are given by $\Lambda\mathbf{C}$, $\mathbf{C}^{-1}\Phi(\mathbf{C}^{-1})'$ for nonsingular $\mathbf{C}$'s. The pair selected (based on $\mathbf{C}$ given Λ, Φ) should be the one with the most meaningful interpretation in terms of the subject matter of the tests. The idea of simple structure is that with 0 factor loadings in certain patterns the component factor scores can be given meaning regardless of Φ. Permitting Φ to be an arbitrary positive definite matrix allows more 0's in Λ.

Another consideration in relating transformations or identification conditions is autonomy or permanence or invariance with regard to certain changes. For example, what happens if a selection of the constituents of a population is made? In case of intelligence tests suppose a selection is made, such as college admittees out of high school students, that can be assumed to involve the primary abilities. One can envisage that the relation between unobserved factor scores $\mathbf{f}$ and observed test scores $\mathbf{x}$ is unaffected by the selection, that is, that the matrix of factor loadings Λ is unchanged. The variances of the errors (and specific factors), the diagonal elements of Ψ, may also be considered as unchanged by the selection because the errors are uncorrelated with the factors (primary abilities).

Suppose there is a "true" model, Λ, Φ, Ψ, and the investigator applies identification conditions that permit him to discover it. Next, suppose there is a selection that results in a new population of factor scores so that their covariance matrix is Φ^*. When does the investigator analyze the new observed covariance matrix $\Psi + \Lambda\Phi^*\Lambda'$ to obtain the original Λ again? If part of the identification conditions are that the factor moment matrix is $\mathbf{I}$, then he will obtain a different factor loading matrix. On the other hand, if the identification conditions are entirely on the factor loadings (specified 0's and 1's) the factor loadings matrix from the analysis is the same as before.

The same consideration is relevant in comparing two populations. It may be reasonable to consider that $\Psi_1 = \Psi_2$, $\Lambda_1 = \Lambda_2$, but $\Phi_1 \neq \Phi_2$. To test the hypothesis that $\Phi_1 = \Phi_2$ one wants to use identification conditions that agree with $\Lambda_1 = \Lambda_2$ (rather than $\Lambda_1 = \Lambda_2\mathbf{C}$). The conditions should be on the factor loadings.

What happens if more tests are added (or deleted)? In addition to observing $\mathbf{X} = \Lambda\mathbf{f} + \mu + \mathbf{U}$, suppose one observes $\mathbf{X}^* = \Lambda^*\mathbf{f} + \mu^* + \mathbf{U}^*$, where $\mathbf{U}^*$ is uncorrelated with $\mathbf{U}$. Since the common factors $\mathbf{f}$ are unchanged, Φ is unchanged. However, the (arbitrary) condition $\Lambda'\Psi^{-1}\Lambda$ being diagonal is changed; use of this type of condition would lead to a rotation of Λ.

Part III. Simultaneous equations models.

17. Simultaneous equations models in econometrics. Simultaneous equations models (or structural equation models) are used to describe the behaviors of sets of economic agents. A familiar example is the pair of demand and supply schedules (as displayed in Figure 5):

(17.1) Demand: α quantity + β price + γ consumer income + δ = random,

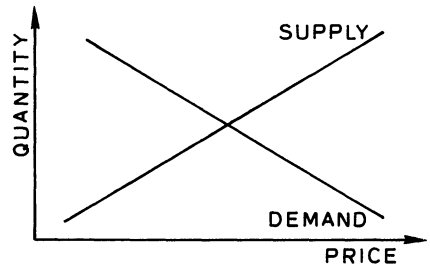

FIG. 5. *Supply and demand schedules.*

(17.2) Supply: λ quantity + ν price + μ cost of raw materials + ϕ = random.

This model indicates that the quantity of goods desired by consumers is a linear function of its price, consumer income and a random term, and the quantity of goods offered by producers is a linear function of the price, cost of raw materials and a random term. Given consumers' income, cost of raw material, and the realizations of the random terms the price and quantity in the market are the intersection of the two lines (solution of the two equations).

Now let us turn to the general model. (In Part III the notation for simultaneous equations is a standard notation of econometrics; unfortunately, it conflicts with the notation of Parts I and II.) The equations are

$$\mathbf{B}\mathbf{y}_t + \Gamma\mathbf{z}_t = \mathbf{u}_t, \quad t = 1, \cdots, T, \tag{17.3}$$

where $\mathbf{B}$ is $G \times G$, Γ is $G \times K$, $\mathbf{y}_t$ is composed of G jointly dependent variables (endogenous), $\mathbf{z}_t$ is composed of K predetermined variables (exogenous and lagged dependent) which are treated as "independent" variables, and $\mathbf{u}_t$ consists of G unobservable random variables with

$$\mathscr{E}\mathbf{u}_t = \mathbf{0}, \quad \mathscr{E}\mathbf{u}_t\mathbf{u}_t' = \mathbf{\Sigma}. \tag{17.4}$$

We require $\mathbf{B}$ to be nonsingular. This model was initiated by Haavelmo (1944) and was developed by Koopmans, Marschak, Hurwicz, Anderson, Rubin, and Leipnik, 1944–1954, at the Cowles Commission for Research in Economics. Each component equation represents the behavior of some group (such as consumers or producers) and has economic meaning.

The set of simultaneous equations (17.3) can be solved for $\mathbf{y}_t$ (because $\mathbf{B}$ is nonsingular):

$$\mathbf{y}_t = \Pi\mathbf{z}_t + \mathbf{v}_t, \tag{17.5}$$

where

$$\Pi = -\mathbf{B}^{-1}\Gamma, \quad \mathbf{v}_t = \mathbf{B}^{-1}\mathbf{u}_t \tag{17.6}$$

with

$$\mathscr{E}\mathbf{v}_t = 0, \quad \mathscr{E}\mathbf{v}_t\mathbf{v}_t' = \mathbf{B}^{-1}\mathbf{\Sigma}(\mathbf{B}')^{-1} = \Omega, \tag{17.7}$$

say. The equation (17.5) is called the *reduced form* of the model. It is a multivar-

iate regression model. In principle it can be determined from the observable $\mathbf{y}_t$, $\mathbf{z}_t$.

18. Identification by specified zeros. The set of simultaneous equations (17.3) can be multiplied on the left by an arbitrarily nonsingular matrix. To determine component equations that are economically meaningful, restrictions must be imposed. In the case of the example of demand and supply the equation describing demand is distinguished by the fact that it includes consumer income and excludes cost of raw materials. The exclusion of the latter amounts to specifying that its coefficient in the demand equation is 0.

We consider identification of an equation by specifying certain coefficients to be 0. It is convenient to treat the first equation. Suppose the variables are numbered so that the first G_1 jointly dependent variables are included in the first equation and the remaining $G_2 = G - G_1$ are not and the first K_1 predetermined variables are included and $K_2 = K - K_1$ are excluded. Then we can partition the coefficient matrices as

$$(\mathbf{B}\ \boldsymbol{\Gamma}) = \begin{pmatrix} \beta' & \mathbf{0} & \gamma' & \mathbf{0} \\ — & — & — & — \end{pmatrix}, \tag{18.1}$$

where β', $\mathbf{0}$, γ', and $\mathbf{0}$ are row vectors with G_1, G_2, K_1, and K_2 components, respectively. The coefficient matrix of the reduced form is partitioned conformably into G_1 and G_2 rows and K_1 and K_2 columns:

$$\boldsymbol{\Pi} = \begin{pmatrix} \boldsymbol{\Pi}_{11} & \boldsymbol{\Pi}_{12} \\ — & — \end{pmatrix}. \tag{18.2}$$

The relation between $\mathbf{B}$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\Pi}$ can be expressed

$$\begin{pmatrix} \gamma' & \mathbf{0} \\ — & — \end{pmatrix} = \boldsymbol{\Gamma} = -\mathbf{B}\boldsymbol{\Pi} = -\begin{pmatrix} \beta' & \mathbf{0} \\ — & — \end{pmatrix}\begin{pmatrix} \boldsymbol{\Pi}_{11} & \boldsymbol{\Pi}_{12} \\ — & — \end{pmatrix} = -\begin{pmatrix} \beta'\boldsymbol{\Pi}_{11} & \beta'\boldsymbol{\Pi}_{12} \\ — & — \end{pmatrix}. \tag{18.3}$$

The upper row of (18.3) yields

$$\beta'\boldsymbol{\Pi}_{12} = \mathbf{0},\ -\beta'\boldsymbol{\Pi}_{11} = \gamma' \tag{18.4}$$

To determine β ($G_1 \times 1$) uniquely except for a constant of proportionality we need

$$\operatorname{rank}(\boldsymbol{\Pi}_{12}) = G_1 - 1. \tag{18.5}$$

This implies

$$K_2 \geq G_1 - 1. \tag{18.6}$$

Addition of G_2 to (18.6) gives the *order condition*

$$G_2 + K_2 \geq G_1 + G_2 - 1 = G - 1. \tag{18.7}$$

The number of specified 0's in an identified equation must equal or exceed the number of equations (or jointly dependent variables) less 1.

It can be shown that when $\mathbf{B}$ is nonsingular (18.5) holds if and only if the

number of specified 0's in the first row is at least $G - 1$ and the rank of the matrix consisting of the columns of $(\mathbf{B}\ \Gamma)$ with specified 0's in the first row is $G - 1$. The condition for a row of $(\mathbf{B}\ \Gamma)$ to be identified (except for a constant of proportionality) corresponds to the condition for a column of Λ to be identified.

19. Estimation of the reduced form. The model (17.5) is a typical multivariate regression model. The observations are

$$\begin{pmatrix}\mathbf{y}_1\\ \mathbf{z}_1\end{pmatrix}, \cdots, \begin{pmatrix}\mathbf{y}_T\\ \mathbf{z}_T\end{pmatrix}. \tag{19.1}$$

The usual estimators of Π and Ω are

$$\mathbf{P} = \textstyle\sum_{t=1}^{T} \mathbf{y}_t\mathbf{z}_t'(\sum_{t=1}^{T} \mathbf{z}_t\mathbf{z}_t')^{-1}, \tag{19.2}$$

$$\hat{\Omega} = (1/T)\ \textstyle\sum_{t=1}^{T} (\mathbf{y}_t - \mathbf{P}\mathbf{z}_t)(\mathbf{y}_t - \mathbf{P}\mathbf{z}_t)'. \tag{19.3}$$

These are maximum likelihood estimators if the $\mathbf{v}_t$ in (17.5) are normally distributed and Π is unrestricted.

If the $\mathbf{z}_t$ are exogenous (regardless of normality), then

$$\mathscr{E}\ \text{vec}\ \mathbf{P} = \text{vec}\ \Pi, \quad \text{cov}(\text{vec}\ \mathbf{P}) = \mathbf{A}^{-1} \otimes \Omega, \tag{19.4}$$

where

$$\mathbf{A} = \textstyle\sum_{t=1}^{T} \mathbf{z}_t\mathbf{z}_t', \tag{19.5}$$

and $\text{vec}(\mathbf{d}_1, \cdots, \mathbf{d}_m) = (\mathbf{d}_1', \cdots, \mathbf{d}_m')'$, and $\otimes$ denotes the Kronecker product.

Futhermore, if the $\mathbf{v}_t$ are normal, then $\mathbf{P}$ is normal and $T\hat{\Omega}$ has the Wishart distribution with covariance matrix Ω and $T - K$ degrees of freedom.

20. Estimation of the structural coefficients. First, consider the estimation of the vector of coefficients β when $K_2 = G_1 - 1$. Let $\mathbf{P}$ be partitioned according to Π:

$$\mathbf{P} = \begin{pmatrix}\mathbf{P}_{11} & \mathbf{P}_{12}\\ \mathbf{P}_{21} & \mathbf{P}_{22}\end{pmatrix}. \tag{20.1}$$

Then the probability is 1 that rank of $\mathbf{P}_{12}$ is $G_1 - 1$ and the equation

$$\hat{\beta}'\mathbf{P}_{12} = \mathbf{0} \tag{20.2}$$

has a nontrivial solution that is unique except for a constant of proportionality. Any solution is a maximum likelihood estimator when the disturbance terms are normally distributed.

If $K_2 \geq G_1$, then the probability is 1 that the rank of $\mathbf{P}_{12}$ is G_1 and the equation (20.2) has only the trivial solution $\hat{\beta} = \mathbf{0}$, which is unsatisfactory. To obtain a suitable estimator we find $\hat{\beta}$ to minimize $\hat{\beta}'\mathbf{P}_{12}$ in a sense to be described later.

Let

$$\mathbf{z}_t = \begin{pmatrix}\mathbf{z}_t^{(1)}\\ \mathbf{z}_t^{(2)}\end{pmatrix}, \tag{20.3}$$

$$\sum_{t=1}^{T} \mathbf{z}_t\mathbf{z}_t' = \mathbf{A} = \begin{pmatrix} \mathbf{A}_{12} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}, \tag{20.4}$$

$$\mathbf{A}_{22.1} = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}, \tag{20.5}$$

$$\mathbf{y}_t = \begin{pmatrix} \mathbf{y}_t^{(1)} \\ \mathbf{y}_t^{(2)} \end{pmatrix}, \tag{20.6}$$

$$\Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}, \tag{20.7}$$

where $\mathbf{z}_t^{(1)}$ and $\mathbf{z}_t^{(2)}$ have K_1 and K_2 components, respectively, and $\mathbf{y}_t^{(1)}$ and $\mathbf{y}_t^{(2)}$ have G_1 and G_2 components, respectively. Now set up the multivariate analysis of variance table for $\mathbf{y}_t^{(1)}$. The first term in Table 1 is the (vector) sum of squares of $\mathbf{y}_t^{(1)}$ due to the effect of $\mathbf{z}_t^{(1)}$. The second term is due to the effect of $\mathbf{z}_t^{(2)}$ beyond the effect of $\mathbf{z}_t^{(1)}$. The two add to $(\mathbf{PAP}')_{11}$, which is the total effect of $\mathbf{z}_t$, the predetermined variables.

We propose to find the vector $\hat{\beta}$ such that the effect of $\mathbf{z}_t^{(2)}$ on $\hat{\beta}'\mathbf{y}_t^{(1)}$ beyond the effect of $\mathbf{z}_t^{(1)}$ is minimized relative to the error sum of squares of $\hat{\beta}'y_t^{(1)}$. We minimize

$$\frac{\hat{\beta}'(\mathbf{P}_{12}\mathbf{A}_{22.1}\mathbf{P}_{12}')\hat{\beta}}{\hat{\beta}'\hat{\Omega}_{11}\hat{\beta}} = \frac{(\hat{\beta}'\mathbf{P}_{12})\mathbf{A}_{22.1}(\hat{\beta}'\mathbf{P}_{12})'}{\hat{\beta}'\hat{\Omega}_{11}\hat{\beta}}. \tag{20.8}$$

This estimator has been called the *least variance ratio* estimator. Under normality and based only on the zero restrictions on the coefficients of this single equation, the estimator is maximum likelihood and is known as the *Limited Information Maximum Likelihood* (LIML) estimator (Anderson and Rubin, 1949).

The algebra of minimizing (20.8) is to find the smallest root, say ν, of

$$|\mathbf{P}_{12}\mathbf{A}_{22.1}\mathbf{P}_{12}' - \lambda\hat{\Omega}_{11}| = 0 \tag{20.9}$$

and the corresponding vector satisfying

$$\mathbf{P}_{12}\mathbf{A}_{22.1}\mathbf{P}_{12}'\hat{\beta} = \nu\hat{\Omega}_{11}\hat{\beta}. \tag{20.10}$$

The vector is normalized according to some rule. A frequently used rule is to set one (nonzero) coefficient equal to 1, for example, $\hat{\beta}_1 = 1$. If we write

$$\beta = \begin{pmatrix} 1 \\ -\beta_* \end{pmatrix}, \quad \hat{\beta} = \begin{pmatrix} 1 \\ -\hat{\beta}_* \end{pmatrix}, \tag{20.11}$$

TABLE 1
Multivariate Analysis of Variance

Source	Sum of Squares
$\mathbf{z}_t^{(1)}$	$\sum_{t=1}^{T} \mathbf{y}_t^{(1)}\mathbf{z}_t^{(1)\prime}\mathbf{A}_{11}^{-1}\sum_{t=1}^{T} \mathbf{z}_t^{(1)}\mathbf{y}_t^{(1)\prime}$
$\mathbf{z}_t^{(2)} \perp \mathbf{z}_t^{(1)}$	$\mathbf{P}_{12}\mathbf{A}_{22.1}\mathbf{P}_{12}'$
Error	$\sum_{t=1}^{T} (\mathbf{y}_t^{(1)} - \mathbf{P}_{11}\mathbf{z}_t^{(1)} - \mathbf{P}_{12}\mathbf{z}_t^{(2)})(\mathbf{y}_t^{(1)} - \mathbf{P}_{11}\mathbf{z}_t^{(1)} - \mathbf{P}_{12}\mathbf{z}_t^{(2)})' = \mathbf{T}\hat{\Omega}_{11}$
Total	$\sum_{t=1}^{T} \mathbf{y}_t^{(1)}\mathbf{y}_t^{(1)\prime}$

$$\mathbf{\Pi}_{12} = \begin{pmatrix} \boldsymbol{\pi}_{12} \\ \mathbf{\Pi}^*_{12} \end{pmatrix}, \quad \mathbf{P}_{12} = \begin{pmatrix} \mathbf{p}_{12} \\ \mathbf{P}^*_{12} \end{pmatrix}, \tag{20.12}$$

$$\hat{\mathbf{\Omega}}_{11} = \begin{pmatrix} \hat{\omega}_{11} & \hat{\boldsymbol{\omega}}^{*\prime}_{11} \\ \hat{\boldsymbol{\omega}}^*_{11} & \mathbf{\Omega}^*_{11} \end{pmatrix}, \tag{20.13}$$

then (20.10) can be replaced by the linear equation

$$(\mathbf{P}^*_{12}\mathbf{A}_{22.1}\mathbf{P}^{*\prime}_{12} - \nu\hat{\mathbf{\Omega}}^*_{11})\hat{\boldsymbol{\beta}}^* = (\mathbf{P}^*_{12}\mathbf{A}_{22.1}\mathbf{p}'_{12} - \nu\hat{\boldsymbol{\omega}}^*_{11}). \tag{20.14}$$

The first component equation in (20.10) has been dropped because it is linearly dependent on the other equations (because ν is a root of (20.9)).

21. Relation to the linear functional relationship. We now show that the model for the single linear functional relationship ($q = 1$) is identical to the model of simultaneous equations in the special case that $G_2 = 0$ (that is, $\mathbf{y}^{(1)}_t = \mathbf{y}_t$) and $z^{(1)}_t \equiv 1$ ($K_1 = 1$). Write the two models as

$$\mathbf{x}_{\alpha j} = \boldsymbol{\mu} + \boldsymbol{\nu}_\alpha + \mathbf{u}_{\alpha j}, \quad \alpha = 1, \cdots, n, \quad j = 1, \cdots, k, \tag{21.1}$$

where

$$\textstyle\sum_{\alpha=1}^{n} \boldsymbol{\nu}_\alpha = \mathbf{0}, \tag{21.2}$$

and

$$\mathbf{y}_t = \mathbf{\Pi}_1\mathbf{z}^{(1)}_t + \mathbf{\Pi}_2\mathbf{z}^{(2)}_t + \mathbf{v}_t, \quad t = 1, \cdots, T, \tag{21.3}$$

where $\mathbf{\Pi} = (\mathbf{\Pi}_1\ \mathbf{\Pi}_2)$. The correspondence between the models is $p \leftrightarrow G = G_1$,

$$\mathbf{x}_{\alpha j} \leftrightarrow \mathbf{y}_t, \quad \mathbf{u}_{\alpha j} \leftrightarrow \mathbf{v}_t, \tag{21.4}$$

$$(\alpha, j) \leftrightarrow t, \quad nk \leftrightarrow T, \tag{21.5}$$

$$\mathbf{\Psi} \leftrightarrow \mathbf{\Omega}. \tag{21.6}$$

We can rewrite the model (21.1) for the linear functional relationship using dummy variables (Anderson, 1958, Section 8.9). Define

$$\mathbf{s}_{\alpha j} = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \leftarrow \alpha\text{th position}, \quad \alpha = 1, \cdots, n-1, \tag{21.7}$$

$$\mathbf{s}_{nj} = \begin{bmatrix} -1 \\ \vdots \\ -1 \end{bmatrix}. \tag{21.8}$$

Then

$$\mu + \nu_\alpha = (\mu, \nu_1, \cdots, \nu_{n-1})\begin{pmatrix} 1 \\ \mathbf{s}_{\alpha j} \end{pmatrix}, \quad \alpha = 1, \cdots, n, \tag{21.9}$$

where j is suppressed on the left. Note that

$$\nu_n = -(\nu_1 + \cdots + \nu_{n-1}). \tag{21.10}$$

The correspondence between the models is

$$1 \leftrightarrow z_t^{(1)}, \quad \mathbf{s}_{\alpha j} \leftrightarrow \mathbf{z}_t^{(2)}, \tag{21.11}$$

$$\mu \leftrightarrow \mathbf{\Pi}_1, \quad (\nu_1, \cdots, \nu_{n-1}) \leftrightarrow \mathbf{\Pi}_2, \tag{21.12}$$

$$1 \leftrightarrow K_1, \quad n - 1 \leftrightarrow K_2, \tag{21.13}$$

$$\mathbf{B}(\nu_1, \cdots, \nu_{n-1}) = \mathbf{0} \leftrightarrow \beta'\mathbf{\Pi}_2 = \mathbf{0}. \tag{21.14}$$

Let $\mathbf{P} = (\mathbf{P}_1\ \mathbf{P}_2)$. In terms of the observable statistics in the two models, we have the correspondence

$$\hat{\mu} = \bar{\mathbf{x}} \leftrightarrow \bar{\mathbf{y}}, \tag{21.15}$$

$$\bar{\mathbf{x}}_\alpha - \bar{\mathbf{x}} \leftrightarrow \mathbf{P}_2. \tag{21.16}$$

Effect Matrix

$$\mathbf{H} = k \textstyle\sum_{\alpha=1}^{n} (\bar{\mathbf{x}}_\alpha - \bar{\mathbf{x}})(\bar{\mathbf{x}}_\alpha - \bar{\mathbf{x}})' \leftrightarrow \mathbf{P}_2\mathbf{A}_{22.1}\mathbf{P}_2' \tag{21.17}$$

Error Matrix

$$\mathbf{G} = \textstyle\sum_{\alpha=1}^{n} \sum_{j=1}^{k} (\mathbf{x}_{\alpha j} - \bar{\mathbf{x}}_\alpha)(\mathbf{x}_{\alpha j} - \bar{\mathbf{x}}_\alpha)' \leftrightarrow T\hat{\Omega} = \textstyle\sum_{t=1}^{T} (\mathbf{y}_t - \mathbf{P}\mathbf{z}_t)(\mathbf{y}_t - \mathbf{P}\mathbf{z}_t)'. \tag{21.18}$$

Then the estimator $\hat{\mathbf{B}}$ of the linear functional relationship for $q = 1$ is identical to the LIML estimator (Anderson, 1951a, 1976).

The correspondence was extended to general q by Anderson (1951a). Suppose assigned 0's form a pattern (after possibly renumbering equations and variables) so that the matrix of coefficients of the simultaneous equations can be partitioned as

$$(\mathbf{B}\mathbf{\Gamma}) = \left(\begin{array}{cccc} \mathbf{B}_1 & \mathbf{0} & \mathbf{\Gamma}_1 & \mathbf{0} \\ — & — & — & — \end{array}\right), \tag{21.19}$$

where $\mathbf{B}_1$ is $q \times G_1$ and $\mathbf{\Gamma}_1$ is $q \times K_1$ and the reduced form is partitioned as (18.2). The matrix $\mathbf{\Pi}_{12}$ should be of rank $G_1 - q$. Then from the generalization from (18.3) we obtain

$$\mathbf{B}_1\mathbf{\Pi}_{12} = \mathbf{0}, \quad \mathbf{B}_1\mathbf{\Pi}_{11} = -\mathbf{\Gamma}_1. \tag{21.20}$$

The first equation in (21.20) corresponds to $\mathbf{B}(\nu_1, \cdots, \nu_{n-1}) = \mathbf{0}$. To eliminate the indeterminacy of the multiplication of (21.20) on the left by an arbitrary nonsingular matrix, $q - 1$ 0's need to be specified in each row of $(\mathbf{B}_1\ \mathbf{\Gamma}_1)$. Hannan (1967) and Chow and Ray-Chaudhuri (1967) repeated some of Anderson's analysis.

It is seen from the preceding discussion that the regression matrix $\mathbf{\Pi}$ is estimated so that a submatrix has a specified rank. This procedure has been called "reduced rank regression." (See Izenman, 1975.) As a matter of fact, Anderson (1951a) obtained the maximum likelihood estimators of a regression submatrix and a matrix of restrictions; the estimators of the linear functional relations in Section 6 were obtained by use of dummy variables as above.

22. Shocks vs. errors. The unobserved random term $\mathbf{u}_t$ in (17.3) is known as a disturbance in equations or a shock. The shock is a stochastic input to the economic system. An alternative probabilistic model is in terms of errors of measurement. Linear relationships among the true values might be written

$$\mathbf{B}\boldsymbol{\eta}_t + \mathbf{\Gamma}\boldsymbol{\zeta}_t = \mathbf{0}. \tag{22.1}$$

The observed variables are

$$\mathbf{y}_t = \boldsymbol{\eta}_t + \mathbf{u}_t^*, \quad \mathbf{z}_t = \boldsymbol{\zeta}_t + \mathbf{v}_t^*, \tag{22.2}$$

where $\mathbf{u}_t^*$ and $\mathbf{v}_t^*$ are random vectors. Substitution into (22.1) yields

$$\mathbf{B}\mathbf{y}_t + \mathbf{\Gamma}\mathbf{z}_t = \mathbf{B}\mathbf{u}_t^* + \mathbf{\Gamma}\mathbf{v}_t^*. \tag{22.3}$$

If $\mathbf{v}_t^* = \mathbf{0}$ with probability 1, the model defined by (22.3) cannot be distinguished from (17.3), but otherwise the model is different. In (22.3) the random term on the right hand side is then correlated with the $\mathbf{z}_t$ term on the left hand side, and the "reduced form" obtained from (22.3) does not have the properties of multiple regression.

The interpretation of the random terms $\mathbf{u}_t^*$ and $\mathbf{v}_t^*$ is that they represent errors of measurement, sometimes called errors in variables. The properties of these models and their use in economics was studied by econometricians, such as Frisch (1934), Reiersøl (1945), Tintner (1945), (1946), (1952), and Geary (1942), (1948). Haavelmo's pathbreaking papers (1943), (1944) introduced the simultaneous equations with disturbances. This approach quickly dominated econometrics; see, for example, Koopmans (1950).

It is possible to construct models with both shocks and errors, but special conditions are needed to effect identification. See, for example, Anderson and Hurwicz (1947), Goldberger (1974), Geraci (1976), and Hsiao (1976).

23. Asymptotic theory as $T \to \infty$. The LIML estimator has an asymptotic normal distribution as the length of the observation series increases. We assume that $(1/T)\mathbf{A}$ approaches a positive definite limit. (If some predetermined variables are lagged endogenous variables, this limit must hold with probability 1). We partition β and $\mathbf{\Pi}_{12}$ as in (20.11) and (20.12). Let

$$\sigma^2 = \beta'\mathbf{\Omega}_{11}\beta = \text{Var } u_{1t}. \tag{23.1}$$

Then

$$(1/\sigma)(\mathbf{\Pi}_{12}^*\mathbf{A}_{22.1}\mathbf{\Pi}_{12}^{*\prime})^{1/2}(\hat{\beta}^* - \beta^*) \to_{\mathcal{L}} N(\mathbf{0}, \mathbf{I}_{G_1-1}). \tag{23.2}$$

We consider $\hat{\beta}^*$ as approximately distributed according to

$$N[\beta^*, \sigma^2(\Pi_{12}^* \mathbf{A}_{22.1} \Pi_{12}^{*\prime})^{-1}] \tag{23.3}$$

(Anderson and Rubin, 1950). The covariance matrix in (23.3) can be estimated consistently.

Because of the correspondence between the LIML estimator and the MLE for the linear functional relationship as outlined in Section 21, this asymptotic theory can be translated for the latter. Suppose the single linear functional relationship is written as

$$0 = \beta' \nu_\alpha = (1 - \beta^{*\prime}) \begin{pmatrix} \nu_{1\alpha} \\ \nu_\alpha^* \end{pmatrix} = \nu_{1\alpha} - \beta^{*\prime} \nu_\alpha^*, \quad \alpha = 1, \cdots, n, \tag{23.4}$$

where

$$\nu_\alpha = \begin{pmatrix} \nu_{1\alpha} \\ \nu_\alpha^* \end{pmatrix}, \quad \alpha = 1, \cdots, n. \tag{23.5}$$

Let n ($\leftrightarrow K$) be fixed and let the number of replications $k \to \infty$ (corresponding to $T/K \to \infty$ for fixed K). Let $\sigma^2 = \beta' \Psi \beta$.

Since $\Pi_{12} \mathbf{A}_{22.1} \Pi_{12}'$ corresponds to $k \sum_{\alpha=1}^n \nu_\alpha \nu_\alpha'$, $\hat{\beta}^*$ here has the approximate distribution

$$N[\beta^*, \sigma^2(k \textstyle\sum_{\alpha=1}^n \nu_\alpha^* \nu_\alpha^{*\prime})^{-1}]. \tag{23.6}$$

24. Two stage least squares. The two stage least squares (TSLS) estimator can be considered as a simplification of the LIML estimator. In (20.13) $\hat{\Omega}_{11}$ is a consistent estimator of Ω_{11}, ν is of probability order $O_p(1)$, and $\mathbf{P}_{12} \mathbf{A}_{22.1} \mathbf{P}_{12}'$ is of probability order $O_p(T)$. Thus, the term $\nu \hat{\Omega}_{11}$ can be omitted without affecting the limiting distribution of $\sqrt{T}(\hat{\beta}^* - \beta^*)$. The TSLS estimator is then obtained from (20.14) as

$$\hat{\beta}_{\text{TSLS}}^* = (\mathbf{P}_{12}^* \mathbf{A}_{22.1} \mathbf{P}_{12}^{*\prime})^{-1} \mathbf{P}_{12}^* \mathbf{A}_{22.1} \mathbf{p}_{12}. \tag{24.1}$$

This estimator was suggested by Basmann (1957) and Theil (1961). It corresponds in the linear functional relationship setup to ordinary least squares on the first coordinate. If some other coefficient of β were set equal to one, the minimization would be in the direction of that coordinate.

25. Distributions of estimators. Econometricians have studied intensively the distributions of TSLS and LIML estimators, particularly in the case of two endogenous variables. Exact distributions have been given by Basmann (1961), (1963), Richardson (1968), Sawa (1969), Mariano and Sawa (1972), Phillips (1980) and Anderson and Sawa (1982). (See also Anderson, 1980.) These have not been very informative because they are usually given in terms of infinite series, the properties of which are unknown or irrelevant.

A more useful approach is by approximating the distributions. Asymptotic expansions of distributions have been made by Sargan and Mikhail (1971), Anderson and Sawa (1973), Anderson (1974), Kunitomo (1980) and others;.

Phillips (1982) has studied the Padé (1892) approach. See also Anderson (1977), (1982).

Tables of the distributions of the TSLS and LIML estimators in the case of two endogenous variables have been given by Anderson and Sawa (1977), (1979), Anderson, Kunitomo, and Sawa, (1982).

Anderson, Kunitomo and Sawa (1983) have graphed densities of the Maximum Likelihood (ML) estimator and the Least Squares (LS) estimator (minimizing in one direction) for the linear functional relationship (Section 2) for the case $p = 2$, $m = q = 1$, $\Psi = \sigma^2 \mathbf{I}_p$ and for various values of β, n, and

$$\delta^2 = (1/\sigma^2) \sum_{\alpha=1}^{n} (\mu_\alpha - \bar{\mu})^2. \tag{25.1}$$

Some of these figures are given here (Figures 6 to 12). Each curve is the density of $(\hat{\beta} - \beta)/\text{ASD}$, where ASD is the asymptotic standard deviation. That is, each density is approximated by the standard normal density. The approximation is very good for the maximum likelihood estimator except when δ^2 is about the same size as n. For given β and n, the accuracy of the approximation increases with the noncentrality parameter δ^2. When $\beta = 0$ (the line in Figure 1 is horizontal), the least squares estimator (minimizing deviations in the vertical direction) has more accuracy than that described by the standard normal density because the least squares estimated line tends to be horizontal. When $\beta \neq 0$, the

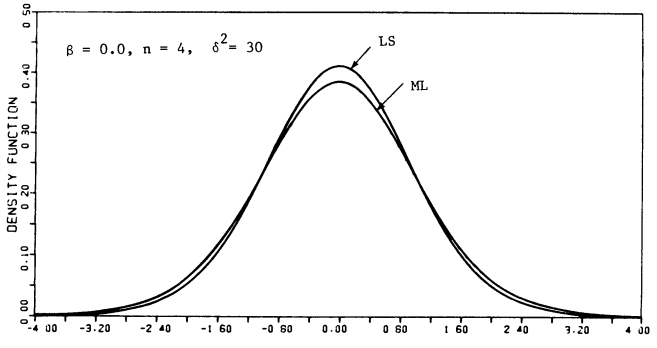

FIG. 6.

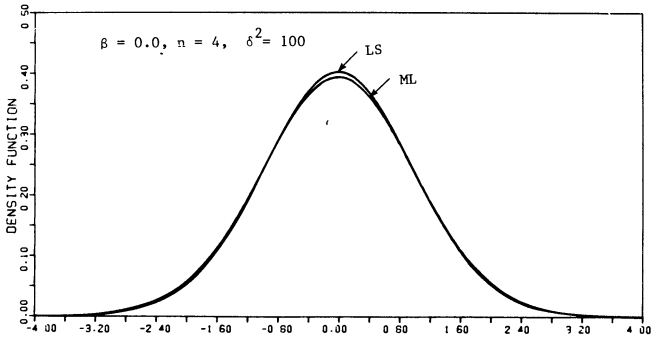

FIG. 7.

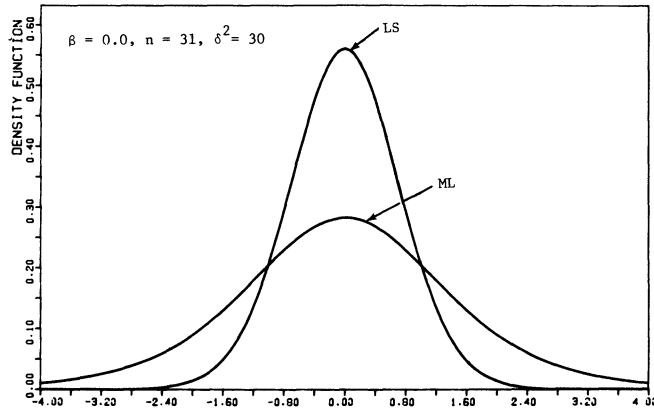

FIG. 8.

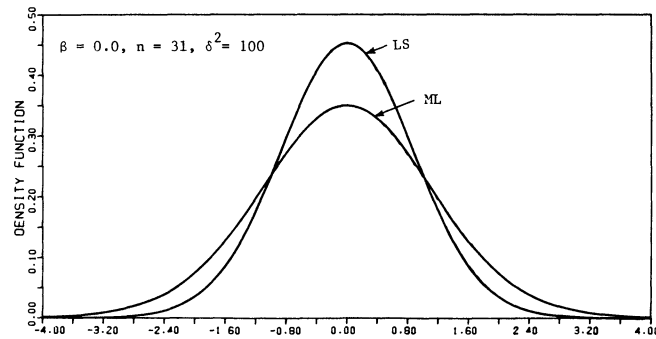

FIG. 9.

least squares estimator is biased towards 0; the bias increases with n (for fixed δ^2) and decreases with δ^2 (for fixed n). An extreme case is indicated in the last figure.

26. Some asymptotic theory for linear functional and structural relationships. We now return to the estimators given in Part I. It should be noted that if $\hat{\mathbf{B}} = (\hat{\mathbf{B}}_1\ \hat{\mathbf{B}}_2)$ is a maximum likelihood estimator of the coefficients of the restrictions (regardless of identification conditions) then $(\mathbf{I}_q\ \hat{\mathbf{B}}_1^{-1}\hat{\mathbf{B}}_2) = (\mathbf{I}_q\ \hat{\mathbf{B}}_2^*)$ is the maximum likelihood estimator of the coefficients when the first submatrix is required to be $\mathbf{I}_q$. The asymptotic distribution of $\hat{\mathbf{B}}_2^*$ can be obtained from the asymptotic distribution of $\hat{\mathbf{B}}$ (by the usual Taylor series expansion). The asymptotic distributions of $\hat{\mathbf{B}}$ and $\hat{\mathbf{\Psi}}$ are obtained from the asymptotic distributions of $\mathbf{C}$ or of $\mathbf{H}$ and $\mathbf{G}$; the asymptotic distribution for the structural relationship holds for the functional relationship.

In the case of the linear functional relationship when $\mathbf{\Psi} = \sigma^2\mathbf{I}_p$ Gleser (1981) found the asymptotic normal distribution of $\hat{\mathbf{B}}_2^*$ when $n \to \infty$, $(1/n)\ \sum_{\alpha=1}^{n} \nu_\alpha \nu_\alpha'$ approaches a limit of rank m, and the fourth order moments of (the identically distributed) $\mathbf{u}_\alpha$ are finite.

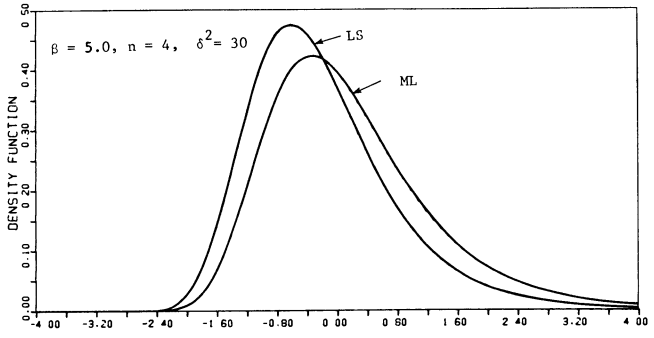

FIG. 10.

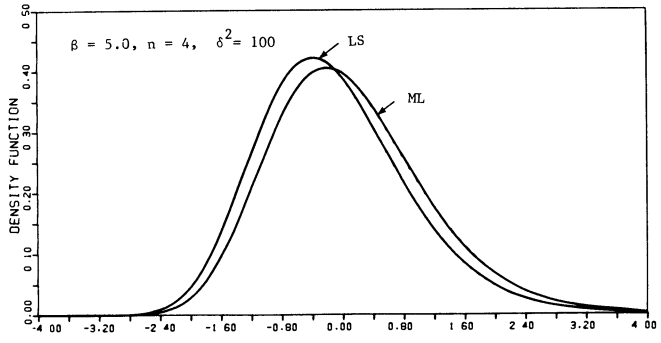

FIG. 11.

Anderson (1963) obtained the asymptotic distribution of the characteristic roots and vectors of the covariance matrix of a sample from a normal distribution when Σ has roots of arbitrary multiplicities. The estimator $\hat{\mathbf{B}} = \mathbf{W}_2'$ for the linear structural relationship is included as a special case. Lawley (1953) gave the asymptotic covariances.

When Ψ is unrestricted and is estimated by replicated observations, as in Sections 6 and 7, the asymptotic theory can depend on $k \to \infty$, $n \to \infty$, or both. Anderson (1951b) gave the asymptotic distribution of the roots of (6.4) and the vectors satisfying (6.5) as $k \to \infty$ for n fixed under general conditions; the estimator $\hat{\mathbf{B}} = \mathbf{Y}_2'$ of the linear functional relationship is included. One aspect of this asymptotic distribution is that $\sqrt{k}\, d_i \to_p 0$, $i = m+1, \cdots, p$. Hence, the maximum likelihood estimator of $\mathbf{B}_2^*$ has the same asymptotic distribution as the least squares estimator

$$(26.1)\qquad \hat{\mathbf{B}}_{\mathrm{LS}}^* = -\sum_{\alpha=1}^{n} (\bar{\mathbf{x}}_\alpha^{(1)} - \bar{\mathbf{x}}^{(1)})(\bar{\mathbf{x}}_\alpha^{(2)} - \bar{\mathbf{x}}^{(2)})'\Big[\sum_{\alpha=1}^{n} (\bar{\mathbf{x}}_\alpha^{(2)} - \bar{\mathbf{x}}^{(2)})(\bar{\mathbf{x}}_\alpha^{(2)} - \bar{\mathbf{x}}^{(2)})'\Big]^{-1},$$

where $\mathbf{x}_\alpha' = (\mathbf{x}_\alpha^{(1)\prime}\mathbf{x}_\alpha^{(2)\prime})$. Here $\hat{\mathbf{B}}_{\mathrm{LS}}^* \to_p \mathbf{B}_2^*$ and

$$(26.2)\qquad \sqrt{k}\ \mathrm{vec}(\mathbf{B}_{\mathrm{LS}}^* - \mathbf{B}_2^*) \to_{\mathscr{L}} N\left[\mathbf{0},\ \Big(\sum_{\alpha=1}^{n} \boldsymbol{\nu}_\alpha^{(2)}\boldsymbol{\nu}_\alpha^{(2)\prime}\Big)^{-1} \otimes (\mathbf{I}_q\ \mathbf{B}_2^*)\Psi\begin{pmatrix}\mathbf{I}_q\\ \mathbf{B}_2^{*\prime}\end{pmatrix}\right],$$

where $\boldsymbol{\nu}_\alpha^{(1)\prime} = (\boldsymbol{\nu}_\alpha^{(1)\prime}\boldsymbol{\nu}_\alpha^{(2)\prime})$. (See Anderson, 1983b.)

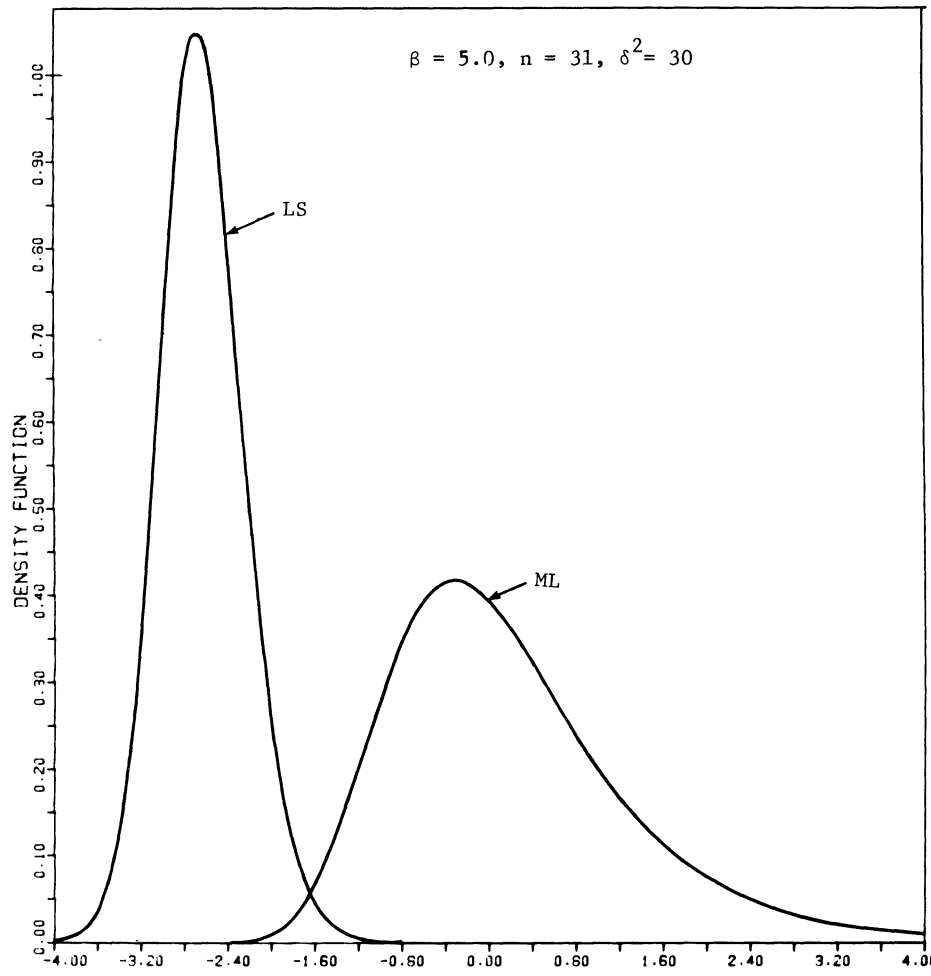

FIG. 12.

When $n \to \infty$, Amemiya and Fuller (1984) have given the asymptotic normal distribution of $\hat{\mathbf{B}}_2^*$ for both the functional and structural case. In the functional model $\hat{\mathbf{\Psi}}_F$ is not a consistent estimator of $\mathbf{\Psi}$, but in the structural model $\hat{\mathbf{\Psi}}_S$ is strongly consistent and asymptotically normal (Amemiya and Fuller, 1984). Fuller (1980) gave other results for $q = 1$. Anderson (1983a) has treated the roots of (6.4) and the vectors satisfying (6.5) in the structural case; $\hat{\mathbf{B}} = \mathbf{Y}_2'$ is included. Kunitomo (1980) and Patefield (1976) have given asymptotic expansions for $q = 1$ and $p = 2$.

Acknowledgments. The author wishes to thank Yasuo Amemiya, Persi Diaconis, Wayne Fuller, Leon Gleser, Lawrence Mayer, P. A. P. Moran, Ingram Olkin, Peter Sprent, and Akimichi Takamura for useful suggestions and Lucia

Alviano, Byoung-Seon Choi, David Criswell, Aaron Han, Debbie Olson and Amy Schwartz for help in preparing the manuscript.

REFERENCES

ADCOCK, R. J. (1978). A problem in least squares. *The Analyst* **5** 53–54.

AMEMIYA, YASUO and FULLER, WAYNE (1984). Estimation for the multivariate errors-in-variables model with estimated error covariance matrix. *Ann. Statist.* **12**, to appear.

ANDERSON, T. W. (1946a). The non-central Wishart distribution and certain problems of multivariate statistics. *Ann. Math. Statist.* **17** 409–431. (Correction, *Ann. Math. Statist.* **35** (1964) 923–924).

ANDERSON, T. W. (1946b). Analysis of multivariate variance. Unpublished.

ANDERSON, T. W. (1951a). Estimating linear restrictions on regression coefficients for multivarate normal distributions. *Ann. Math. Statist.* **22** 327–351. (Correction, *Ann. Statist.* **8** (1980) 1400).

ANDERSON, T. W. (1951b). The asymptotic distribution of certain characteristic roots and vectors. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, (Jerzy Neyman, ed.) 105–130. University of California.

ANDERSON, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

ANDERSON, T. W. (1959). Some scaling models and estimation procedures in the latent class model. *Probability and Statistics: The Harold Cramér Volume*, 9–38. (Ulf Grenander, ed.), Almqvist and Wiksell, Stockholm.

ANDERSON, T. W. (1963). Asymptotic theory for principal component analysis. *Ann. Math. Statist.* **34** 122–148.

ANDERSON, T. W. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *Ann. Statist.* **1** 135–141.

ANDERSON, T. W. (1974). An asymptotic expansion of the distribution of the limited information maximum likelihood estimate of a coefficient in a simultaneous equation system. *J. Amer. Statist. Assoc.* **69** 565–573.

ANDERSON, T. W. (1976). Estimation of linear functional relationships: approximate distributions and connections with simultaneous equations in econometrics. *J. Roy. Statist. Soc. B* **38** 1–36.

ANDERSON, T. W. (1977). Asymptotic expansions of the distributions of estimates in simultaneous equations for alternative parameter sequences. *Econometrica* **45** 509–518.

ANDERSON, T. W. (1980). Recent results on the estimation of a linear functional relationship. *Multivariate Analysis V* 23–34. (P. R. Krishnaiah, ed.), North-Holland Publishing Co., Amsterdam.

ANDERSON, T. W. (1982). Some recent developments on the distributions of single-equation estimators, 109–122. (W. Hildenbrand, ed.) Cambridge University Press, New York.

ANDERSON, T. W. (1983a). Linear structural relationships with unknown error covariance matrix. Unpublished.

ANDERSON, T. W. (1983b). Asymptotic distribution of an estimator of linear functional relationships. Unpublished.

ANDERSON, T. W., and HURWICZ, LEONID (1947). Errors and shocks in economic relationships. *Bulletin of the International Statistical Institute* **31(5)** 23–25. (Also *Econometrica Supp.* **17** (1949) 23–25.)

ANDERSON, T. W., KUNITOMO, N., and SAWA, T. (1982). Evaluation of the distribution function of the limited information maximum likelihood estimator. *Econometrica* **50** 1009–1027.

ANDERSON, T. W., KUNITOMO, N., and SAWA, T. (1983). Comparison of the densities of the TSLS and LIMLK estimators. *Global Econometrics, Essays in Honor of Lawrence R. Klein.* 103–124 (F. Gerard Adams and Bert G. Hickman, eds.) M.I.T. Press.

ANDERSON, T. W., and RUBIN, HERMAN (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *Ann. Math. Statist.* **20** 46–63. (Reprinted in

Readings in Econometric Theory 358–375. (J. Malcolm Dowling and Fred R. Glahe, eds.), Colorado Associated University Press, 1970.)

ANDERSON, T. W., and RUBIN, HERMAN (1950). The asymptotic properties of estimates of the parameters of a single equation in a complete system of stochastic equations. *Ann. Math. Statist.* **21** 570–582. (Reprinted in *Readings in Econometric Theory* 376–388. (J. Malcolm Dowling and Fred R. Glahe, eds.), Colorado Associated University Press, 1970.)

ANDERSON, T. W., and RUBIN, HERMAN (1956). Statistical inference in factor analysis. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **5** 111–150. (Jerzy Neyman, ed.). University of California Press.

ANDERSON, T. W., and SAWA, T. (1973). Distribution of estimates of coefficients of a single equation in a simultaneous system and their asymptotic expansions. *Econometrica* **41** 683–714.

ANDERSON, T. W., and SAWA, T. (1977). Tables of the distribution of the maximum likelihood estimate of the slope coefficient and approximation. Technical Report No. 234, Institute for Mathematical Studies in the Social Sciences, the Economics Series, Stanford University, April, 1977.

ANDERSON, T. W., and SAWA, T. (1979). Evaluation of the distribution function of the two-stage least squares estimate. *Econometrica* **47** 163–182.

ANDERSON, T. W. and SAWA, T. (1982). Exact and approximate distributions of the maximum likelihood estimator of a slope coefficient. *J. Roy. Statist. Soc. B* **44** 52–62.

ANDERSON, T. W., STEIN, CHARLES, and ZAMAN, ASAD (1982). Best invariant estimator of a direction parameter with application to linear functional relationships and factor analysis. Technical Report No. 381, Institute for Mathematical Studies in the Social Sciences, the Economics Series, Stanford University, August, 1982.

BARTLETT, M. S. (1947). Multivariate analysis. *J. Roy. Statist. Soc. Suppl.* **9** 176–197.

BARTLETT, M. S. (1948a). A note on the statistical estimation of demand and supply relations from time series. *Econometrica* **16** 323–329.

BARTLETT, M. S. (1948b). Internal and external factor analysis. *British J. Psych. (Statist. Sec.)*, **1** 73–81.

BASMANN, R. L. (1957). A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica* **25** 77–83.

BASMANN, R. L. (1961). A note on the exact finite sample frequency functions of generalized classical linear estimators in two leading overidentified cases. *J. Amer. Statist. Assoc.* **56** 619–636.

BASMANN, R. L. (1963). A note on the exact finite sample frequency functions of generalized classical linear estimators in a leading three equation case. *J. Amer. Statist. Assoc.* **58** 161–171.

BHARGAVA, A. K. (1979). Estimation of a linear transformation and an associated distributional problem. *J. Statist. Plan. Infer.* **3** 19–26.

BROWNE, M. W. (1974). Generalized least squares estimates in the analysis of covariance structures. *South African Statist. J.* **8** 1–24.

CHAN, N. N. (1980). Estimating linear functional relationships. *Recent Developments in Statistical Inference and Data Analysis*, 27–32. (K. Matusita, ed.). North-Holland, Amsterdam.

CHOW, GREGORY, C., and RAY-CHAUDHURI, D. K. (1967). An alternative proof of Hannan's theorem on canonical conditions and multiple equation systems. *Econometrica* **35** 139–142.

COPAS, J. B. (1972). The likelihood surface in the linear functional relationship problem. *J. Roy. Statist. Soc. Ser. B* **34** 190–202.

DAHM, P. FRED, and FULLER, WAYNE A. (1981). Generalized least squares estimation of the functional multivariate linear errors in variables model. Unpublished.

DEMPSTER, A. P., LAIRD, N. M., and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B* (with discussion) **39** 1–38.

DENT, B. (1935). On observations of points connected by a linear relation. *Proc. of the Physical Soc.* **47** 92–106.

DOLBY, GEOFFREY, R. (1976). The ultrastructural relation: a synthesis of the functional and structural relations. *Biometrika* **63** 39–50.

ECKART, CARL, and YOUNG, GALE (1936). The approximation of one matrix by another of lower rank. *Psychometrika* **1** 211–218.

FISHER, R. A. (1938). The statistical utilization of multiple measurements. *Ann. Eugen.* **8** 376–386.

FRISCH, RAGNAR (1934). *Statistical Confluence Analysis by Means of Complete Regression Systems.* Universitetets Økonomiske Institutt, Oslo [Publikasjon Nr. 5. Separate impression *Nordic Statist. J.* **5** 1933].

FULLER, WAYNE A. (1980). Properties of some estimators for the errors-in-variables model. *Ann. Statist.* **8** 407–422.

FULLER, WAYNE A., PANTULA, G. SASTRY, and AMEMIYA, YASUO (1982). The covariance matrix of estimators for the factor model. Unpublished.

GARNETT, J. C. M. (1919). On certain independent factors in mental measurement. *Proc. Royal Soc. London* **46** 91–111.

GEARY, R. C. (1942). Inherent relations between random variables. *Proc. Roy. Irish Acad.* **47** 63–76.

GEARY, R. C. (1948). Studies in relations between economic time series. *J. Roy. Statist. Soc. B* **10** 140–158.

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

GINI, C. (1921). Sull'interpolazaione di una retta quando i valori della variabile independente sono affetti da errori accidentali. *Metron* **1** 63–82.

GLESER, LEON JAY (1981). Estimation in a multivariate 'errors in variables' regression model: large sample results. *Ann. Statist.* **9** 24–44.

GLESER, LEON JAY, and WATSON, GEOFFREY S. (1973). Estimation of a linear transformation. *Biometrika* **60** 525–534.

GOLDBERGER, ARTHUR S. (1974). Unobservable variables in econometrics. *Frontiers in Econometrics*, 193–213 (P. Zarembka, ed.). Academic, New York.

HAAVELMO, T. (1943). The structural implication of a system of simultaneous equations. *Econometrica* **11** 1–12.

HAAVELMO, TRYGVE (1944). The probability approach in econometrics. *Econometrica* **12**(Supp) 1–118.

HANNAN, E. J. (1967). Canonical correlation and multiple equation systems in economics. *Econometrica* **35** 123–138.

HEALY, JOHN D. (1980). Maximum likelihood estimation of a multivariate linear functional relationship. *J. Multivariate Anal.* **10** 243–251.

HÖSCHEL, H. P. (1978). Generalized least squares estimators of linear functional relations with known error-covariance. (German and Russian summaries). *Math. Operationsforsch. Statist., Ser. Statist.* **9** 9–26.

HOTELLING, HAROLD (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psych.* **24** 417–441, 498–520.

HOWE, W. G. (1955). Some contributions to factor analysis. Report No. ORNL-1919. Oak Ridge National Laboratory, Oak Ridge, Tennessee.

HSIAO, CHENG (1976). Identification and estimation of simultaneous equation models with measurement error. *Internat. Econom. Rev.* **17** 319–339. Corrigendum (1977). *Internat. Econom. Rev.* **18** 252.

IZENMAN, ALAN JULIAN (1975). Reduced-rank regression for the multivariate linear model. *J. Multivariate Anal.* **5** 248–264.

JENNRICH, ROBERT I., and THAYER, DOROTHY T. (1973). A note on Lawley's formulas for standard errors in maximum likelihood factor analysis. *Psychometrika* **38** 571–580.

JÖRESKOG, KARL G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* **34** 183–202.

JÖRESKOG, KARL G., and GOLDBERGER, ARTHUR S. (1972). Factor analysis by generalized least squares. *Psychometrika* **37** 243–260.

KENDALL, MAURICE G., and STUART, ALAN (1979). *The Advanced Theory of Statistics, Vol. 2, Inference and Relationship, Fourth Edition.* Griffin, London.

KLOTZ, JEROME, and PUTTER, JOSEPH (1969). Maximum likelihood estimation of the multivariate covariance components for the balanced one-way layout. *Ann. Math. Statist.* **33** 273–289.

KOOPMANS, T. C. (1937). *Linear Regression Analysis of Economic Time Series.* Bohn, Haarlem, Netherlands.

KOOPMANS, T. C. ed (1950). *Statistical Inference in Dynamic Economic Models.* Wiley, New York.

KOOPMANS, T. C., and REIERSØL, O. (1950). The identification of structural characteristics. *Ann. Math. Statist.* **21** 165–181.

KUMMEL, C. H. (1879). Reduction of observed equations which contain more than one observed quantity. *The Analyst* **6** 97–105.

KUNITOMO, NAOTO (1980). Asymptotic expansions of the distributions of estimators in a linear functional relationship and simultaneous equations. *J. Amer. Statist. Assoc.* **75** 693–700.

LAWLEY, D. N. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proc. Roy. Soc. Edinburgh, Sect. A* **60** 64–82.

LAWLEY, D. N. (1941). Further investigations in factor estimation. *Proc. Roy. Soc. Edinburgh, Sect. A* **61** 176–185.

LAWLEY, D. N. (1943). The application of the maximum likelihood method to factor analysis. *British J. Psych.* **33** 172–175.

LAWLEY, D. N. (1953). A modified method of estimation in factor analysis and some large sample results. *Uppsala Symposium on Psych. Factor Anal. 17–19 March 1953 (Nordisk Psykoligi's Monograph Series* **3** 35–42. Almqvist and Wicksell, Stockholm.

LAWLEY, D. N. (1958). Estimation in factor analysis under various initial assumptions. *British J. Statist. Psych.* **11** 1–12.

LAWLEY, D. N. (1967). Some new results in maximum likelihood factor analysis. *Proc. Roy. Soc. Edinburgh, Sect. A* **87** 256–264.

LAWLEY, D. N., and MAXWELL, A. E. (1971). *Factor Analysis as a Statistical Method, Second Edition.* American Elsevier, New York.

LINDLEY, D. V. (1947). Regression lines and the linear functional relationship. *J. Roy. Statist. Soc. Supp.* **9** 218–244.

LINDLEY, D. V., and EL-SAYYAD, G. M. (1968). The Bayesian estimation of a linear functional relationship. *J. Roy. Statist. Soc. Ser. B* **30** 190–202.

MADANSKY, ALBERT (1959). The fitting of straight lines when both variables are subject to error. *J. Amer. Statist Assoc.* **55** 173–205.

MALINVAUD, E. (1964). *Statistical Methods of Econometrics.* North-Holland, Amsterdam.

MARIANO, ROBERTO S., and SAWA, T. (1972). The exact finite-sample distribution of the limited-information maximum likelihood estimator in the case of two included endogenous variables. *J. Amer. Statist. Assoc.* **67** 159–163.

MORAN, P. A. P. (1971). Estimating structural and functional relationships. *J. Multivariate Anal.* **1** 232–255.

MORIMUNE, KIMIO, and KUNITOMO, NAOTO (1980). Improving the maximum likelihood estimate in linear functional relationships for alternative parameter sequences. *J. Amer. Statist. Assoc.* **75** 230–237.

MORRIS, BLAIR, and OLKIN, INGRAM (1964). Some estimation and testing problems for factor analysis models. Unpublished.

NUSSBAUM, M. (1976). Maximum likelihood and least squares estimation of linear functional relationships. (German and Russian summaries). *Math. Operationsforsch. Statist., Ser. Statist.* **7** 23–49.

PADÉ, H. (1892). Sur la representation approchee d'une fonction par des fractions rationelles. *Ann. Sci. Ecole Norm. Sup., Supp.* **9** 1–93.

PATEFIELD, W. M. (1976). On the validity of approximate distributions arising in fitting a linear functional relationship. *J. Statist. Comput. Simulation* **5** 43–60.

PATEFIELD, W. M. (1981). Multivariate linear relationships: maximum likelihood estimation and regression bounds. *J. Roy. Statist. Soc. Ser. B* **43** 342–352.

PHILLIPS, P. C. B. (1980). The exact finite sample density of instrumental variable estimators in an equation with $n + 1$ endogenous variables. *Econometrica* **48** 861–878.

PHILLIPS, P. C. B. (1982). A new approach to small sample theory. Cowles Foundation, Yale University, mimeograph.

PEARSON, K. (1901). On lines and planes of closest fit to systems of points in space. *Philos. Mag.* **2** 559–572.

RAO, C. R. (1955). Estimation and tests of significance in factor analysis. *Psychometrika* **20** 93–111.

RAO, C. R. (1973). *Linear Statistical Inference and its Application, second edition.* Wiley, New York.

REIERSØL, O. (1945). Confluence analysis by means of instrumental sets of variables. *Arkiv för Matematik, Astonomi och Fysk* **32** 1–119.

REIERSØL. O. (1950). On the identifiability of parameters in Thurstone's multiple factor analysis. *Psychometrika* **15** 121–149.

RICHARDSON, D. H. (1968). The exact distribution of a structural coefficient estimator. *J. Amer. Statist. Assoc.* **63** 1214–1226.

RUBIN, DONALD B., and THAYER, DOROTHY T. (1982). EM algorithms for ML factor analysis. *Psychometrika* **47** 69–76.

SARGAN, J. D., and MIKHAIL, W. M. (1971). A general approximation to the distribution of instrumental variables estimates. *Econometrica* **39** 131–169.

SAWA, T. (1969). The exact finite sampling distribution of ordinary least squares and two stage least squares estimator. *J. Amer. Statist. Assoc.* **64** 923–936.

SOLARI, M. E. (1969). The maximum likelihood solution of the problem of estimating a linear functional relationship. *J. Roy. Statist. Soc. Ser. B* **31** 372–375.

SPEARMAN, CHARLES (1904). General intelligence, objectively determined and measured. *Amer. J. Psych.* **15** 201–293.

SPRENT, P. (1966). A generalized least-squares approach to linear functional relationships (with discussion). *J. Roy. Statist. Soc. Ser. B* **28** 278–297.

THEIL, H. (assisted by J. S. CRAMER, H. MOERMAN, and A. RUSSCHEN) (1961). *Economic Forecasts and Policy, second edition.* North-Holland, Amsterdam. Contributions to Economic Analysis No. XV. (First published 1958).

THEOBALD, C. M. (1975). An inequality with application to multivariate analysis. *Biometrika* **62** 461–466.

THOMSON, GODFREY H. (1934). Hotelling's method modified to give Spearman's '*G*'. *J. Educ. Psych.* **25** 336–374.

THURSTONE, L. L. (1935). *The Vectors of Mind.* Univ. of Chicago Press.

THURSTONE, L. L. (1947). *Multiple-Factor Analysis.* Univ. of Chicago Press.

TINTNER, GERHARD (1945). A note on rank, multicollinearity and multiple regression. *Ann. Math. Statist.* **16** 304–308.

TINTNER, GERHARD (1946). Multiple regression for systems of equations. *Econometrica* **14** 5–36.

TINTNER, GERHARD (1950). A test for linear relations between weighted regression coefficients. *J. Roy. Statist. Soc. Ser. B* **12** 273–277.

TINTNER, GERHARD (1952). *Econometrics.* Wiley, New York.

TUKEY, J. W. (1951). Components in regression. *Biometrics* **7** 33–69.

VILLEGAS, C. (1961). Maximum likelihood estimation of a linear functional relationship. *Ann. Math. Statist.* **32** 1048–1062.

VAN UVEN, M. J. (1930). Adjustment of N points (in n-dimensional space). II, *Koninklyke Academie van Wetenschappen to Amsterdam, Proc. Sect. Sci.* **33** 307–326.

WHITTLE, P. (1952). On principal components and least squares methods of factor analysis. *Skandinavisk Aktuarietidskrift* **35** 223–239.

WU, C. F. JEFF (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11** 95–103.

DEPARTMENT OF ECONOMICS
COLUMBIA UNIVERSITY
NEW YORK, NEW YORK 10027