Multivariate functional and structural relationships.

Theobald, C. M.

Award date:
1975

Awarding institution:
University of Bath

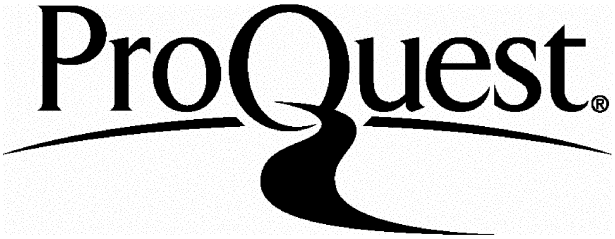
MULTIVARIATE FUNCTIONAL AND STRUCTURAL RELATIONSHIPS

submitted by C.M. Theobald

for the degree of Ph.D.

of the University of Bath

1975



ABSTRACT

This dissertation is concerned with sets of random variables connected by relationships which may be interpreted as exact linear relations between unobserved 'true' values which are obscured by random 'errors' or 'departures'. The great majority of published work on this topic relates to two random variables connected by a single relation; examples are given here in which it is necessary to consider more than two variates and more than one relation between them. Methods are proposed for representing such relationships between an arbitrary number of random variables, and connexions are established with other statistical models, in particular factor analysis. After a review of methods of estimation proposed for certain cases of varying generality, a fairly comprehensive treatment is given of the estimation of such relationship by the methods of maximum likelihood and generalized least-squares; the large-sample behaviour of the estimators is considered, and connexions with the technique of canonical analysis are established. In the course of this study, some inequalities for matrix traces are derived which are of wider mathematical interest. One of the procedures developed here is applied to a problem of comparing different instruments designed to measure the same property, examining their relative calibrations and relative precision. The data are also used to illustrate graphical techniques developed for testing the assumptions of the calibration model.

ACKNOWLEDGEMENTS

The author wishes to thank Professor G.B. Wetherill for proposing the topic of this thesis and for his suggestions and encouragement in the course of the work, and Professor V. Barnett for useful discussions in its later stages; also Dr. J.R. Mallinson for writing and running most of the computer programs.

CONTENTS

CONTENTS (continued) Page

1. INTRODUCTION

We are concerned in this dissertation with relationships between random variables which have the characteristics (i) that they may be expressed as the sum of an unobserved 'true' or 'underlying' variable and a random variable interpretable as an 'error' or 'departure' from the true value, and (ii) that the true variables are exactly linearly related. If the true variables are random the linear relationships are called *structural*, if mathematical variables they are *functional* relationships.

It is usually assumed that the errors are independent of the true values and independent between observations, and that their variances and covariances are the same for each observation. Also, the true values connected by a structural relationship are usually taken to be independent and identically distributed for different observations.

The topic must be distinguished from regression and multiple regression, in which only one variate is observed with 'error', and from multivariate regression, in which at least one variable is observed without error.

The great majority of the work on this subject is concerned with relationship between two variables. After some examples which illustrate contexts in which structural and functional relationships arise and the problems associated with their estimation, we give various mathematical representations of such relationships and go on to consider their estimation by maximum likelihood (ML), generalized least-squares (GLS), and other methods. We concentrate on the first two methods here since they are considered almost exclusively in later chapters. Extensive discussions of the various methods are given by Madansky (1959), Kendall and Stuart (1967, Ch. 29), and Moran (1971).

1.1 Some examples

(a) The strength and hardness of artillery shells

Madansky (1959) records the measurements of yield strength (Y_1) and

hardness (Y_2) of steel artillery shells. He assumes a linear relationship

$$\xi_2 = \alpha + \beta\xi_1$$

between the corresponding 'true' measurements ξ_1 and ξ_2 on each shell. The true measurements cannot be observed directly, only random variables Y_1, Y_2 expressible as

$$Y_1 = \xi_1 + \varepsilon_1 \quad , \quad Y_2 = \xi_2 + \varepsilon_2 \quad ,$$

where ε_1 and ε_2 are error terms resulting from inhomogeneity of the steel and errors of measurement. If we take ε_1 and ε_2 to be independently and normally distributed with zero means and variances σ_{11}, σ_{22} respectively, and suppose that there are n shells with true values of yield strength ξ_{1j} ($j = 1,\ldots,n$), then the unknown parameters in this functional relationship model are α, β, σ_{11}, σ_{22} and the ξ_{1j}.

The method of maximum likelihood does not give consistent estimators for this model without further assumptions. Indeed, a too-simple analysis of the problem led first to the belief that the application of ML gave a nonsensical result, and later that there was no such solution. We shall consider the theory for this and other functional relationship models in Section 1.4.

(b) The comparison of two measures of the impact strength of timber

Williams (1959, p.200) gives the results of a test of impact strength applied to two specimens from each of 109 planks of Northern Silver Ash. One specimen from each plank was tested radially, the other tangentially to the growth rings, giving readings Y_1 and Y_2 respectively. As in the last example, Y_i was considered as being made up of a true measurement ξ_i and a random error ε_i ($i = 1,2$), ξ_1 and ξ_2 being linearly related and the errors ε_1 and ε_2 being independent. It was of interest

to test whether the data were consistent with the relationship $\xi_2 = \alpha+\xi_1$.

An important difference from the last example is that since the two readings were made with the same method of test it is reasonable to assume that the error variances σ_{11} and σ_{22} are equal. With this assumption, ML gives consistent estimators for the parameters of the relationship but not for the common error variance.

(c) The comparative calibration of measuring instruments

The estimation of the functional relationship in the last example may be regarded as the calibration of one measurement of a property of the wood against another. The idea may be extended to the calibration of $p(\geq 2)$ instruments for measuring the same property using readings on a common group of n specimens. Expressing the true values of the remaining $p-1$ instruments in terms of the first, we have the $p-1$ linear functional relationships

$$\xi_i = \alpha_i + \beta_i \xi_1 \qquad (i = 2,3,\ldots,p); \qquad (1.1)$$

these $p-1$ equations together define a line in p-space. In Williams' (1969) representation of the same problem, the line on which the true measurements lie is defined by expressing each as a linear function of a hypothetical standard measurement. We observe the associated random variables Y_i which contain errors of measurement ε_i, so that

$$Y_i = \xi_i + \varepsilon_i \qquad (i = 1,\ldots,p). \qquad (1.2)$$

If $p>2$ and we take the ε_i to be uncorrelated between different specimens and different instruments, and to have mean zero and variances σ_{ii} $(i = 1,\ldots,p)$, then, in contrast to the case $p = 2$, we may obtain consistent estimators for the α_i, β_i and σ_{ii}.

Barnett (1969) also considers a structural-relationship version of the same problem; instead of the true measurements being mathematical variables, they are taken to be random variables distributed normally over the population of possible specimens. If we write them as X_i $(i=1,\ldots,p)$ then equations (1.1) and (1.2) become

$$X_i = \alpha_i + \beta_i X_1 \qquad (i = 2,3,\ldots,p)$$

and

$$Y_i = X_i + \varepsilon_i \qquad (i = 1,\ldots,p) .$$

If we take the ε_i to be distributed as $N(0,\sigma_{ii})$ independently of each other and of the X_i, then the parameters to be estimated are the α_i and β_i $(i = 2,3,\ldots,p)$, the σ_{ii} $(i = 1,\ldots,p)$, and the mean μ and variance σ^2 of X_1.

It is not possible to estimate all the parameters of the model when there are only two instruments to be compared, but there is a ML solution for $p>2$ which has an explicit form when $p = 3$.

Barnett applies this model to the comparison of two instruments for measuring human lung capacity. The instruments were assessed by taking measurements on a common group of 72 patients using two operatives; the total of four combinations of instrument and operative makes it possible to estimate the relationships consistently. Barnett estimates the parameters using some consistent moment estimators based on the exact ML solution in the case $p = 3$.

Various methods of representing these calibration models are considered in Section 2.3(a). In Chapter 6 the ML estimators for the structural case are considered and Barnett's data are re-analysed.

1.2 Representing the relationships

We give in this Section various notations for structural and

functional relationships, emphasising those which lead to useful p-variate generalizations; such generalizations are considered in Sections 2.1 and 2.2.

1.2.1. Structural relationships

We define the vector of observations $\underline{Y}$ as $(Y_1\ Y_2)'$; the i th element of $\underline{Y}$ is assumed to be made up of a 'true' variate X_i and an 'error' or 'departure' from the true value ε_i, so writing $\underline{X}$ for $(X_1\ X_2)'$ and $\underline{\varepsilon}$ for $(\varepsilon_1\ \varepsilon_2)'$ we have

$$\underline{Y} = \underline{X} + \underline{\varepsilon}\ .$$

The elements of the true variate are taken to be exactly linearly related; the usual method of writing this is

$$X_2 = \alpha + \beta X_1\ . \tag{1.3}$$

Writing μ_i for $E(X_i)$ $(i = 1,2)$, we have $\mu_2 = \alpha + \beta\,\mu_1$, so an alternative to (1.3) is

$$X_2 - \mu_2 = \beta(X_1 - \mu_1). \tag{1.4}$$

Measuring the X_i from their means has a number of advantages: it leads to simpler p-variate generalizations, and its symmetry makes estimation slightly easier and emphasizes the distinction between the structural relationship and regression models; the resemblance has given rise to much confusion. Another way of writing (1.4) is

$$\underline{\gamma}'(\underline{X} - \underline{\mu}) = 0, \tag{1.5}$$

where $\underline{\mu} = (\mu_1\ \mu_2)'$, $\underline{\gamma} = (\gamma_1\ \gamma_2)'$ and $\beta = -\gamma_1/\gamma_2$. Here $\underline{\gamma}$ is defined only up to a constant factor; if we wish to define it uniquely then the constraint $\gamma_2 = -1$ corresponding to (1.4) is only one of the more useful possibilities. Other equivalent conditions with useful generalizations are that the dispersion matrix of $\underline{X}$ has rank 1 or that we may write

$$X_i = \mu_i + \lambda_i f \qquad (i = 1,2), \tag{1.6}$$

where f has zero mean and unit variance. The latter follows from (1.4) if we write λ_1^2 for $V(X_1)$, λ_2 for $\beta\lambda_1$ and f for $(X_1-\mu_1)/\lambda_1$. The symmetry of (1.6) makes it particularly convenient. If X_1 and therefore X_2 are assumed to be normally distributed this is equivalent to f in (1.6) having a standard normal distribution.

The error vectors $\underline{\varepsilon}$ are usually taken to be independent of the true variates, and both true variates and error vectors corresponding to different observations are assumed independent. The dispersion matrix, $\underline{\Sigma} = (\sigma_{hi})$, of $\underline{\varepsilon}$ is usually assumed constant. Whether it is also diagonal depends on the context: this is the usual assumption if ε_1 and ε_2 are errors of measurement, but if they represent, say, biological variation then this is not likely to be appropriate.

1.2.2. Functional relationships

A functional relationship between two variates Y_1 and Y_2 is an exact linear relationship between their expected values. If we write $\underline{\xi} = (\xi_1\ \xi_2)'$ for the expected value of $\underline{Y} = (Y_1\ Y_2)'$ then the vector $\underline{\varepsilon}$ of 'errors' or 'departures' from $\underline{\xi}$ is defined by

$$\underline{Y} = \underline{\xi} + \underline{\varepsilon}\ .$$

Unlike the structural case, we cannot consider the true values corresponding to different observations as being a sample from some population; we must define the functional relationship with respect to the true but unknown values $\underline{\xi}_1,\ldots,\underline{\xi}_n$ for our sample of size n. For these n vectors to be collinear we must have for some α and β

$$\xi_{2j} = \alpha + \beta\xi_{1j} \qquad (j = 1,\ldots,n). \qquad (1.7)$$

In this case we have to estimate $\alpha,\beta,\xi_{11},\ldots,\xi_{1n}$ and the parameters defining the distribution of $\underline{\varepsilon}$. By analogy with (1.4) we may measure the $\underline{\xi}_j$ about their mean, although in this case it must be the 'sample'

mean $\bar{\underline{\xi}} = n^{-1} \Sigma \underline{\xi}_j$. We then have

$$\xi_{2j} - \bar{\xi}_2 = \beta(\xi_{1j} - \bar{\xi}_1) \qquad (j = 1,\ldots,n),$$

and we have to estimate $\bar{\xi}_2$ instead of α. This again is a special case of the set of equations

$$\underline{\gamma}'(\underline{\xi}_j - \bar{\underline{\xi}}) = 0 \qquad (j = 1,\ldots,n)$$

in which a constraint has to be applied to $\underline{\gamma}$ to make it unique. Equivalent conditions are that the matrix $\Sigma \underline{\xi}_j\underline{\xi}_j'$ has rank 1 or that we may write

$$\underline{\xi}_j = \bar{\underline{\xi}} + \underline{\lambda}\phi_j \qquad (j = 1,\ldots,n),$$

in which case the ϕ_j are subject to the constraint $\Sigma\phi_j = 0$. We may also fix the scale of the ϕ_j by adding the condition $\Sigma\phi_j^2 = n$.

The vectors of departures $\underline{\varepsilon}_j$ have zero means and are usually taken to be independent, or at least uncorrelated, with a common dispersion matrix $\underline{\Sigma}$. The correlation between ε_1 and ε_2 is usually taken as zero if they represent errors of measurement.

1.3 Maximum likelihood estimation of structural relationships

We consider first a structural-relationship model in which both X_1 and $\underline{\varepsilon}$ are normally distributed. If we use the representation (1.6) then the vectors of observations $\underline{Y}_1,\ldots,\underline{Y}_n$ are independently distributed as $N_2(\underline{\mu},\underline{\Sigma}+\underline{\lambda}\underline{\lambda}')$. The problem immediately arises that only five parameters are needed to define a bivariate normal distribution, while our model contains seven, or, if $\sigma_{12} = 0$, six. It is impossible to estimate the parameters of the model without making some extra assumptions, since different values of the parameters may give rise to the same dispersion matrix for $\underline{Y}$; the parameters are said to be non-identifiable in such a case.

The usual way out of this difficulty is to suppose that as well as

having $\sigma_{12} = 0$ we know the ratio $\sigma_{11} : \sigma_{22}$, so that we have only five unknown parameters. If we scale the variates so that $\sigma_{11} = \sigma_{22}$ then the solution amounts to fitting a line in the direction of the first component in a principal component analysis. This is a special case of the identifiable situation, with which we deal in Section 4.2.1, in which $\underline{\Sigma}$ is known up to a constant factor. Other identifiable cases in which solutions have been found are those in which $\sigma_{12} = 0$ and one or both of σ_{11} and σ_{22} are known (see, for example, Moran, 1971). The last is included in the case of known $\underline{\Sigma}$ given in Section 4.2.

Reiersøl (1950) shows that the problem of non-identifiability arises from the assumption of normality in the distributions of X_1 and $\underline{\varepsilon}$, in the sense that if $\underline{\varepsilon}$ is normal the slope parameter of the relationship is identifiable if and only if X_1 is non-normal. Thus it is possible in principle to estimate the parameters of the model provided we can be sure that X_1 is non-normal. A method of Geary which uses non-zero mixed cumulants of order greater than 2 is given in Section 1.5.

If the structural relationship is constrained to pass through a particular point, which we may take without loss of generality to be the origin, then a consistent estimate of the slope is $\bar{y}_2/\bar{y}_1$, provided that $\mu_1 \neq 0$. This is not the ML solution; an attempt in Section 4.2.4 to find this solution is unsuccessful.

It is of interest to examine whether the problem of non-identifiability extends to structural relationships between more than two variates. We stated in Example 1.1(c) that the parameters are identifiable in the important case of p-1 relationships between p variates with independent errors if $p \geq 3$. This is because there are 3p independent parameters while a p-variate normal distribution is defined by $\frac{1}{2}$ p(p+3) parameters; the difference, $\frac{1}{2}$ p(p-3), is non-negative when $p \geq 3$. The ML estimators in the

case of 3 variates have been given independently by Teissier (1955), Lord and Novick (1968, pp.216-219) and Barnett (1969). However, the parameters of a single relationship between 3 variates with independent departures are not identifiable; if we write the relationship as

$$X_3 - \mu_3 = \beta_1(X_1 - \mu_1) + \beta_2 (X_2 - \mu_2) ,$$

then there are 11 parameters to be estimated (3 means, 3 error variances, the variances and covariance of X_1 and X_2, β_1 and β_2) but only nine parameters are needed to define a trivariate normal distribution. A criterion for the identifiability of structural relationships between p variates with normal distributions and independent departures is given in Section 2.4. The estimation of structural relationships between p variates is considered in detail in Chapter 4.

If there are replicate observations, that is, several observations on $\underline{Y}$ at the same value of $\underline{X}$, then there is enough information to estimate all the parameters of the model. According to Moran (1971), no ML solution has published for this case although various consistent estimators have been suggested in the case p = 2. We show in Section 4.2.2 that such a solution exists when there are equal numbers of replicates at each value of $\underline{X}$ and $\underline{\Sigma}$ is a general positive-definite matrix, but the problem appears to be more difficult if there are unequal numbers of replicates or if $\underline{\Sigma}$ is diagonal.

1.4 Maximum likelihood estimation of functional relationships

If we use the representation (1.7) and assume that ε_1 and ε_2 are normal and independent then there are n+4 parameters: $\alpha,\beta,\xi_{11},\ldots,\xi_{1n}$, σ_{11} and σ_{22}. For the data (y_{1j},y_{2j}) (j = 1,...,n), Lindley (1947) equates to zero the first derivatives of the likelihood with respect to these parameters and obtains the remarkable result

$$\hat{\sigma}_{22} = \hat{\beta}^2 \hat{\sigma}_{11}. \qquad (1.8)$$

This relationship between the estimators of the slope of the line and the error variances is quite unacceptable, for there is no reason to suppose that the true values of the parameters are related in this way; it has been interpreted as a failure of the ML method. However, Solari (1969) has shown that the likelihood has two stationary points, both giving (1.8) but one having a higher value than the other, and that the higher point is not a maximum but a saddle point. She shows also that the likelihood tends to infinity if either $q_1 = \Sigma(y_{1j}-\xi_{1j})^2 = 0$ and σ_{11} tends to zero or $q_2 = \Sigma(y_{2j}-\xi_{2j})^2 = 0$ and σ_{22} tends to zero, and that the likelihood has essential singularities at all points at which either $q_1 = \sigma_{11} = 0$ or $q_2 = \sigma_{22} = 0$, so no ML solution exists. The problem is not confined to a relationship between two variates; we show in Section 2.4 that for relationships between any number of variates the likelihood is unbounded.

Copas (1972) shows that the above arguments are invalid in that they assume that the y_{1j} and y_{2j} are recorded with complete accuracy. The likelihood function is a product of probability densities which are only approximate in that they ignore the grouping error in the observations; the approximation is excellent so long as this grouping error is small compared with the standard deviation of the corresponding departure term, but invalid near parameter points at which the standard deviation is zero (see, for example, Kempthorne, 1966). The singularities found by Solari occur at points of just this type. Copas re-analyses the problem assuming that y_{1j} and y_{2j} are both recorded with a grouping interval of length h, and shows that as h tends to zero the ML estimator of β tends to the slope of the regression of y_2 on y_1 if $\Sigma(y_{2j}-\bar{y}_2)^2 < \Sigma\ (y_{1j}-\bar{y}_1)^2$ and the reciprocal of the slope of the regression of y_1 on y_2 if this inequality is reversed; in each case the error variance corresponding to the variable treated as

the regressor is estimated as zero. This estimator of β is not consistent if $n^{-1}\Sigma(\xi_{1j}-\bar{\xi}_1)^2$ converges to a finite limit, but it is clear from the structural case that no solution can be in general, since there could be no consistent solution if the ξ_{1j} were in fact a random sample from a normal distribution. However, the limits of $\bar{\underline{y}}$ and of the sums of squares and products are consistently estimated.

As in the structural case we can obtain a consistent estimator of β (in fact, the same estimator) if we assume knowledge of the ratio $\sigma_{11} : \sigma_{22}$, but even then the estimators of σ_{11} and σ_{22} converge to half their true values. The problem arises because a new parameter ξ_{1j} is introduced for each observation $\underline{y}_j$. Neyman and Scott (1948) call such parameters, which are specific to the distributions of individual observations, 'incidental'; parameters which appear in the distributions of all vectors of observations are 'structural'. The usual asymptotic properties of ML do not apply in the presence of incidental parameters. The general situation in which $\underline{\Sigma}$ is known up to a constant factor is treated in Section 5.1.1.

Unlike the structural case, there is no difficulty here in estimating relationships through the origin; we simply use raw rather than corrected sums of squares and products.

If there are replicate observations then consistent estimators may be constructed by finding the ML estimators given the value of $\underline{\Sigma}$ and estimating $\underline{\Sigma}$ using the within-groups sums of squares and products divided by their degrees of freedom; this is equivalent to the generalized least-squares solution described in the next Section. The full ML solution, given in Chapter 5, differs from this; it estimates the parameters of the relationships consistently, but whether $\underline{\Sigma}$ is estimated consistently depends on the way in which the total number of observations tends to infinity; if we have only a finite number of replicate observations at each $\underline{\xi}_j$ then

we are again plagued with incidental parameters.

1.5 Other methods of estimation

Because of the difficulties associated with ML estimation, various more-or-less ad hoc methods of estimation have been suggested which depend on having various extra pieces of information as well as the n pairs of observations (y_{1j}, y_{2j}) $(j = 1,\ldots,n)$.

1.5.1 Generalized least-squares

Sprent's (1966) generalized least-squares procedure may be regarded as an extension of the ML solution for a functional relationship for which the observations are independent with a common known dispersion matrix $\underline{\Sigma}$. It has been shown to be equivalent to ML in many cases where the error variances and covariances are known. We show in Section 5.3 that a further extension of the procedure to multiple relationships between any number of variates may be seen as a simplification of the ML method when $\underline{\Sigma}$ has to be estimated from replicate observations.

Suppose that we postulate a homogeneous functional relationship $\xi_{2j} = \beta\xi_{1j}$ $(j = 1,\ldots,n)$ between the elements of the mean vectors $\underline{\xi}_j$ of n bivariate observations $\underline{y}_j$ $(j = 1,\ldots,n)$ which are uncorrelated with a common known dispersion matrix $\underline{\Sigma}$. Sprent observes that the linear function $z = y_2 - \beta y_1$ has mean zero and variance independent of $\underline{\xi}$, and is thus what Williams (1955) has called a 'null variate'. If, for ease of generalization, we write the relationship as $\underline{\gamma}'\underline{\xi}_j = 0$ $(j = 1,\ldots,n)$ and define z as $\underline{\gamma}'\underline{y}$, then the generalized least-squares principle is to estimate $\underline{\gamma}$ by minimizing

$$U = \sum_j z_j^2 \{V(z_j)\}^{-1} = \underline{\gamma}'\underline{B}\underline{\gamma}/\underline{\gamma}'\underline{\Sigma}\underline{\gamma} \quad ,$$

where $\underline{B} = \Sigma\, \underline{y}_j\, \underline{y}_j'$. The minimum of U is $\delta_2(\underline{B})$, the smaller root of the determinantal equation $|\underline{B}-\delta\, \underline{\Sigma}| = 0$, and it is attained when $\{\underline{B} - \delta_2(\underline{B})\underline{\Sigma}\}\underline{\gamma} = \underline{0}$. In this case Sprent (1969, pp.41-42) shows that if the $\underline{y}_j$ are normally

distributed the procedure is equivalent to ML.

He suggests extensions of the method in various directions: if the $\underline{y}_j$, instead of being uncorrelated with equal dispersion matrices, have correlated departures, then the null variates also become correlated. If we write $\underline{\Sigma}_o$ for the 2n x 2n dispersion matrix of the vector $(\underline{y}_1', \ldots, \underline{y}_n')'$ then we may partition $\underline{\Sigma}_o$ into 2 x 2 submatrices

$$\underline{\Sigma}_{jj'} = C(\underline{y}_j, \underline{y}_{j'}) \qquad (j,j' = 1,\ldots,n).$$

The null variates z_j now have covariances $\theta_{jj'} = \underline{\gamma}'\underline{\Sigma}_{jj'}\underline{\gamma}$ $(j = 1,\ldots,n)$; if we write $\underline{\Theta}$ for $(\theta_{jj'})$ then the procedure is to estimate β by minimizing

$$U = \underline{z}' \, \underline{\Theta}^{-1} \, \underline{z} \tag{1.9}$$

where $\underline{z} = (z_1, \ldots, z_n)'$. Dolby (1972) has shown that this is also equivalent to ML for estimating a relationship through the origin. Relationships not constrained in this way may be estimated by replacing the z_j in (1.9) by $z_j - \alpha (j=1,\ldots,n)$; the covariances $\theta_{jj'}$ are unaffected, so (1.9) becomes

$$(\underline{z} - \alpha\underline{1})' \, \underline{\Theta}^{-1} \, (\underline{z} - \alpha\underline{1}) \tag{1.10}$$

where $\underline{1}$ is an n-vector of 1s. The minimum of (1.10) with respect to α for given β is

$$\underline{z}' \, \underline{\Theta}^{-1} \, \underline{z} - (\underline{1}' \, \underline{\Theta}^{-1} \, \underline{z})^2 / \underline{1}' \, \underline{\Theta}^{-1} \, \underline{1},$$

attained when $\alpha = \underline{1}' \, \underline{\Theta}^{-1} \, \underline{z} / \underline{1}' \, \underline{\Theta}^{-1} \, \underline{1}$.

The inclusion of the constant term in the relationship may be seen as a special case of another extension of the method, that to the estimation of a single functional relationship between p variates. When the vectors of observations are uncorrelated and have equal dispersion matrices the solution is a simple extension of that in the case p=2. A further, intuitive, extension to more than one relationship (Sprent 1969, p.91) is shown in Section 5.3 to be equivalent to ML under the assumption of

normality.

If we have uncorrelated vectors of observations with a common dispersion matrix Σ, then Sprent suggests estimating Σ by an appropriate multiple of the matrix of within-group sums of squares and products. We show in Section 5.1.2 that under the assumption of normality this gives the same estimators of the relationships as ML, although the estimator of Σ is not the same.

1.5.2. The method of cumulants

We have seen in Section 1.3 that a structural relationship is identifiable if X_1 is non-normal. In this case X_1 may have a finite non-zero cumulant of order greater than 2. Geary (1942, 1943) shows that if $\kappa\ (c_1,c_2)$ is the bivariate cumulant of y_1 and y_2 of order c_1,c_2 then $\kappa(c_1,c_2+1) = \beta\kappa(c_1+1,c_2)$, so the ratio of the corresponding k-statistics is a consistent estimator of β provided that $\kappa(c_1+1,c_2) \neq 0$. In the choice of cumulants to be used we have to bear in mind that the variance of the estimators increases rapidly with order and that odd-order cumulants are zero for a symmetric distribution, so it seems best to choose $c_1 = 1$, $c_2 = 2$ or $c_1 = 2$, $c_2 = 1$. The method breaks down when the distribution of (y_1,y_2) is normal because all cumulants of order 3 or above are then zero.

1.5.3. Grouping methods

Wald's (1940) two-group method for estimating a linear relationship is very simple in principle. The n observations are divided into two equal groups and the line estimated by joining the centroids of the two groups. The errors $\varepsilon_1,\varepsilon_2$ are taken to be independent with finite variances and the X_{1j} may be random or mathematical. Wald's first condition for the method to be consistent is that, writing $\bar{X}_1'$, $\bar{X}_1''$ for the means of the X_{1j} in the two groups,

$$\lim_{n\to\infty}\ \inf.\ |\bar{X}_1' - \bar{X}_1''| > 0$$

if the X_{1j} have fixed values or that the same inequality hold in probability if the X_{1j} are random variables; the second is that points are allocated to the two groups in such a way that the distributions of ε_1 and ε_2 are unaffected.

The difficulty of the method lies in ensuring that the conditions are satisfied. Sometimes there is some extra information available for grouping. For example, in growth studies we could group according to age, but in this case we might be able to make better use of the time variable by regressing the other variables on it.

Bartlett (1949) shows that where it is possible to apply the grouping method, its efficiency may be improved by dividing the data into three groups and discarding the observations in the centre group.

1.5.4. Instrumental variables

An instrumental variable is some variable U which is known to be correlated with the unobservable true value X_1 but not with the departures ε_1, ε_2 from X_1 and X_2. The estimator

$$b = \sum_j (y_{2j} - \bar{y}_2)u_j \Big/ \sum (y_{1j} - \bar{y}_1)u_j \qquad (1.11)$$

tends to β under these assumptions. The efficiency of the method depends on the correlation between U and X_1; in the best case U and X_1 are linearly related and (1.11) is obviously justified as the ratio of the regression slopes of Y_2 and Y_1 on U. Geary (1949) assumes a trivariate normal distribution for Y_1, Y_2 and U.

The grouping methods may be seen as special cases of (1.11) : Wald's two-group estimator corresponds to U taking the values ± 1 in equal numbers; Bartlett's three-group method is obtained when U takes values -1, 0 and 1.

Barnett (1969) points out that if we are considering a structural relationship and U is also assumed to have a linear structural relationship with Y_1, then we are in the identifiable case of 3 variates and 2 relationships.

2. MULTIVARIATE RELATIONSHIPS

In this chapter we define and provide alternative systems of parameters for multiple functional and structural relationships between any number of random variables, and go on to give some examples, some connexions with other statistical models and a review of the methods proposed for estimating such relationships.

The essential features of a set of linear functional or structural relationships are that we observe a p-vector of random variables $\underline{Y}$ which may be expressed as the sum of a 'true' p-vector $\underline{X}$, whose elements are exactly linearly related, and a 'departure' or 'error of measurement' $\underline{\varepsilon}$. The relationships are structural if $\underline{X}$ is regarded as a random vector over some population; if the components of $\underline{X}$ are mathematical variables taking different values for each observation (or, with replication, for each group of observations) then the relationships are functional.

If the linear relationships between the components of $\underline{X}$ are $p-r$ in number then $\underline{X}$ is confined to an r-dimensional hyperplane which may be defined by specifying one point on the surface and either $p-r$ linearly independent vectors orthogonal to the hyperplane or r linearly independent vectors parallel to it. (See, for example, Sprent,1968,and Kendall,1961,pp. 6-7). To these two ways of specifying a hyperplane correspond equivalent representations of functional and structural relationships: the first is the more appropriate to the term 'relationships'; the second — because of its symmetry — is more convenient when problems of estimation are to be considered. The two representations are given below, first for structural and then for functional relationships.

2.1 Representing structural relationships

We suppose that the observed vector $\underline{Y}$ is given by

$$\underline{Y} = \underline{X} + \underline{\varepsilon}, \tag{2.1}$$

where $\underline{X}$ is a random p-vector whose distribution is degenerate and $\underline{\varepsilon}$ is a random p-vector of departures from $\underline{X}$, usually taken to be independent of $\underline{X}$. The vector $\underline{X}$ is confined to a hyperplane of dimension r and its mean vector $\underline{\mu}$ must lie in this hyperplane. The $p-r$ linear structural relationships are $p-r$ independent homogeneous linear constraints on $\underline{X}-\underline{\mu}$. In what we shall call the *Type I* representation the constraints are given by

$$\underline{\Gamma}'(\underline{X}-\underline{\mu}) = \underline{0}, \tag{2.2}$$

where $\underline{\Gamma}$ is a $p \times (p-r)$ matrix of full rank. The columns of $\underline{\Gamma}$ span the subspace orthogonal to $\underline{X}-\underline{\mu}$.

A convenient way to make $\underline{\Gamma}$ unique is to assume that the first r elements of $\underline{X}$ are linearly independent (this may require some reordering) and express the remaining $p-r$ in terms of them. Thus

$$\underline{X}_b - \underline{\mu}_b = \underline{A}(\underline{X}_a - \underline{\mu}_a), \tag{2.3}$$

where $\underline{X}_a$ consists of the first r elements of $\underline{X}$ and $\underline{X}_b$ of the remaining $p-r$, $\underline{\mu}$ is partitioned similarly, and $\underline{A}$ is a $(p-r) \times r$ matrix. That is, $\underline{\Gamma}'$ is required to have the form

$$\underline{\Gamma}' = (-\underline{A}\ \underline{I}_{p-r}). \tag{2.4}$$

Equation (2.3) may also be written

$$\underline{X}_b = \underline{A}\,\underline{X}_a + \underline{\alpha}, \tag{2.5}$$

where $\underline{\alpha} = \underline{\mu}_b - \underline{A}\underline{\mu}_a$; this may be appropriate if there is some asymmetry between $\underline{X}_a$ and $\underline{X}_b$.

If the hyperplane is constrained to pass through the origin then equations (2.2) and (2.3) become respectively $\underline{\Gamma}'\underline{X} = \underline{0}$ and $\underline{X}_b = \underline{A}\underline{X}_a$; the same relationships apply to $\underline{\mu}$.

In the *Type II* representation $\underline{X}-\underline{\mu}$ is expressed as a linear combination, with random coefficients, of r linearly independent p-vectors. That is

$$\underline{X} = \underline{\mu} + \underline{\Lambda}\,\underline{f}, \tag{2.6}$$

where $\underline{\Lambda}$ is $p \times r$ and of full rank and $\underline{f}$ is an r-vector of linearly independent random variables with zero means. The columns of $\underline{\Lambda}$ span the subspace containing $\underline{X} - \underline{\mu}$. We may put $\underline{X}$ in the form (2.6) if and only if its dispersion matrix has rank r. For if (2.6) holds the dispersion matrix is $\underline{\Lambda}\,D(\underline{f})\,\underline{\Lambda}'$, which has rank r; conversely, if $D(\underline{X})$ has rank r it has r linearly independent columns and the corresponding r elements of $\underline{X}$ are linearly independent. The remaining $p - r$ elements are linear functions of them, so these r elements minus their means serve as $\underline{f}$.

Comparing equations (2.2) and (2.6) and using the linear independence of the elements of $\underline{f}$, we find that the two representations are related by the equation

$$\underline{\Gamma}'\underline{\Lambda} = \underline{0}. \tag{2.7}$$

If $\underline{\Gamma}'$ has the form (2.4) and $\underline{\Lambda}'$ is partitioned as $(\underline{\Lambda}_1' \; \underline{\Lambda}_2')$ with $\underline{\Lambda}_1$ $r \times r$ then

$$\underline{A}\,\underline{\Lambda}_1 = \underline{\Lambda}_2, \tag{2.8}$$

and this is independent of any postmultiplication of $\underline{\Lambda}$ by a nonsingular matrix.

Without loss of generality the components of $\underline{f}$ in (2.6) may be taken to be uncorrelated with unit variances by, say, postmultiplying $\underline{\Lambda}$ by the symmetric square root of the dispersion matrix of $\underline{f}$. The matrix $\underline{\Lambda}$ is then unique up to postmultiplication by an orthogonal $r \times r$ matrix; the product $\underline{\Lambda}\,\underline{\Lambda}'$ is unique.

If the 'error' vectors $\underline{\varepsilon}$ corresponding to each observation are uncorrelated with zero means and a common positive-definite dispersion matrix $\underline{\Sigma}$, then we may conveniently define a unique $\underline{\Lambda}$ as follows: let $\underline{\Theta} = \underline{\Lambda}\,\underline{\Lambda}'$ and let $\omega_1, \ldots, \omega_r$ be the nonzero roots in descending order of

$$|\underline{\Theta} - \omega \underline{\Sigma}| = 0 . \tag{2.9}$$

If $\underline{\Omega} = \text{diag}(\omega_1, \ldots, \omega_r, 0, \ldots, 0)$ is $p \times p$ then there is a nonsingular $p \times p$ matrix $\underline{N}_{\underline{\Theta}}$ such that

$$\underline{\Sigma} = \underline{N}_{\underline{\Theta}} \underline{N}'_{\underline{\Theta}} , \quad \underline{\Theta} = \underline{N}_{\underline{\Theta}} \underline{\Omega} \underline{N}'_{\underline{\Theta}} = \underline{N}_{\underline{\Theta}1} \underline{\Omega}_1 \underline{N}'_{\underline{\Theta}1} \tag{2.10}$$

where $\underline{N}_{\underline{\Theta}1}$ consists of the first r columns of $\underline{N}_{\underline{\Theta}}$ and $\underline{\Omega}_1 = \text{diag}(\omega_1, \ldots, \omega_r)$ (Rao, 1973, p.41). If the positive roots of (2.9) are distinct then $\underline{N}_{\underline{\Theta}1}$ is unique up to changes of sign of its columns, and a unique $\underline{\Lambda}$ is given by

$$\underline{\Lambda}_o = \underline{N}_{\underline{\Theta}1} \text{ diag}(\omega_1^{\frac{1}{2}}, \ldots, \omega_r^{\frac{1}{2}}).$$

With this choice of $\underline{\Lambda}$ we have

$$\underline{\Lambda}'_o \underline{\Sigma}^{-1} \underline{\Lambda}_o = \underline{\Omega}_1 . \tag{2.11}$$

Thus we have chosen $\underline{\Lambda}_o$ so that $\underline{\Lambda}'_o \underline{\Sigma}^{-1} \underline{\Lambda}_o$ is a diagonal matrix whose positive elements are the nonzero roots of (2.9) arranged in descending order.

A set of linear structural relationships constrained to pass through the origin is given by

$$\underline{X} = \underline{\Lambda} \underline{f}^* ,$$

where $\underline{f}^*$ is a random r-vector with unit dispersion matrix and mean $\underline{\nu}$, not necessarily zero. Equivalently, we demand that $\underline{\mu}$ in equation (2.6) lie in the range space of $\underline{\Lambda}$, that is, that $\underline{\mu} = \underline{\Lambda} \underline{a}$ for some r-vector $\underline{a}$. Since $\underline{\Lambda}$ and $\underline{\Lambda}\underline{\Lambda}'$ have the same range space (Rao, 1973, p.27) the condition is also that $\underline{\mu} = \underline{\Theta} \underline{c}$ for some p-vector $\underline{c}$.

If we wish to assume that $\underline{X}$ follows a p-variate normal distribution then $\underline{f}$ may be taken to be distributed as $N_r(\underline{0}, \underline{I})$. This is equivalent to the assumption that $\underline{X}$ has the distribution $N_p(\underline{\mu}, \underline{\Theta})$, where $\underline{\Theta}(= \underline{\Lambda}\underline{\Lambda}')$ has rank r (Rao, 1973, p.521).

2.2 Representing functional relationships

We again assume that equation (2.1) holds, except that we write

$\underline{\xi}$ rather than $\underline{X}$ for the 'true' values since they are to be considered as unknown parameters. Rather than refer to a population of values of $\underline{\xi}$ we consider the individual $\underline{\xi}_j$ $(j = 1, \ldots, n)$ corresponding to $n(>r+1)$ observations (or groups of replicate observations) on $\underline{Y}$. The $\underline{\xi}_j$ are subject to $p-r$ functional relationships if there exists a point $\underline{\xi}_o$ and a $p \times (p-r)$ matrix $\underline{\Gamma}$ of full rank such that

$$\underline{\Gamma}'(\underline{\xi}_j - \underline{\xi}_o) = \underline{0} \qquad (j = 1, \ldots, n). \tag{2.12}$$

By analogy with Section 2.1 we shall call this the *Type I* representation of functional relationships. The point $\underline{\xi}_o$ may be anywhere on the hyperplane containing the $\underline{\xi}_j$. For the sake of uniqueness we may take $\underline{\xi}_o$ to be $\bar{\underline{\xi}}$, the mean of the $\underline{\xi}_j$, defined as $n^{-1} \Sigma \underline{\xi}_j$ or, if there are m_j replicate observations at $\underline{\xi}_j$ $(j = 1, \ldots, n)$, as $\Sigma m_j \underline{\xi}_j / \Sigma m_j$. (Note that we are assuming an equal number of replications, m_j, on each component of $\underline{Y}$ corresponding to $\underline{\xi}_j$).

As before we may define a unique $\underline{\Gamma}$ by giving it the form of (2.4) (possibly after some re-ordering of the elements of $\underline{\xi}$), so that

$$\underline{\xi}_{bj} - \bar{\underline{\xi}}_b = \underline{A}(\underline{\xi}_{aj} - \bar{\underline{\xi}}_a) \qquad (j = 1, \ldots, n), \tag{2.13}$$

where $\underline{\xi}_{aj}$ consists of the first r elements of $\underline{\xi}_j$ and $\underline{\xi}_{bj}$ of the remaining $p-r$ and $\bar{\underline{\xi}}$ is partitioned similarly. This may be rewritten as

$$\underline{\xi}_{bj} = \underline{A}\ \underline{\xi}_{aj} + \underline{\alpha} \qquad (j = 1, \ldots, n), \tag{2.14}$$

where $\underline{\alpha} = \bar{\underline{\xi}}_b - \underline{A}\ \bar{\underline{\xi}}_a$. For a hyperplane through the origin (2.12) and (2.13) become respectively $\underline{\Gamma}' \underline{\xi}_j = \underline{0}$ $(j = 1, \ldots, n)$ and $\underline{\xi}_{bj} = \underline{A}\, \underline{\xi}_{aj}$ $(j = 1, \ldots, n)$.

In the *Type II* representation of a set of functional relationships we assume that there exists a point $\underline{\xi}_o$, a $p \times r$ matrix $\underline{\Lambda}$ of full rank, and n r-vectors $\underline{\phi}_j$ $(j = 1, \ldots, n)$ such that

$$\underline{\xi}_j = \underline{\xi}_o + \underline{\Lambda}\, \underline{\phi}_j \qquad (j = 1, \ldots, n). \tag{2.15}$$

Corresponding to the linear independence of the elements of $\underline{f}$ in (2.6) we here require that there is no vector $\underline{a}$ such that $\underline{a}'\underline{\phi}_j = 0$ for all j. Thus the rows of the $r \times n$ matrix $(\underline{\phi}_1 \ \ldots \ \underline{\phi}_n)$ must be linearly independent. Comparing equations (2.12) and (2.15) and using the linear independence of the rows of $(\underline{\phi}_1 \ \ldots \ \underline{\phi}_n)$ we find that equations (2.7) and (2.8) are also satisfied for functional relationships.

Equation (2.15) is unchanged by the addition of a constant vector, $\underline{b}$ say, to each $\underline{\phi}_j$ and the subtraction of $\underline{\Lambda}\underline{b}$ from $\underline{\xi}_o$. This redundancy in the parameters may be removed by defining $\bar{\underline{\phi}}$ in the same manner as $\bar{\underline{\xi}}$ and setting the former equal to zero, so that $\underline{\xi}_o$ becomes $\bar{\underline{\xi}}$. Another source of redundancy is that postmultiplication of $\underline{\Lambda}$ by any nonsingular $r \times r$ matrix and premultiplication of each $\underline{\phi}_j$ by the inverse also leaves (2.15) unchanged. We could, by analogy with the condition $D(\underline{f}) = \underline{I}_r$ in the structural case, apply restrictions to the $\underline{\phi}_j$ such as

$$\Sigma m_j \underline{\phi}_j \underline{\phi}_j' = \Sigma m_j \underline{I}_r , \qquad (2.16)$$

so that the arbitrariness is reduced to an orthogonal transformation, but this has the disadvantage that $\underline{\Lambda}$ depends on the particular values of the $\underline{\phi}_j$. A particular problem is that if we wished to consider the consistency of a set of estimators in an unreplicated case under various types of limiting behaviour of the $\underline{\xi}_j$ we might assume $\Sigma(\underline{\xi}_j - \bar{\underline{\xi}})(\underline{\xi}_j - \bar{\underline{\xi}})'$ (which equals $\underline{\Lambda}\Sigma\underline{\phi}_j\underline{\phi}_j'\underline{\Lambda}'$) to be of order n^s $(s > 0)$. It would be better for $\Sigma\underline{\phi}_j\underline{\phi}_j'$, a function of the incidental parameters alone, to share this behaviour than for $\underline{\Lambda}$ to converge to zero or diverge.

Instead of imposing conditions such as (2.16) on the $\underline{\phi}_j$, we shall follow, and make slightly more precise, the development given by Rao (1973, p.559), which is mathematically similar to the device used to define a unique $\underline{\Lambda}$ in Section 2.1. We again assume that the vectors

of departures $\underline{\epsilon}_j$ are uncorrelated with zero means and common positive-definite dispersion matrix $\underline{\Sigma}$. Rao employs a linear transformation $\underline{Y} \to \underline{\Sigma}^{-\frac{1}{2}} \underline{Y}$ in order to give the vector of departures unit dispersion. However, it is not clear whether the symbol $\underline{\Sigma}^{\frac{1}{2}}$ is intended to denote the symmetric square root of $\underline{\Sigma}$ or merely a matrix $\underline{N}$ satisfying

$$\underline{\Sigma} = \underline{N}\underline{N}' ; \tag{2.17}$$

any solution of (2.17) would have the required property. We shall consider generally the transformation to $\underline{U} = \underline{N}^{-1} \underline{Y}$ and $\underline{\zeta} = \underline{N}^{-1} \underline{\xi}$ for any solution of (2.17) and then choose a particularly convenient solution. The $\underline{\zeta}_j$ have the same structure as (2.15), say

$$\underline{\zeta}_j = \bar{\underline{\zeta}} + \underline{S}\, \underline{\phi}_j \qquad (j = 1, \ldots, n) \tag{2.18}$$

where $\underline{S} = \underline{N}^{-1} \underline{\Lambda}$. Without loss of generality we may take the columns of $\underline{S}$ to be orthonormal, that is, choose $\underline{S}$ to satisfy

$$\underline{S}'\underline{S} = \underline{I}_r . \tag{2.19}$$

For a given $\underline{N}$ this defines $\underline{S}$, and therefore $\underline{\Lambda}$, to within postmultiplication by an orthogonal $r \times r$ matrix, and this remaining arbitrariness causes little difficulty in estimation since $\hat{\underline{\Lambda}}\, \hat{\underline{\phi}}_j$ or $\hat{\underline{\xi}}_j$ is of more interest than $\hat{\underline{\phi}}_j$ itself. In terms of $\underline{\Lambda}$, equation (2.19) becomes

$$\underline{\Lambda}'\underline{\Sigma}^{-1}\underline{\Lambda} = \underline{I}_r , \tag{2.20}$$

in contrast to (2.11).

If $\underline{\Sigma}$ is a general positive-definite $p \times p$ matrix then it is convenient, when considering the estimation of $\underline{\Lambda}$ and $\underline{\phi}_j$, to define a particular solution of (2.17) in terms of the simultaneous reduction of $\underline{\Sigma}$ and the between-groups SSP matrix $\underline{B}$ defined by

$$\underline{B} = \Sigma(\underline{y}_j - \bar{\underline{y}})(\underline{y}_j - \bar{\underline{y}})' \tag{2.21}$$

or, if there is replication, by

$$\underline{B} = \Sigma\, m_j(\bar{\underline{y}}_j - \bar{\underline{y}})(\bar{\underline{y}}_j - \bar{\underline{y}})' . \tag{2.22}$$

Let the equation

$$|\underline{B} - \delta \underline{\Sigma}| = 0 \qquad (2.23)$$

have roots $\delta_i(\underline{B})$ $(i = 1, \ldots, p)$ in descending order and let $\underline{\Delta}_{\underline{B}} = \text{diag}\{\delta_1(\underline{B}), \ldots, \delta_p(\underline{B})\}$; then there exists a solution $\underline{N}_{\underline{B}}$ of (2.17) which also satisfies

$$\underline{B} = \underline{N}_{\underline{B}} \, \underline{\Delta}_{\underline{B}} \, \underline{N}_{\underline{B}}' \; . \qquad (2.24)$$

If $n > p$, then with probability one $\underline{B}$ is non-singular, the roots are distinct and $\underline{N}_{\underline{B}}$ is unique up to changes of sign of its columns. The choice of this solution of (2.17) is convenient because the within-groups SSP matrix of $\underline{U}$ is diagonal and because the estimators of $\underline{\Lambda}$ involve $\underline{N}_{\underline{B}}$.

If the departures from different elements of $\underline{\xi}$ are uncorrelated, so that $\underline{\Sigma}$ is diagonal with nonzero elements σ_{ii} $(i = 1, \ldots, p)$, then an obvious alternative to the above transformation is to use

$$\underline{N}_d = \text{diag}(\sigma_{11}^{\frac{1}{2}}, \ldots, \sigma_{pp}^{\frac{1}{2}}). \qquad (2.25)$$

A set of linear relationships through the origin is given by

$$\underline{\xi}_j = \underline{\Lambda} \, \underline{\phi}_j^* \quad (j = 1, \ldots, n), \qquad (2.26)$$

with the same condition of linear independence on the $\underline{\phi}_j^*$ as on the $\underline{\phi}_j$ in (2.15). The same method of defining $\underline{\Lambda}$ to within postmultiplication by an orthogonal matrix applies here as before, except that $\underline{B}$ is replaced by the raw between-groups SSP matrix

$$\underline{B}^* = \Sigma \, \underline{y}_j \, \underline{y}_j' \qquad \text{or} \qquad \Sigma \, m_j \, \bar{\underline{y}}_j \, \bar{\underline{y}}_j' \; ; \qquad (2.27)$$

the condition $n > p$ for $\underline{B}$ to be nonsingular (a.s.) becomes $n \geq p$ for $\underline{B}^*$.

2.3 Some examples of multivariate relationships

We give here some examples of situations in which $\underline{X}$, or $\underline{\xi}$, has in general more than two elements, using our own notation wherever it conflicts with the authors'. Of the two representations of functional and structural relationships given above, Type I is the more appropriate when there is some asymmetry among the elements of $\underline{X}$ or $\underline{\xi}$.

(a) Comparative calibration

We return to the problem introduced in Section 1.1 of a calibration

experiment in which the same property is measured on each of n specimens using p measuring instruments. In Barnett's (1969) formulation, each reading is made up of a normally distributed 'true' value X_i and an independent error distributed as $N(0, \sigma_i^2)$ $(i = 1, \ldots, p)$; all errors are independent and the true measurements on any pair of instruments are linearly related. He has

$$X_i = \alpha_i + \beta_i X_1 \qquad (i = 2, 3, \ldots, p); \tag{2.28}$$

instruments 2, 3, ..., p are each related to instrument 1, which may be chosen arbitrarily. This is a Type I representation of $p - 1$ structural relationships in the form of equation (2.5) with $r = 1$, $\underline{\alpha} = (\alpha_2, \ldots, \alpha_p)'$ and $\underline{A} = (\beta_2, \ldots, \beta_p)'$. However, the symmetry of the relationships between the instruments may be expressed by supposing that each true measurement is a linear function of a hypothetical standard measurement f, distributed as $N(0, 1)$ over the population of possible specimens. Thus we may write

$$\underline{X} = \underline{\mu} + \underline{\lambda} f \ , \tag{2.29}$$

where $\underline{\mu} = E(\underline{X})$ and $\underline{\lambda}$ is a p-vector of unknown 'calibration factors'; this is of Type II. The only condition required for the parameters to be unique is on the signs of the λ_i; the model is sensible only if each λ_i is positive.

Williams' (1969, 1973) functional-relationship version of the calibration problem has a Type II representation; his model for the vector of measurements on the j^{th} specimen amounts to

$$\underline{\xi}_j = \underline{\xi}_o + \underline{\lambda} \phi_j \qquad (j = 1, \ldots, n). \tag{2.30}$$

In the earlier paper the ϕ_j are regarded as the values of an unknown standard measurement and are subject to the constraints

$$\Sigma \phi_j = 0 \ , \quad \Sigma \phi_j^2 = 1.$$

The first condition sets $\underline{\xi}_o$ equal to $\bar{\underline{\xi}}$ and the second is similar to (2.16). As in Barnett's model, all errors are taken to be independent,

those on the i^{th} instrument being distributed as $N(0, \sigma_i^2)$ $(i = 1, \ldots, p)$. The suggested maximum likelihood solution when the σ_i are unknown is unworkable, as we explain in Section 2.4(a). In the later paper restrictions are applied to the structural rather than the incidental parameters. They are

$$\Sigma \lambda_i^2/\sigma_i^2 = 1 \quad , \quad \Sigma \mu_i \lambda_i/\sigma_i^2 = 0 \ ,$$

the first being (2.20) with $r = 1$.

Taylor (1973) considers a calibration problem in which several instruments are compared with a standard. One variable, Y_o, has a special position and we are interested in the functional relationships, assumed linear, between each of the others, Y_i $(i = 1, \ldots, q)$, and Y_o. All errors of observation are taken to be independent, and there may be unequal numbers of replications on the different instruments, a complication we shall ignore when considering the estimation of functional relationships.

(b) Homogeneous strain in a glacier

Gleser and Watson (1973) consider a problem in which markers are placed in a glacier at points $\underline{\gamma}_1, \ldots, \underline{\gamma}_n$. Here the elements of the 3-vector $\underline{\gamma}_j$ are the Cartesian co-ordinates of the j^{th} marker. The positions are observed with error and recorded as $\underline{u}_1, \ldots, \underline{u}_n$. A later survey, using similar techniques, yields measurements $\underline{v}_j$ of the true positions $\underline{\delta}_j$ reached by the markers $(j = 1, \ldots, n)$. If the deformation in the meantime is merely a homogeneous strain then $\underline{\delta}_j = \underline{A}\,\underline{\gamma}_j$ $(j = 1, \ldots, n)$, where $\underline{A}$ is 3×3. This is a Type I representation of linear functional relationships through the origin with $p = 6$, $r = 3$, $\underline{\xi}_j' = (\underline{\gamma}_j', \underline{\delta}_j')$, and $\underline{y}_j' = (\underline{u}_j', \underline{v}_j')$; the other representation would hardly be appropriate. The authors assume that $\underline{u}_j$ and $\underline{v}_j$ are independently distributed as $N_3(\underline{\gamma}_j, \tau \underline{\Upsilon})$ and $N_3(\underline{\delta}_j, \tau \underline{\Upsilon})$ respectively, where $\underline{\Upsilon}$ is a known positive-definite matrix but τ is unknown. They state that 'obvious' linear transformations

$\underline{u} \to \underline{T}^{-\frac{1}{2}} \underline{u}$ and $\underline{v} \to \underline{T}^{-\frac{1}{2}} \underline{v}$ reduce the problem to the case in which $\underline{T} = \underline{I}$, but do not say what is meant by the notation $\underline{T}^{\frac{1}{2}}$.

(c) Growth and size allometry

Sprent (1972) defines allometry as the study of differences in the shape of organisms associated with size. Individual organisms change their size and shape with time, but there are also differences between members of a species with the same age. These two aspects of the subject are known respectively as growth and size allometry. Suppose that $z_1, \ldots, z_p$ are measurements of the size of different parts of an organism. For the study of just two characteristics a widely-used model is the *simple allometry equation* introduced by Huxley (1924); the measurements z_1, z_2 are assumed to satisfy, at least approximately, the relationship

$$z_1 = k z_2^{\beta}$$

or, equivalently,

$$\log z_1 = \log k + \beta \log z_2. \tag{2.31}$$

Some authors use the term *allometry* for this relationship. When we come to consider all p characteristics simultaneously a problem arises as to the appropriate generalization of equation (2.31). According to Hopkins (1966) it is the set of $p-1$ relationships

$$\log z_1 = \log k_2 + \beta_2 \log z_2 = \ldots = \log k_p + \beta_p \log z_p.$$

This is equivalent to

$$\log z_i = \mu_i + \lambda_i f \quad (i = 1, \ldots, p), \tag{2.32}$$

where $\beta_i = \lambda_1/\lambda_i$ and f is a general size factor. Thus, apart from random biological variation and measurement errors, the logarithms of the z_i lie on a line in p dimensions. Hopkins applies this model to some size data on the weights of nine organs of rats. He first screens the measurements for non-linearity in their bivariate relationships and deletes three of them; a structural-relationship model with independent departures is found to give an adequate fit to

the remaining six. However, the method of fitting the line is somewhat suspect, in particular because it is not scale-invariant.

Teissier (1955) has also considered the problem of fitting a line to the logarithms of the dimensions of organisms. He discusses the use of the first principal component as well as a set of $p-1$ structural relationships with independent departures, and decides in favour of the latter, deriving an explicit solution when $p=3$. As an example he applies this model to some data on eight measurements of the dimensions of the crab Oxyrhynchid.

The assumptions made by these two authors about the logarithms of size measurements are equivalent to those made in Example (a) about different measurements of the same property. In the case of biological variation, however, it is doubtful whether the hypothesis of independent departures between different measurements is appropriate. The assumption can be checked when there are replicate observations.

Sprent (1968) agrees with Hopkins that the proper generalization of the simple allometry equation is to a line in p dimensions, but considers that hyperplanes of higher dimension are also worth studying. In cases where Σ is known or estimable he recommends testing for $p-1$ linear relationships, then $p-2$, and so on until a significant result is obtained. However, he shows that in growth allometry if each characteristic has a growth curve which can be expressed as a polynomial of order r in time then there will be $p-r$ approximate linear relationships between the p characteristics; this justifies the restriction of the simple allometry concept to a line rather than merely any hyperplane of dimension less than p.

Regarding assumptions about departures from the hyperplane, he argues (Sprent,1972) that in growth allometry longitudinal studies involving repeated measurements on the same organism may lead to autocorrelated

departures, so it is advantageous to measure different organisms at each time. Thus the data he presents in the 1968 paper on the wood growth, girth increment and reproductive growth of apple trees are observations on different trees in each of five years. He considers a functional-relationship model in which the mean vectors $\underline{\xi}_j$ for each year are subject to one or two linear constraints, and is able to estimate the dispersion matrix of the departures by measuring six trees each year. The data are found to be consistent with a single linear relationship between the logarithms of the characteristics, but not with a simple allometry relationship which would involve a second equation. The relationship is estimated by an extension of the generalized least-squares method of Section 1.5 to be described in Section 2.5. In this analysis the years are used only for grouping the data; the precise values taken by the time variable and even their order are ignored. Bartlett (1948, 1966) argues that with data of this type we should carry out a canonical analysis of the logarithms of the characteristics with years, the regressor variables being, for example, polynomials in time, and the small canonical roots being relevant to any linear functional relationships. If the number of regressor variables is given an upper limit then there are no incidental parameters, and in the absence of replication the dispersion matrix of the departures can be estimated from the residuals after removing the systematic variation. Clearly, we should expect Bartlett's analysis to be the more efficient since it makes greater use of the available data. On the other hand, it depends on being able to represent satisfactorily the growth curves of the characteristics with the chosen set of regressor variables.

(d) Relations between economic time series

Attempts have been made to determine fundamental relationships

between economic time series by assuming that linear functional relationships exist between sets of economic variables which are all subject to random disturbances. The hope is that such relationships may be interpreted as, for instance, demand functions, supply functions, or production functions (Tintner, 1952, p.121). Some of the models proposed ignore the order in which observations are taken, since they allow neither temporal dependence of the 'true values' of the variables nor the possibility of autocorrelated departures. They are functional relationship models similar mathematically to Sprent's model for growth allometry, and the criticisms made by Bartlett apply no less forcibly here. Other objections are that the dispersion matrix of the departures has to be assumed since replication is seldom possible in economic series, that lagged relationships between the true values cannot be incorporated, and that the assumption of zero autocorrelations may not be realistic. Nevertheless, the methods of estimation developed for such models are of interest since they apply to functional relationship models employed in more appropriate contexts.

Koopmans (1937) applies the method of maximum likelihood to the problem of estimating a single functional relationship when the vector of departures is normally distributed with dispersion matrix known to within a constant factor and there is no replication. As in the bivariate case the constant of proportionality is not consistently estimated because of the presence of incidental parameters.

Under the title 'weighted regression' Tintner (1945) considers the estimation of $p-r$ linear functional relationships when there is no replication but the vector of departures is normal with known dispersion matrix. The method is applied (Tintner, 1946) in an attempt to find a demand and a supply function for agricultural products in

the United States of America. A curious aspect of the analysis is that time is included as one of the variables 'subject to error'. The error variances are estimated by Tintner's 'variate-difference' method.

Bartlett (1948), in criticizing Tintner's approach to the analysis of economic time series, draws attention to the ambiguous role played by time, which is sometimes brought in as a dependent variable but is tacitly assumed to be a regressor variable in the variate-difference method. He re-analyses some data on the supply and demand for cotton yarn in the years 1924-38 to which Tintner's method had been applied by Lomax (1948), fitting polynomials up to the fifth degree in time to represent the systematic variation in the four variables considered, and treating the remaining variability as error. There is no assumption of independence of errors between variables. A canonical analysis of the total and residual CSSP matrices yields two non-significant latent roots. Equating to zero the corresponding canonical variables gives relationships which, when rearranged, may be interpreted as demand and supply equations. The series is, however, too short for reliable estimation of the coefficients.

2.4 Connexions with other models

Some of the statistical models that have been proposed in the field of psychological testing are special cases of the Type II representation of functional and structural relationships. When we come to consider the estimation of such relationships we shall note the connexions between the estimation procedures and the methods of canonical analysis and principal component analysis. However, we are concerned here with statistical models rather than techniques.

(a) Factor analysis

Equation (2.6) expresses the 'true' values of variates connected by a set of linear structural relationships as a linear combination

with random coefficients of r linearly independent p-vectors. If the elements of the vector $\underline{f}$ are assumed to follow, independently, a standard normal distribution and the departures are mutually independent, then we have the standard model of factor analysis, except that it is usual to ignore $\underline{\mu}$ and subtract the sample mean from each observation (Lawley and Maxwell, 1971, p.6). The columns of $\underline{\Lambda}$ are factor loadings and the elements of $\underline{f}$ represent an individual's score on each factor. It is usual to assume that replicate observations are not available, because in many types of psychological experiment the subject would remember what he had done on previous occasions.

A comparison of the numbers of parameters and minimal sufficient statistics suggests that the model is identifiable if

$$(p - r)^2 \geq p + r \; ; \qquad (2.33)$$

there are $p(r + 2)$ parameters in $\underline{\mu}$, $\underline{\Lambda}$ and $\underline{\Sigma}$ subject to $\frac{1}{2}r(r - 1)$ independent conditions given by (2.10), while the number of minimal sufficient statistics is $\frac{1}{2}p(p + 3)$. The method used in Section 2.1 for defining a unique $\underline{\Lambda}$ is a generalization to non-diagonal $\underline{\Sigma}$ of that used in factor analysis (Lawley and Maxwell, 1971, pp. 7,8). The postmultiplication of $\underline{\Lambda}$ by an orthogonal matrix which condition (2.10) excludes corresponds to a rotation of the factors.

Another model of factor analysis which used to be considered takes each individual's factor scores to be unknown parameters, and is thus equivalent to the Type II representation of a set of functional relationships (equation 2.15). Lawley's (1941) Method II of factor analysis is an attempt to apply the method of maximum likelihood to this model, but he observes that his iterative procedures either do not converge or tend to unacceptable solutions in which one or more of the error variances vanish. Kendall (1950) questions whether the method is capable of giving a satisfactory solution in such cases. In the discussion of

Kendall's paper, Bartlett (1950) points out the connexion with structural and functional relationships and shows that the breakdown of Lawley's Method II follows from the failure of maximum likelihood in estimating functional relationships without replication. Anderson and Rubin (1956) show that the likelihood for this model is unbounded, one of the points made by Solari (1969) in the case $p = 2$. Williams' (1969) suggested solution to the problem of comparative calibration with unknown error variances (Section 2.3(a)) must be unworkable because it is Lawley's Method II restricted to a single common factor.

Madansky (1964) applies the method of instrumental variables to factor analysis by rewriting the factor model as a set of structural relationships. Lawley and Maxwell (1973) use the factor model in carrying out multiple linear regression when there are independent errors of measurement in the regressor variables.

(b) Covariance structures

Jöreskog (1970, 1973) has applied the method of maximum likelihood to sets of independent multivariate normal variables with a very general structure. Here we shall follow his notation even where it conflicts with our own, because of its complexity and the paucity of the Greek Alphabet. He considers an $N \times p$ data matrix $\underline{X}$ whose rows are independently normally distributed with a common dispersion matrix of the form

$$\underline{B}(\underline{\Lambda}\,\underline{\Phi}\,\underline{\Lambda}' + \underline{\Psi}^2)\underline{B}' + \underline{\Theta}^2 \,, \tag{2.34}$$

and whose mean vectors are given by

$$E(\underline{X}) = \underline{A}\,\underline{\Xi}\,\underline{P} \,. \tag{2.35}$$

Here $\underline{A}$ is an $N \times g$ matrix with $g \leqslant N$ and $\underline{P}$ is an $h \times p$ matrix with $h \leqslant p$, both being fixed and of full rank. The matrices $\underline{\Xi}$, $\underline{B}$, $\underline{\Lambda}$, the symmetric matrix $\underline{\Phi}$ and the diagonal matrices $\underline{\Psi}$ and $\underline{\Theta}$ are parameter matrices. Their nonzero elements are allowed to be of three types:

(i) fixed elements; (ii) constrained parameters that are unknown but equal to one or more other parameters; (iii) free parameters that are unknown and not constrained to be equal to any other parameter. There is clearly a great deal of indeterminacy in the general model; restrictions must be imposed to make the parameters indentifiable.

Among the many special cases of the model are multivariate regression, multivariate analysis of variance and some of its generalizations, and factor analysis. It follows that unreplicated structural relationships with independent errors may be estimated provided that condition (2.33) is fulfilled; replicated structural relationships are excluded, since observations corresponding to the same value of the 'true' random p-vector $\underline{X}$ are correlated. Functional relationships are also excluded since their mean vectors are not of the required form. Despite this, the model is of value to our study for its use of constrained parameters. For example, the model for comparative calibration given by equation (2.29) corresponds to equations (2.34) and (2.35) with $g = 1$, $h = p$, $\underline{A} = (1, \ldots, 1)'$, $\underline{\Xi} = \underline{\mu}'$, $\underline{P} = \underline{B} = \underline{I}$, $\underline{\Lambda} = \underline{\lambda}$, $\underline{\Phi} = 1$, $\underline{\Psi} = \mathrm{diag}(\sigma_1, \ldots, \sigma_p)$, and $\underline{\Theta} = \underline{0}$. The hypothesis that any pair of instruments has true values differing only by a constant corresponds to the constraint that the elements of $\underline{\Lambda}$ are equal. The further constraint that the elements of $\underline{\Xi}$ are equal implies that the true values are identical.

A computer program for the analysis of covariance structures by maximum likelihood is given by Jöreskog et al. (1971); it is based on the function-minimization method of Fletcher and Powell (1963).

2.5 Some methods of estimation

Little appears to have been published on the estimation of multivariate structural relationships beyond the observation that factor analysis techniques can be used when the departures are all independent

and there is no replication. For the case $r = 1$, Barnett (1969) shows that a number of asymptotically unbiased estimators may be obtained by considering the variates three at a time.

In the functional case most of the work on estimation is concerned with a single relationship and a common dispersion matrix for the departures known to within at least a constant factor.

Koopmans (1937) considers maximum likelihood (ML) estimation when the n vectors of observations $\underline{y}_1, \ldots, \underline{y}_n$ are independent and normally distributed with common dispersion matrix $\underline{\Sigma}$ and mean vectors $\underline{\xi}_1, \ldots, \underline{\xi}_n$ subject to the single relationship $\underline{\gamma}'(\underline{\xi}_j - \underline{\xi}_o) = 0$ $(j = 1, \ldots, n)$. The log-likelihood is

$$-\tfrac{1}{2} n \log|\underline{\Sigma}| - \tfrac{1}{2} \sum_j (\underline{y}_j - \underline{\xi}_j)' \underline{\Sigma}^{-1} (\underline{y}_j - \underline{\xi}_j) , \qquad (2.36)$$

so maximization with respect to the $\underline{\xi}_j$ and $\underline{\gamma}$ is equivalent to minimization of

$$\sum_j (\underline{y}_j - \underline{\xi}_j)' \underline{\Sigma}^{-1} (\underline{y}_j - \underline{\xi}_j). \qquad (2.37)$$

He shows that the ML estimator of $\underline{\gamma}$ minimizes the ratio

$$\underline{\gamma}'\underline{B}\underline{\gamma} / \underline{\gamma}'\underline{\Sigma}\underline{\gamma} , \qquad (2.38)$$

where $\underline{B}$ is given by (2.21); the minimum is $\delta_p(\underline{B})$, the smallest root of (2.23), attained when $\{\underline{B} - \delta_p(\underline{B})\underline{\Sigma}\}\underline{\gamma} = \underline{0}$; $\underline{\xi}_j$ is estimated by $\underline{y}_j - \underline{\Sigma}\hat{\underline{\gamma}}\hat{\underline{\gamma}}'(\hat{\underline{\gamma}}'\underline{\Sigma}\hat{\underline{\gamma}})^{-1}(\underline{y}_j - \bar{\underline{y}})$. If $\underline{\Sigma}$ is known to within a constant factor, say $\underline{\Sigma} = \tau\underline{\Upsilon}$ where $\underline{\Upsilon}$ is known but τ is an unknown scalar, then τ is estimated by $(np)^{-1}\phi_p$ where ϕ_p is the smallest root of $|\underline{B} - \phi\underline{\Upsilon}| = 0$. The estimator of $\underline{\gamma}$ is consistent but that of τ tends in probability to τ/p.

Kendall and Stuart (1967, p.392) give the special case of Koopmans' solution when $\underline{\Upsilon}$ is diagonal, and extend Geary's (1942, 1943) method of cumulants to a single linear relationship between p variates, not all of which are normally distributed.

Villegas (1961) extends Koopmans' results for a single relationship to a situation in which the vectors of observations $\underline{y}_j$ are not independent but are normally distributed with means $\underline{\xi}_j$ and covariances $C(\underline{y}_j, \underline{y}_{j'}) = c_{jj'}\underline{\Sigma}$ $(j,j' = 1, \ldots, n)$, where the $c_{jj'}$ are known coefficients and $\underline{\Sigma}$ is either known or unbiasedly estimable independently of the $\underline{y}_j$. The model is appropriate when the data arise from an (in general incomplete) block design. The estimators of $\underline{\xi}_j$ and $\underline{\gamma}$ are similar to those of Koopmans, but if we include the block effects as well as $\underline{\Sigma}$ among the parameters to be estimated by ML then we obtain an inconsistent estimator of the latter.

Tintner (1952) generalizes Koopmans' work to $p-r$ linear relationships, except that he requires $\underline{\Sigma}$ to be known completely. He argues that only $r(p-r)$ of the elements of $\underline{\Gamma}$ (equation 2.12) are independent parameters, since we may express $p-r$ of the variables in $\underline{\xi}$ in terms of the remaining r. Thus $(p-r)^2$ conditions are required to define $\underline{\Gamma}$ uniquely; of these $\frac{1}{2}(p-r)(p-r+1)$ are given by

$$\underline{\Gamma}'\underline{\Sigma}\,\underline{\Gamma} = \underline{I}_{p-r} \tag{2.39}$$

Rather curiously, the remaining $\frac{1}{2}(p-r)(p-r+1)$ conditions are applied not to $\underline{\Gamma}$ but to the set of Lagrange multipliers introduced in order to maximize the likelihood with respect to the $\underline{\xi}_j$ subject to (2.12); the effect of these constraints is to make $\underline{\Gamma}'\underline{B}\,\underline{\Gamma}$ diagonal. He finds that the ML estimator of $\underline{\xi}_j$ for given $\underline{\Gamma}$ satisfies

$$\underline{y}_j - \underline{\xi}_j = \underline{\Sigma}\,\underline{\Gamma}\,\underline{\Gamma}'\underline{y}_j \,, \tag{2.40}$$

while $\hat{\underline{\xi}}_o$ is given by $\bar{\underline{y}}$. Substituting into (2.37) gives

$$tr(\underline{\Gamma}'\underline{B}\,\underline{\Gamma}) \,, \tag{2.41}$$

which he minimizes subject to his two sets of constraints. If $\delta_1(\underline{B}), \ldots, \delta_p(\underline{B})$ are the roots in descending order of (2.23) and $\underline{\Delta}_{\underline{B}} = \mathrm{diag}\{\delta_1(\underline{B}), \ldots, \delta_p(\underline{B})\}$, then there exists a $p \times p$ matrix $\underline{K}$ such that

$$\underline{K}'\underline{\Sigma}\,\underline{K} = \underline{I} \,, \quad \underline{K}'\underline{B}\,\underline{K} = \underline{\Delta}_{\underline{B}} \tag{2.42}$$

(Rao, 1973, p.41); if $\underline{K}$ is partitioned as

$$\underline{K} = (\underline{K}_1 \ \underline{K}_2) \ , \tag{2.43}$$

where $\underline{K}_1$ is $p \times r$, then Tintner's solution is

$$\hat{\underline{\Gamma}} = \underline{K}_2 \ . \tag{2.44}$$

A comparison of (2.42) with (2.17) and (2.24) shows that

$$\underline{K} = (\underline{N}_B')^{-1} = \underline{\Sigma}^{-1} \underline{N}_B \ , \tag{2.45}$$

in the sense that a solution of (2.42) yields a solution of the other two equations and conversely. The colums of $\underline{K}_2$ are the vectors of coefficients of the last $p - r$ canonical variates in a canonical analysis in which the elements of the second set of variables denote membership or non-membership of the n populations. The $p - r$ relationships are thus obtained by equating to zero the canonical variates corresponding to the $p - r$ smallest roots. If $n > p$ then the $\delta_i(\underline{B})$ are distinct with probability one, and $\underline{K}$ is unique apart from possible changes of sign of its columns. From (2.40) the estimator of $\underline{\xi}_j$ is given by

$$\hat{\underline{\xi}}_j = \underline{y}_j - \underline{\Sigma} \ \underline{K}_2 \underline{K}_2' \ (\underline{y}_j - \bar{\underline{y}}) = \bar{\underline{y}} + \underline{\Sigma} \ \underline{K}_1 \underline{K}_1' \ (\underline{y}_j - \bar{\underline{y}}) \ . \tag{2.46}$$

Sprent's (1966) generalized least-squares (GLS) procedure for estimating a single linear relationship extends immediately from two variates (Section 1.5) to p if we take $\underline{y}$, $\underline{\xi}$ and $\underline{\gamma}$ to be p-vectors and make corresponding changes in $\underline{\Sigma}_{jj}$ and $\underline{\Sigma}_o$. Under the assumptions made by Villegas (1961), $\underline{\Sigma}_o$ has the form of a Kronecker product $\underline{C} \otimes \underline{\Sigma}$, where $\underline{C} = (c_{jj'})$; if the $\underline{y}_j$ are uncorrelated with a common known dispersion matrix $\underline{\Sigma}$ this becomes $\underline{I}_n \otimes \underline{\Sigma}$, and the GLS estimator of $\underline{\gamma}$ minimizes

$$U = \underline{\gamma}' \underline{B} \ \underline{\gamma} / \underline{\gamma}' \underline{\Sigma} \ \underline{\gamma} \ . \tag{2.47}$$

Since this is the same as (2.38), the minimum of U is $\delta_p(\underline{B})$, attained when $\underline{\gamma} \propto \underline{k}_p$, the final column of $\underline{K}$. The condition $\tilde{\underline{\gamma}}' \underline{\Sigma} \ \tilde{\underline{\gamma}} = 1$ makes $\tilde{\underline{\gamma}}$ equal $\underline{k}_p$.

Sprent (1969, p.91) extends his method to $p-r$ relationships by arguing that it is reasonable to estimate the coefficients of these relationships by the latent vectors corresponding to the $p-r$ smallest roots of (2.23); given the value of r, this is the same as Tintner's ML solution for normal populations (equation 2.44). If $\underline{\Sigma}$ is unknown, and uncorrelated replicate observations $\underline{y}_{je}$ ($j = 1, \ldots, n$; $e = 1, \ldots, m_j$) are available, then the GLS estimator of $\underline{\Sigma}$ is

$$\tilde{\underline{\Sigma}} = (M - n)^{-1} \underline{W} , \qquad (2.48)$$

where $M = \Sigma m_j$ and $\underline{W}$ is the within-groups SSP matrix given by

$$\underline{W} = \sum_j \sum_e (\underline{y}_{je} - \bar{\underline{y}}_j)(\underline{y}_{je} - \bar{\underline{y}}_j)' . \qquad (2.49)$$

Unless the number of relationships is specified in our model, we may wish to carry out a test of the hypothesis of $p-r$ functional relationships against the alternative that there are none; such a test may be obtained from a test of the significance of the last $p-r$ roots in a canonical analysis. Sprent suggests using that given by Bartlett (1947, p.179) when $\underline{\Sigma}$ is estimated by (2.48).

Gleser and Watson (1973) consider the case in which $p = 2r$, there are r relationships through the origin, and $\underline{\Sigma} = \tau \underline{I}$ with τ an unknown parameter; they use the Type I representation in the form of (2.4). Their ML estimators for $\underline{\xi}_j$ and $\underline{A}$ may be deduced from Tintner's results by replacing $\underline{B}$ by the raw between-groups SSP matrix $\underline{B}^*$ and deleting $\bar{\underline{y}}$ in (2.46); τ is estimated by $(np)^{-1} (\phi_{r+1} + \ldots + \phi_p)$ where ϕ_i is the i^{th} latent root of $\underline{B}^*$. They give conditions under which $\hat{\underline{A}}$ is consistent, but show that $\hat{\tau}$ tends in probability to $\tau/2$. They find also that the same estimator of $\underline{A}$ may be obtained using what is effectively an extension of Sprent's GLS procedure.

Williams' (1969) model for the problem of comparative calibration (Section 2.3(a)) is in our terms a Type II representation of $p-1$ linear

functional relationships with $\underline{\Sigma}$ a diagonal matrix. We have noted already that there is no ML solution when $\underline{\Sigma}$ is unknown. For the case in which $\underline{\Sigma}$ is known up to a constant factor, Williams shows that $\hat{\underline{\Lambda}}$ is proportional to $\underline{\Sigma}\underline{k}$ where $\underline{k}$ is one of the columns of $\underline{K}$ defined in (2.42). However, he takes it to be $\underline{k}_p$ instead of $\underline{k}_1$. The estimator of ϕ_j is proportional to $\underline{k}_1'(\underline{y}_j - \bar{\underline{y}})$ and $\underline{\xi}_j$ is estimated by $\bar{\underline{y}} + \underline{\Sigma}\,\underline{k}_1\underline{k}_1'\,(\underline{y}_j - \bar{\underline{y}})$.

We may summarize the results on functional relationships in this section as follows:

(i) One ML estimator for $\underline{\Gamma}$ is $\hat{\underline{\Gamma}} = \underline{K}_2$, and this is consistent in some sense when $r = p - 1$ and when there are $\frac{1}{2}p$ homogeneous relationships.

(ii) For relationships through the origin we have

$$\hat{\underline{\xi}}_j = \underline{\Sigma}\,\underline{K}_1\underline{K}_1'\,\underline{y}_j \qquad (2.50)$$

with $\underline{K}_1$ defined relative to the raw between-groups SSP matrix; otherwise $\hat{\underline{\xi}}_o = \bar{\underline{y}}$ and $\hat{\underline{\xi}}_j$ is given by (2.46).

(iii) If $\underline{\Sigma}$ is given by $\tau\underline{\Upsilon}$, where $\underline{\Upsilon}$ is known, then for $r = \frac{1}{2}p$, $\underline{\Upsilon} = I$ and for $r = p - 1$

$$\hat{\tau} = (np)^{-1}(\phi_{r+1} + \ldots + \phi_p), \qquad (2.51)$$

where ϕ_i is the i^{th} root of $|\underline{B}^* - \phi\underline{\Upsilon}| = 0$ or $|\underline{B} - \phi\underline{\Upsilon}| = 0$. When $r = \frac{1}{2}p$, $\hat{\tau}$ tends to $\frac{1}{2}\tau$; when $r = p - 1$, $\hat{\tau}$ tends to τ/p.

(iv) The GLS estimator for $\underline{\Gamma}$ is the same as the ML estimator when there is a single relationship; the procedure may be extended to the case $r = \frac{1}{2}p$ and again gives the same estimator as ML.

It therefore seems reasonable to expect that:

(i) The ML estimator of $\underline{\Gamma}$ is consistent for any number of relationships, homogeneous or otherwise.

(ii) If $\underline{\Sigma}$ is known up to a constant factor τ then $\hat{\tau}$ is given by (2.51) and tends to $\tau(1 - r/p)$ in probability as $n \to \infty$.

(iii) The GLS procedure may be extended to any number of relationships and coincides with ML if $\underline{\Sigma}$ is known up to at least a constant factor.

3. SOME INEQUALITIES FOR TRACES OF MATRIX PRODUCTS

In this chapter we present some inequalities for the traces of products of pairs of symmetric matrices. The final Theorem is applied in Chapter 4 to the problem of estimating structural relationships by maximum likelihood.

For the basic inequality we require the following notation: let $\underline{C}$ and $\underline{D}$ be real symmetric $p \times p$ matrices having orthogonal reductions

$$\underline{P}' \underline{C} \underline{P} = \underline{\Lambda}_{\underline{C}}, \quad \underline{Q}' \underline{D} \underline{Q} = \underline{\Lambda}_{\underline{D}},$$

where $\underline{\Lambda}_{\underline{C}} = \text{diag}\{\lambda_1(\underline{C}),\ldots,\lambda_p(\underline{C})\}$ and $\underline{\Lambda}_{\underline{D}} = \text{diag}\{\lambda_1(\underline{D}),\ldots,\lambda_p(\underline{D})\}$, and the sets of latent roots $\lambda_i(\underline{C})$, $\lambda_i(\underline{D})$ $(i=1,\ldots,p)$ are each in descending order. (Thus we may say that the latent vectors of $\underline{C},\underline{D}$ are ordered with respect to their latent roots.) Let the multiplicities of the roots $\lambda_i(\underline{C})$ and $\lambda_i(\underline{D})$ in order of occurrence be respectively $m_g(\underline{C})(g=1,\ldots,c)$ and $m_h(\underline{D})(h=1,\ldots,d)$, so that

$$\sum_{g=1}^{c} m_g(\underline{C}) = \sum_{h=1}^{d} m_h(\underline{D}) = p.$$

The following inequality of Richter (1958) has been given a necessary and sufficient condition for equality by Theobald (1975 a):

THEOREM 3.1. With the above notation

$$\text{tr}(\underline{C}\underline{D}) \leqslant \text{tr}(\underline{\Lambda}_{\underline{C}}\underline{\Lambda}_{\underline{D}})$$

or, equivalently,

$$\sum \lambda_i(\underline{C}\underline{D}) \leqslant \sum \lambda_i(\underline{C})\lambda_i(\underline{D}), \tag{3.1}$$

with equality if and only if

$$\underline{P}' \underline{Q} = \underline{G}' \underline{H},$$

where $\underline{G}$, $\underline{H}$ take the form

$$\underline{G} = \begin{pmatrix} \underline{G}_1 & & & \\ & \underline{G}_2 & & \\ & & \ddots & \\ & & & \underline{G}_c \end{pmatrix}, \qquad \underline{H} = \begin{pmatrix} \underline{H}_1 & & & \\ & \underline{H}_2 & & \\ & & \ddots & \\ & & & \underline{H}_d \end{pmatrix}, \tag{3.2}$$

and $\underline{G}_g$, $\underline{H}_h$ are orthogonal matrices of orders $m_g(\underline{C})$, $m_h(\underline{D})$ respectively ($g=1,\ldots,c;h=1,\ldots,d$). An equivalent condition is that there exists a common set of latent vectors for $\underline{C}$ and $\underline{D}$ which are ordered with respect to both $\{\lambda_i(\underline{C})\}$ and $\{\lambda_i(\underline{D})\}$.

Proof. Let $\underline{R} = \underline{P}'\underline{Q}$. Then $\underline{R}$ is orthogonal and

$$\operatorname{tr}(\underline{C}\underline{D}) = \operatorname{tr}(\underline{\Lambda}_{\underline{C}}\underline{R}\underline{\Lambda}_{\underline{D}}\underline{R}') = \sum_{i=1}^{p}\sum_{j=1}^{p} \lambda_i(\underline{C})r_{ij}^2\,\lambda_j(\underline{D}),$$

where $\underline{R} = (r_{ij})$. The matrix whose ij^{th} element is r_{ij}^2 is doubly stochastic, so it is sufficient to show that, for any doubly stochastic matrix $\underline{S} = (s_{ij})$,

$$\sum_{i=1}^{p}\sum_{j=1}^{p} \lambda_i(\underline{C})s_{ij}\,\lambda_j(\underline{D}) \leq \sum_{i=1}^{p} \lambda_i(\underline{C})\lambda_i(\underline{D}) \quad . \tag{3.3}$$

For any pair (s_{ue},s_{ev}) in which $u,v>e$ and both elements are positive, let

$$t_{uv}^{(e)} = \min\,(s_{ue},s_{ev})$$

and consider the operation of adding $t_{uv}^{(e)}$ to s_{ee} and s_{uv} and subtracting the same amount from s_{ue} and s_{ev}. This reduces to zero at least one off-diagonal element of $\underline{S}$, preserves the doubly stochastic character of the matrix, and increases the left-hand side of (3.3) by the non-negative amount

$$t_{uv}^{(e)}\,\{\lambda_e(\underline{C}) - \lambda_u(\underline{C})\}\{\lambda_e(\underline{D}) - \lambda_v(\underline{D})\} \quad .$$

If we clear the off-diagonal elements of the first row and column, then the second, and so on, it requires at most $(p-1)^2$ such operations to reduce $\underline{S}$ to the identity matrix. (There are $p(p-1)$ off-diagonal elements, but clearing the final non-zero element in each row also clears the coresponding column.)

Equality is attained when

$$\operatorname{tr}(\underline{\Lambda}_{\underline{C}}\underline{R}\underline{\Lambda}_{\underline{D}}\underline{R}') = \operatorname{tr}(\underline{\Lambda}_{\underline{C}}\underline{\Lambda}_{\underline{D}}). \tag{3.4}$$

It follows from (3.3) that the only products of roots which may appear on the left-hand side of (3.4) are those of the pairs $(\lambda_i(\underline{C}),\lambda_i(\underline{D}))$ $(i=1,\ldots,p)$. Thus $\underline{R}$ represents the combined effect of a rotation $\underline{G}$ of the columns of $\underline{P}$ which permutes the $\lambda_i(\underline{C})$ while preserving their order and a similar rotation $\underline{H}$ of the columns of $\underline{Q}$. So $\underline{R} = \underline{G}'\underline{H}$, where $\underline{G}$ and $\underline{H}$ satisfy

$$\underline{G}\,\underline{\Lambda}_{\underline{C}}\,\underline{G}' = \underline{\Lambda}_{\underline{C}}\ , \quad \underline{H}\,\underline{\Lambda}_{\underline{D}}\,\underline{H}' = \underline{\Lambda}_{\underline{D}}\ .$$

The general forms for orthogonal matrices satisfying these conditions are given by (3.2).

Let $\underline{T}$ be an orthogonal matrix whose columns form a common set of latent vectors for $\underline{C}$ and $\underline{D}$ and are ordered with respect to the roots of both $\underline{C}$ and $\underline{D}$. Then

$$\mathrm{tr}(\underline{C}\underline{D}) = \mathrm{tr}(\underline{T}\,\underline{\Lambda}_{\underline{C}}\,\underline{T}'\,\underline{T}\,\underline{\Lambda}_{\underline{D}}\,\underline{T}') = \mathrm{tr}(\underline{\Lambda}_{\underline{C}}\underline{\Lambda}_{\underline{D}}).$$

Conversely, if there exist $\underline{G},\underline{H}$ satisfying (3.2), then the columns of $\underline{P}\underline{G}'$ form the required set of latent vectors of $\underline{C}$ and $\underline{D}$. This completes the proof.

The inequality (3.1) is the restriction to real symmetric matrices of a similar result for Hermitian matrices given by Richter (1958) with an analytic proof. Mirsky (1959) gives an algebraic proof of the same result. It may also be derived from an inequality of Marcus (1956), using Theorem 368 of Hardy et. al. (1952, p.261). (The latter Theorem is obtained if, in (3.3), $\underline{S}$ has the form of a permutation matrix.) The above proof and the necessary and sufficient conditions for equality are taken from Theobald (1975 a). We may deduce the following:

COROLLARY 3.2. If the roots of $\underline{C}$ are distinct then the condition for equality is that the (unique) ordered set of latent vectors of $\underline{C}$ also form an ordered set of latent vectors for $\underline{D}$. If $\underline{D}$ also has distinct roots the condition is that the ordered sets of vectors are identical.

COROLLARY 3.3. If $\underline{D}$ is non-singular then

$$\mathrm{tr}(\underline{C}\,\underline{D}^{-1}) \geqslant \mathrm{tr}\,(\underline{\Lambda}_{\underline{C}}\,\underline{\Lambda}_{\underline{D}}^{-1}) = \Sigma\,\lambda_i(\underline{C})/\lambda_i(\underline{D}), \qquad (3.5)$$

with the same condition for equality as before.

The proof is immediate on replacing $\underline{D}$ by $-\underline{D}^{-1}$.

The final result in this chapter involves the simultaneous reduction of symmetric matrices to diagonal form. Let $\underline{D}$ now be positive definite and let $\underline{E}$ be a symmetric p×p matrix. Let the roots in descending order of the equations

$$|\underline{C} - \delta\, \underline{D}| = 0 \quad , \qquad |\underline{E} - \delta\, \underline{D}| = 0 \tag{3.6}$$

be respectively $\delta_i(\underline{C})$ and $\delta_i(\underline{E})$ (i=1,...,p), and let $\underline{\Delta}_{\underline{C}}$ be diag $\{\delta_1(\underline{C}),\ldots,\delta_p(\underline{C})\}$ etc. It follows (Rao, 1973, p.41) that there exist non-singular p×p matrices $\underline{N}_{\underline{C}}, \underline{N}_{\underline{E}}$ such that

$$\underline{D} = \underline{N}_{\underline{C}}\, \underline{N}_{\underline{C}}' \quad , \quad \underline{C} = \underline{N}_{\underline{C}}\, \underline{\Delta}_{\underline{C}}\, \underline{N}_{\underline{C}}' \quad , \tag{3.7}$$

$$\underline{D} = \underline{N}_{\underline{E}}\, \underline{N}_{\underline{E}}' \quad , \quad \underline{E} = \underline{N}_{\underline{E}}\, \underline{\Delta}_{\underline{E}}\, \underline{N}_{\underline{E}}' \quad . \tag{3.8}$$

Any column of $\underline{N}_{\underline{C}}$, $\underline{N}_{\underline{E}}$ which corresponds to a simple root of (3.6) is unique apart from reversal of sign. It is easily verified that $\underline{N}_{\underline{E}}^{-1}\, \underline{N}_{\underline{C}}$ is orthogonal.

From Theobald (1975 b) we take the following:

THEOREM 3.4. With the above notation

$$\mathrm{tr}\{\underline{C}(\underline{D}+\underline{E})^{-1}\} \geqslant \mathrm{tr}\{\underline{\Delta}_{\underline{C}}(\underline{I}+\underline{\Delta}_{\underline{E}})^{-1}\} = \Sigma\, \delta_i(\underline{C})\{1+\delta_i(\underline{E})\}^{-1} \quad ,$$

with equality if and only if (3.7) and (3.8) have a common solution.

Proof. We have

$$\mathrm{tr}\{\underline{C}(\underline{D}+\underline{E})^{-1}\} = \mathrm{tr}\left[\{\underline{N}_{\underline{E}}^{-1}\, \underline{N}_{\underline{C}}\, \underline{\Delta}_{\underline{C}}\, \underline{N}_{\underline{C}}'\, (\underline{N}_{\underline{E}}')^{-1}\}(\underline{I}+\underline{\Delta}_{\underline{E}})^{-1}\right]$$

$$\geqslant \mathrm{tr}\{\underline{\Delta}_{\underline{C}}(\underline{I}+\underline{\Delta}_{\underline{E}})^{-1}\}$$

by Corollary 3.3, since, if $\underline{R}$ is any p×p orthogonal matrix, $\underline{R}\, \underline{\Delta}_{\underline{C}}\, \underline{R}'$ has the same latent roots as $\underline{\Delta}_{\underline{C}}$. In terms of Theorem 3.1, $\underline{P} = \underline{N}_{\underline{E}}^{-1}\, \underline{N}_{\underline{C}}$, $\underline{Q} = \underline{I}$, and the condition for equality is that

$$\underline{N}_{\underline{E}}^{-1}\, \underline{N}_{\underline{C}} = \underline{H}_{\underline{E}}'\, \underline{G}_{\underline{C}} \quad , \tag{3.9}$$

where $\underline{G}_C$, $\underline{H}_E$ have the form of (3.2), the orders of the sub-matrices being given by the multiplicities of the roots of (3.6). Equation (3.9) may be rewritten

$$\underline{N}_C \, \underline{G}'_C = \underline{N}_E \, \underline{H}'_E \, . \tag{3.10}$$

But the orthogonal transformations $\underline{N}_C \to \underline{N}_C \, \underline{G}'_C$ and $\underline{N}_E \to \underline{N}_E \, \underline{H}'_E$ are precisely those which leave (3.7) and (3.8) unaffected, since the forms $\underline{G}_C$ and $\underline{H}_E$ are such as to satisfy

$$\underline{G}'_C \, \underline{\Delta}_C \, \underline{G}_C = \underline{\Delta}_C \quad \text{and} \quad \underline{H}'_E \, \underline{\Delta}_E \, \underline{H}_E = \underline{\Delta}_E \, .$$

Thus (3.10) amounts to the statement that (3.7) and (3.8) have a common solution.

4. ESTIMATING STRUCTURAL RELATIONSHIPS

We consider here the maximum likelihood estimation of $p - r$ linear structural relationships among the elements of a random p-vector $\underline{X}$ under the following assumptions:

(i) The vector of 'true' values, $\underline{X}$, and the vector of departures from $\underline{X}$, $\underline{\varepsilon}$, are normally distributed independently of each other.

(ii) At each of n values of $\underline{X}$ there are m replicate observations $\underline{y}_{je}$ $(j = 1, \ldots, n\ ;\ e = 1, \ldots, m)$ on $\underline{Y}$; the corresponding vectors of departures $\underline{\varepsilon}_{je}$ are identically distributed as $N_p(\underline{0},\underline{\Sigma})$, with $\underline{\Sigma}$ positive definite, and independent, both within and between groups.

Note that we assume equal numbers of replicates in each group; the likelihood is much more complicated with unequal numbers. We shall find it convenient first to maximize the likelihood given the value of $\underline{\Sigma}$, and then to consider separately different assumptions about $\underline{\Sigma}$: that it is known up to a constant factor, that it is a general positive-definite matrix, and that it is diagonal. We shall indicate those cases for which it is necessary to have replication. Tests for the number of relationships are given and the asymptotic behaviour of the estimators is examined.

4.1 The likelihood

We have shown in Section 2.1 that our assumptions about $\underline{X}$ amount to the hypothesis that it is distributed as $N_p(\underline{\mu},\underline{\Theta})$ where $\underline{\Theta}$ has rank r. For observations in the j^{th} group (2.1) becomes

$$\underline{Y}_{je} = \underline{X}_j + \underline{\varepsilon}_{je} \qquad (j = 1, \ldots, n\ ;\ e = 1, \ldots, m),$$

so that the vectors $\underline{Y}_{je}$ are jointly normal with mean $\underline{\mu}$ and second moments given by

$$D(\underline{Y}_{je}) = \underline{\Sigma} + \underline{\Theta}\ ,\quad C(\underline{Y}_{je}, \underline{Y}_{je'}) = \underline{\Theta} \quad (j = 1, \ldots, n\ ;\ e \neq e'),$$

$$C(\underline{Y}_{je}, \underline{Y}_{j'e'}) = \underline{0} \qquad (j \neq j')\ .$$

The mp-vectors $\underline{T}_j$ defined by

$$\underline{T}_j' = (\underline{y}_{j1}', \ldots, \underline{y}_{jm}') \qquad (j = 1, \ldots, n) \tag{4.1}$$

are therefore independently and normally distributed with common mean vector $\underline{1}_m \otimes \underline{\mu}$ and dispersion matrix

$$D(\underline{T}) = \underline{I}_m \otimes \underline{\Sigma} + \underline{J}_m \otimes \underline{\Theta} ,$$

where $\underline{1}_m$ is an m-vector with all elements equal to unity and $\underline{J}_m = \underline{1}_m \underline{1}_m'$. For the density function of $\underline{T}_j$ we require

$$|D(\underline{T})| = |\underline{\Sigma}|^{m-1} |\underline{\Sigma} + m\,\underline{\Theta}| \tag{4.2}$$

(Rao, 1973, p.68). It is straightforward to verify that

$$D^{-1}(\underline{T}) = \underline{I}_m \otimes \underline{\Sigma}^{-1} - \underline{J}_m \otimes (\underline{\Sigma} + m\underline{\Theta})^{-1} \underline{\Theta}\, \underline{\Sigma}^{-1} .$$

The covariance form in $\underline{T}_j$ is

$$\begin{aligned} &(\underline{T}_j - \underline{1}_m \otimes \underline{\mu})' D^{-1}(\underline{T}) (\underline{T}_j - \underline{1}_m \otimes \underline{\mu}) \\ &= \sum_e (\underline{y}_{je} - \underline{\mu})' \underline{\Sigma}^{-1} (\underline{y}_{je} - \underline{\mu}) - m^2 (\bar{\underline{y}}_j - \underline{\mu})' (\underline{\Sigma} + m\underline{\Theta})^{-1} \underline{\Theta}\, \underline{\Sigma}^{-1} (\bar{\underline{y}}_j - \underline{\mu}) \\ &= \sum_e (\underline{y}_{je} - \bar{\underline{y}}_j)' \underline{\Sigma}^{-1} (\underline{y}_{je} - \bar{\underline{y}}_j) + m(\bar{\underline{y}}_j - \underline{\mu})' (\underline{\Sigma} + m\underline{\Theta})^{-1} (\bar{\underline{y}}_j - \underline{\mu}). \end{aligned} \tag{4.3}$$

The log-likelihood is therefore, from (4.2) and (4.3),

$$\begin{aligned} L = -\tfrac{1}{2}\Big[&n(m-1)\log|\underline{\Sigma}| + n \log|\underline{\Sigma} + m\,\underline{\Theta}| + \operatorname{tr}(\underline{W}\underline{\Sigma}^{-1}) + \operatorname{tr}\{\underline{B}(\underline{\Sigma} + m\underline{\Theta})^{-1}\} \\ &+ m\,n\,(\bar{\underline{y}} - \underline{\mu})' (\underline{\Sigma} + m\underline{\Theta})^{-1} (\bar{\underline{y}} - \underline{\mu})\Big] , \end{aligned} \tag{4.4}$$

where $\underline{B}$ and $\underline{W}$ are given by (2.22) and (2.49) with m_j equal to m.

4.2 Maximum likelihood estimation

The following procedure for maximizing the likelihood with respect to $\underline{\mu}$ and $\underline{\Theta}$ is an extension of that given by Theobald (1975b).

In (4.4) the final term in the square brackets is positive definite, so we have

$$\hat{\underline{\mu}} = \bar{\underline{y}} . \tag{4.5}$$

Given the value of $\underline{\Sigma}$, we have to minimize with respect to $\underline{\Theta}$

$$F_1 = n(m-1)\log|\underline{\Sigma}| + n \log|\underline{\Sigma} + m\,\underline{\Theta}| + \operatorname{tr}(\underline{W}\underline{\Sigma}^{-1}) + \operatorname{tr}\{\underline{B}(\underline{\Sigma} + m\,\underline{\Theta})^{-1}\} . \tag{4.6}$$

Now, we have from (2.10), (2.17) and (2.24) that

$$\underline{\Sigma} = \underline{N}_{\underline{\Theta}}\, \underline{N}_{\underline{\Theta}}' \quad , \quad \underline{\Theta} = \underline{N}_{\underline{\Theta}}\, \underline{\Omega}\, \underline{N}_{\underline{\Theta}}' \, , \tag{4.7}$$

$$\underline{\Sigma} = \underline{N}_{\underline{B}}\, \underline{N}_{\underline{B}}' \quad , \quad \underline{B} = \underline{N}_{\underline{B}}\, \underline{\Delta}_{\underline{B}}\, \underline{N}_{\underline{B}}' \, , \tag{4.8}$$

where $\underline{\Omega}$ is the diagonal matrix of the roots ω_i of $|\underline{\Theta} - \omega\underline{\Sigma}| = 0$ and $\underline{\Delta}_{\underline{B}}$ the diagonal matrix of the roots $\delta_i(\underline{B})$ of $|\underline{B} - \delta\underline{\Sigma}| = 0$.

Theorem 3.4 gives

$$\begin{aligned} \operatorname{tr}\{\underline{B}(\underline{\Sigma} + m\underline{\Theta})^{-1}\} &\geqslant \operatorname{tr}\{\underline{\Delta}_{\underline{B}}(\underline{I} + m\underline{\Omega})^{-1}\} \\ &= \sum_{i=1}^{r} \delta_i(\underline{B})/(1 + m\omega_i) + \sum_{i=r+1}^{p} \delta_i(\underline{B}) \end{aligned} \tag{4.9}$$

since $\omega_{r+1} = \ldots = \omega_p = 0$. The minimum of (4.6) with respect to $\underline{\Theta}$ for given $\underline{\Omega}$ is thus

$$n\,m \log|\underline{\Sigma}| + \sum_{i=1}^{r} \{n \log(1 + m\omega_i) + \delta_i(\underline{B})/(1 + m\omega_i)\} + \sum_{i=r+1}^{p} \delta_i(\underline{B}) + \operatorname{tr}(\underline{W}\underline{\Sigma}^{-1}) \, ,$$

since $|\underline{\Sigma} + m\underline{\Theta}| = |\underline{\Sigma}||\underline{I} + m\underline{\Omega}|$. The i^{th} term in braces is minimized with respect to ω_i when $1 + m\omega_i = \delta_i(\underline{B})/n$, so $\hat{\omega}_i = \{\delta_i(\underline{B})/n - 1\}/m$ provided that $\delta_i(\underline{B}) > n \quad (i = 1, \ldots, r)$. Thus if

$$\delta_r(\underline{B}) > n \tag{4.10}$$

we have

$$\hat{\underline{\Omega}}_1 = \frac{1}{m}\left(\frac{1}{n}\,\underline{\Delta}_{\underline{B}1} - \underline{I}_r\right), \tag{4.11}$$

where

$$\underline{\Delta}_{\underline{B}1} = \operatorname{diag}\{\delta_1(\underline{B}), \ldots, \delta_r(\underline{B})\} \, ; \tag{4.12}$$

if condition (4.10) does not hold for a particular set of data we may wish to consider a smaller value for r. In (4.9) the condition for equality is that (4.7) and (4.8) have a common solution; if the $\delta_i(\underline{B})$ are distinct (as they are almost surely if $n > p$) then so are the $\hat{\omega}_i$ $(i = 1, \ldots, r)$, and equality occurs if and only if $\underline{N}_{\underline{\Theta}1} = \underline{N}_{\underline{B}1}$, where $\underline{N}_{\underline{B}1}$ consists of the first r columns of $\underline{N}_{\underline{B}}$. The ML estimator of $\underline{\Theta}$ is

$$\begin{aligned} \hat{\underline{\Theta}} = \hat{\underline{N}}_{\underline{\Theta}1}\, \hat{\underline{\Omega}}_1 \hat{\underline{N}}_{\underline{\Theta}1}' &= \frac{1}{m}\, \underline{N}_{\underline{B}1}\left(\frac{1}{n}\,\underline{\Delta}_{\underline{B}1} - \underline{I}_r\right)\underline{N}_{\underline{B}1}' \\ &= \frac{1}{m} \sum_{i=1}^{r} \left\{\frac{1}{n}\,\delta_i(\underline{B}) - 1\right\} \underline{n}_i\, \underline{n}_i' \, , \end{aligned} \tag{4.13}$$

where the $\underline{n}_i$ are the columns of $\underline{N}_B$. A corresponding estimator of $\underline{\Lambda}$ is any $p \times r$ matrix $\hat{\underline{\Lambda}}$ satisfying $\hat{\underline{\Lambda}}\hat{\underline{\Lambda}}' = \underline{\Theta}$, and the unique $\underline{\Lambda}_o$ satisfying (2.11) is estimated by

$$\hat{\underline{\Lambda}}_o = m^{-\frac{1}{2}} \underline{N}_{B1} \left(\frac{1}{n} \underline{\Delta}_{B1} - \underline{I}_r \right)^{\frac{1}{2}} = m^{-\frac{1}{2}} \sum_{i=1}^{r} \left\{ \frac{1}{n} \delta_i (\underline{B}) - 1 \right\}^{\frac{1}{2}} \underline{n}_i . \qquad (4.14)$$

These estimators are such that the possibility that $m = 1$ need not be excluded when $\underline{\Sigma}$ is known.

For the Type I representation of the relationships we require a $p \times (p - r)$ matrix $\hat{\underline{\Gamma}}$ of full rank which satisfies $\hat{\underline{\Gamma}}'\hat{\underline{\Lambda}} = \underline{0}$ (equation 2.7); (4.14) shows that this is equivalent to $\hat{\underline{\Gamma}}'\underline{N}_{B1} = \underline{0}$. Now, we have from (2.45) that $\underline{K}'\underline{N}_B = \underline{I}$, so one solution is

$$\hat{\underline{\Gamma}} = \underline{K}_2 , \qquad (4.15)$$

where $\underline{K}_2$ is defined by (2.43). This is precisely Tintner's ML solution (2.44) for functional relationships. The general solution is

$$\hat{\underline{\Gamma}} = \underline{K}_2 \underline{R} , \qquad (4.16)$$

where $\underline{R}$ is any nonsingular $(p - r) \times (p - r)$ matrix, for we have $\hat{\underline{\Gamma}}'\underline{N}_{B1} = \underline{R}'\underline{K}_2'\underline{N}_{B1} = \underline{0}$; conversely, if $\hat{\underline{\Gamma}}'\underline{N}_{B1} = \underline{0}$ then the columns of $\hat{\underline{\Gamma}}$ and $\underline{K}_2$ both form a basis for the null space of $\underline{N}_{B1}$ and may therefore be expressed in terms of each other, that is, we may write $\hat{\underline{\Gamma}} = \underline{K}_2\underline{R}$ with $\underline{R}$ nonsingular. Finally, the maximum of the log-likelihood with respect to $\underline{\mu}$ and $\underline{\Theta}$ for given $\underline{\Sigma}$ is $-\frac{1}{2} F_2$, where

$$F_2 = n m \log|\underline{\Sigma}| + n \sum_{i=1}^{r} \log\{\delta_i(\underline{B})/n\} + \operatorname{tr}(\underline{W}\underline{\Sigma}^{-1}) + \sum_{i=r+1}^{p} \delta_i(\underline{B}) + n r \qquad (4.17)$$

We next consider the minimization of (4.17) under various assumptions about the parametric form of $\underline{\Sigma}$.

4.2.1 $\underline{\Sigma}$ known up to a constant factor

We suppose that $\underline{\Sigma}$ may be written as $\tau\underline{\Upsilon}$ where $\underline{\Upsilon}$ is a known positive-definite matrix and τ is unknown. Writing ϕ_i for the i^{th} root of

$|\underline{B} - \phi\underline{T}| = 0$ we have $\phi_i = \tau\,\delta_i(\underline{B})$ $(i = 1, \ldots, p)$, and (4.17) becomes, apart from terms not involving τ,

$$n(mp - r)\log\tau + \frac{1}{\tau}\{tr(\underline{W}\underline{T}^{-1}) + tr(\underline{\Phi}_2)\} ,$$

where

$$\underline{\Phi}_2 = diag(\phi_{r+1}, \ldots, \phi_p) , \tag{4.18}$$

and this is maximized with respect to τ at

$$\hat{\tau} = \{n(mp - r)\}^{-1}\{tr(\underline{W}\underline{T}^{-1}) + tr(\underline{\Phi}_2)\} . \tag{4.19}$$

The estimators of $\underline{\Theta}$, $\underline{\Lambda}_o$ and $\underline{\Gamma}$ are unaffected apart from the replacement of $\underline{\Sigma}$ by $\hat{\tau}\underline{T}$. Again the case $m = 1$ is not excluded. The maximum of the log-likelihood is

$$-\frac{1}{2}n\{m\log|\underline{T}| + \sum_{i=1}^{r}\log(\phi_i/n) + (mp - r)\log\hat{\tau} + mp\} . \tag{4.20}$$

4.2.2 $\underline{\Sigma}$ a general positive-definite matrix

To find the stationary points of F_1 (equation 4.6) we equate to zero the matrix of derivatives

$$\partial F_1/\partial\underline{\Sigma} = n(m - 1)\underline{\Sigma}^{-1} + n(\underline{\Sigma} + m\underline{\Theta})^{-1} - \underline{\Sigma}^{-1}\underline{W}\underline{\Sigma}^{-1} - (\underline{\Sigma} + m\underline{\Theta})^{-1}\underline{B}(\underline{\Sigma} + m\underline{\Theta})^{-1} . \tag{4.21}$$

The matrix derivative used is that of Dwyer (1967) rather than that given by, say, Rao (1973, p.72), the distinction being that Dwyer's derivative is defined relative to the position of a matrix element as well as its scalar value. From (4.8) and (4.13) we have

$$\hat{\underline{\Sigma}} + m\,\hat{\underline{\Theta}} = \hat{\underline{N}}_{\underline{B}}\begin{pmatrix}\frac{1}{n}\hat{\underline{\Delta}}_{\underline{B}1} & \underline{0}\\ \underline{0} & \underline{I}\end{pmatrix}\hat{\underline{N}}_{\underline{B}}' .$$

Substituting this into (4.21), premultiplying the result by $\hat{\underline{N}}_{\underline{B}}'$, and postmultiplying by $\hat{\underline{N}}_{\underline{B}}$ gives

$$n(m - 1)\underline{I} + n\begin{pmatrix}n\hat{\underline{\Delta}}_{\underline{B}1} & \underline{0}\\ \underline{0} & \underline{I}\end{pmatrix} = \hat{\underline{K}}'\underline{W}\,\hat{\underline{K}} + \begin{pmatrix}n^2\hat{\underline{\Delta}}_{\underline{B}1} & \underline{0}\\ \underline{0} & \hat{\underline{\Delta}}_{\underline{B}2}\end{pmatrix} ,$$

where $\hat{\underline{K}}' = \hat{\underline{N}}_{\underline{B}}^{-1}$ and

$$\underline{\Delta}_{\underline{B}2} = \mathrm{diag}\{\delta_{r+1}(\underline{B}), \ldots, \delta_p(\underline{B})\} , \tag{4.22}$$

so that we have

$$\hat{\underline{K}}'\underline{W}\hat{\underline{K}} = \begin{pmatrix} n(m-1)\,\underline{I} & \underline{0} \\ \underline{0} & nm\underline{I} - \hat{\underline{\Delta}}_{\underline{B}2} \end{pmatrix} \tag{4.23}$$

Thus $\hat{\underline{K}}$ diagonalizes $\underline{W}$ as well as $\hat{\underline{\Sigma}}$ and $\underline{B}$, which suggests that $\hat{\underline{\Delta}}_{\underline{B}}$ and $\hat{\underline{K}}$ may be simply expressed in terms of the simultaneous reduction of $\underline{W}$ and $\underline{B}$ to diagonal form. So let $\psi_1, \ldots, \psi_p$ be the roots in descending order of

$$|\underline{B} - \psi\underline{W}| = 0 , \tag{4.24}$$

let

$$\underline{\Psi} = \mathrm{diag}(\psi_1, \ldots, \psi_p),\ \underline{\Psi}_1 = \mathrm{diag}(\psi_1, \ldots, \psi_r),\ \underline{\Psi}_2 = \mathrm{diag}(\psi_{r+1}, \ldots, \psi_p), \tag{4.25}$$

and let $\underline{G}$ be any $p \times p$ matrix such that

$$\underline{G}'\underline{W}\underline{G} = \underline{I}, \quad \underline{G}'\underline{B}\underline{G} = \underline{\Psi} . \tag{4.26}$$

The ψ_i must also be the roots of $|\hat{\underline{K}}'\underline{B}\hat{\underline{K}} - \psi\hat{\underline{K}}'\underline{W}\hat{\underline{K}}| = 0$, that is, using (2.42) and (4.23), of

$$\left| \hat{\underline{\Delta}}_{\underline{B}} - \psi \begin{pmatrix} n(m-1)\,\underline{I} & \underline{0} \\ \underline{0} & nm\underline{I} - \hat{\underline{\Delta}}_{\underline{B}2} \end{pmatrix} \right| = 0$$

or $$\prod_{i=1}^{r}\{\hat{\delta}_i(\underline{B}) - n(m-1)\psi\} . \prod_{i=r+1}^{p}\{\hat{\delta}_i(\underline{B}) - nm\psi + \psi\hat{\delta}_i(\underline{B})\} = 0 .$$

Thus for some permutation π of $(1, \ldots, p)$ we have

$$\left.\begin{aligned} \hat{\delta}_i(\underline{B}) &= n(m-1)\,\psi_{\pi(i)} \quad (i = 1, \ldots, r), \\ \hat{\delta}_i(\underline{B}) &= nm\psi_{\pi(i)}(1 + \psi_{\pi(i)})^{-1} \quad (i = r+1, \ldots, p), \end{aligned}\right\} \tag{4.27}$$

subject to condition (4.10) which becomes

$$(m-1)\,\psi_{\pi(r)} > 1 . \tag{4.28}$$

To find the permutation which maximizes the likelihood we substitute (4.23) into (4.17) to give

$$n \log (|\hat{\underline{\Sigma}}|^m \, |\hat{\underline{\Delta}}_{B1}|) + \text{constant terms.}$$

Therefore, since $|\underline{B}| = |\hat{\underline{\Sigma}}| \, |\hat{\underline{\Delta}}_{\underline{B}}|$, we have to maximize $|\hat{\underline{\Delta}}_{\underline{B}}|^m \, |\hat{\underline{\Delta}}_{B1}|^{-1}$, which may be shown, using (4.27), to be equivalent to choosing π to maximize

$$\prod_{i=1}^{r} (1 + \psi_{\pi(i)})^m / \psi_{\pi(i)} \, . \tag{4.29}$$

Now the function $(1 + \psi)^m / \psi$ is increasing if $(m-1)\psi > 1$, and we have already had to apply this condition to $\psi_{\pi(r)}$ (equation 4.28) and therefore to $\psi_{\pi(i)}$ $(i = 1, \ldots, r)$. So, if $\psi_r > \psi_{r+1}$ (as it almost surely is if $n > p$), (4.29) is maximized when and only when π maps the set $(1, \ldots, r)$ onto itself. Equations (4.27) then become

$$\hat{\underline{\Delta}}_{B1} = n(m-1)\underline{\Psi}_1 \, , \quad \hat{\underline{\Delta}}_{B2} = n\,m\underline{\Psi}_2(\underline{I} + \underline{\Psi}_2)^{-1} \, , \tag{4.30}$$

while condition (4.28) becomes

$$(m-1)\psi_r > 1 \, . \tag{4.31}$$

To obtain $\hat{\underline{K}}$ in terms of $\underline{\Psi}$ and $\underline{G}$ we let $\underline{H}$ be $\underline{G}^{-1}\hat{\underline{K}}$; (4.23), (4.26) and (4.30) then give

$$\underline{H}'\underline{H} = n \begin{pmatrix} (m-1)\underline{I} & \underline{0} \\ \underline{0} & m(\underline{I} + \underline{\Psi}_2)^{-1} \end{pmatrix} , \tag{4.32}$$

while (4.30) and the second parts of (2.42) and (4.26) lead to

$$\underline{H}'\underline{\Psi}\underline{H} = \hat{\underline{\Delta}}_{\underline{B}} = n \begin{pmatrix} (m-1)\underline{\Psi}_1 & \underline{0} \\ \underline{0} & m\underline{\Psi}_2(\underline{I} + \underline{\Psi}_2)^{-1} \end{pmatrix} . \tag{4.33}$$

One solution of (4.32) and (4.33) is

$$\underline{H}_d = n^{\frac{1}{2}} \begin{pmatrix} (m-1)^{\frac{1}{2}}\underline{I} & \underline{0} \\ \underline{0} & m^{\frac{1}{2}}(\underline{I} + \underline{\Psi}_2)^{-\frac{1}{2}} \end{pmatrix} . \tag{4.34}$$

If $\underline{J}$ is a $p \times p$ matrix satisfying $\underline{H} = \underline{J}\underline{H}_d$, then (4.32) shows that $\underline{J}$ is orthogonal and (4.33) that $\underline{J}$ commutes with $\underline{\Psi}$ (and therefore with $\underline{H}_d$). This implies that if $\underline{J} = (j_{hi})$ then $j_{hi} = 0$ whenever $\psi_h \neq \psi_i$. In particular, if $\underline{J}$ is partitioned as

$$\underline{J} = \begin{pmatrix} \underline{J}_{11} & \underline{J}_{12} \\ \underline{J}_{21} & \underline{J}_{22} \end{pmatrix}$$

with $\underline{J}_{11}$ $r \times r$, then $\underline{J}_{12}$ and $\underline{J}_{21}$ are zero, since we have assumed that $\psi_r > \psi_{r+1}$. It follows that $\underline{J}_{11}$ and $\underline{J}_{22}$ are orthogonal and commute with $\underline{\Psi}_1$ and $\underline{\Psi}_2$ respectively. Since $\hat{\underline{K}} = \underline{G}\underline{H} = \underline{G}\underline{J}\underline{H}_d$ and $\hat{\underline{N}}_B = (\hat{\underline{K}}')^{-1} = (\underline{G}')^{-1}\underline{J}\underline{H}_d^{-1}$, we have

$$\hat{\underline{K}}_1 = \{n(m-1)\}^{\frac{1}{2}}\,\underline{G}_1\,\underline{J}_{11}\,, \qquad \hat{\underline{K}}_2 = (n\,m)^{\frac{1}{2}}\,\underline{G}_2\,\underline{J}_{22}(\underline{I} + \underline{\Psi}_2)^{-\frac{1}{2}} \tag{4.35}$$

and

$$\hat{\underline{N}}_{B1} = \{n(m-1)\}^{-\frac{1}{2}}\,\underline{L}_1\,\underline{J}_{11}\,, \qquad \hat{\underline{N}}_{B2} = (n\,m)^{-\frac{1}{2}}\,\underline{L}_2\,\underline{J}_{22}(\underline{I} + \underline{\Psi}_2)^{\frac{1}{2}}, \tag{4.36}$$

where

$$\underline{G} = (\underline{G}_1 \quad \underline{G}_2)\,, \tag{4.37}$$

$$\underline{L} = (\underline{G}')^{-1} = \underline{W}\,\underline{G} = (\underline{W}\,\underline{G}_1 \quad \underline{W}\,\underline{G}_2) = (\underline{L}_1 \quad \underline{L}_2)\,, \tag{4.38}$$

and $\underline{G}_1$ and $\underline{L}_1$ are $p \times r$. Substituting (4.30) and (4.36) into (4.13), (4.14), and then the first part of (4.8) we have respectively

$$\hat{\underline{\Theta}} = \{n\,m(m-1)\}^{-1}\underline{L}_1\{(m-1)\,\underline{\Psi}_1 - \underline{I}\}\,\underline{L}_1'\,, \tag{4.39}$$

$$\hat{\underline{\Lambda}}_o = \{n\,m(m-1)\}^{-\frac{1}{2}}\underline{L}_1\{(m-1)\,\underline{\Psi}_1 - \underline{I}\}^{\frac{1}{2}}\,, \tag{4.40}$$

and

$$\hat{\underline{\Sigma}} = \hat{\underline{N}}_{B1}\,\hat{\underline{N}}_{B1}' + \hat{\underline{N}}_{B2}\,\hat{\underline{N}}_{B2}' = \{n(m-1)\}^{-1}\underline{L}_1\underline{L}_1' + (n\,m)^{-1}\underline{L}_2(\underline{I} + \underline{\Psi}_2)\underline{L}_2'\,,$$

provided that $\psi_r > \psi_{r+1}$ and (4.31) holds; in (4.40) we have to assume that the ψ_i are distinct so that $\underline{J}_{11} = \underline{I}$.

The ML estimator of $\underline{\Gamma}$ is, following (4.16), $\hat{\underline{\Gamma}} = \hat{\underline{K}}_2\,\underline{R}$ with $\underline{R}$ an arbitrary nonsingular $(p-r) \times (p-r)$ matrix, and this becomes

$$\hat{\underline{\Gamma}} = \underline{G}_2 \underline{R} , \tag{4.42}$$

since the other terms in the second part of (4.35) can be subsumed into $\underline{R}$. As in the case of known $\underline{\Sigma}$ the relationships are estimated by equating to zero the last $p-r$ canonical variates; in this case the corresponding canonical analysis uses $\{n(m-1)\}^{-1}\underline{W}$ rather than the unknown $\underline{\Sigma}$ to represent within-group variation. Similarly, (4.39) and (4.40) are obtained from (4.13) and (4.14) when $\underline{\Sigma}$ is replaced by $\{n(m-1)\}^{-1}\underline{W}$, although it is not the ML estimator.

The maximum of the log-likelihood in this case is

$$-\tfrac{1}{2}n\Big[m\{\log|\underline{W}| + \log|\underline{I} + \underline{\Psi}_2| - p \log n - (p-r) \log m + p\} + \log|\underline{\Psi}_1| - r(m-1) \log(m-1)\Big]. \tag{4.43}$$

We must obviously have $m > 1$ in this case.

4.2.3 $\underline{\Sigma}$ a diagonal matrix

There appears to be no explicit estimator for $\underline{\Sigma}$ in this case; maximization of the likelihood must be carried out numerically with respect to the diagonal elements σ_{ii} $(i = 1, \ldots, p)$. Otherwise the estimation procedure can be simplified by defining the matrix $\underline{\Sigma}^{\frac{1}{2}}$ as the diagonal square root of $\underline{\Sigma}$ and $\underline{C}$ as $\underline{\Sigma}^{\frac{1}{2}}\underline{K}$. It then follows from (2.42) that $\underline{C}$ is orthogonal and that the $\delta_i(\underline{B})$ and the columns of $\underline{C}$ are respectively the latent roots and vectors of the matrix $\underline{\Sigma}^{-\frac{1}{2}}\underline{B}\underline{\Sigma}^{-\frac{1}{2}}$. The matrix $\underline{N}_B$ is given by $\underline{\Sigma}^{\frac{1}{2}}\underline{C}$. The columns of $\underline{C}$ may be viewed as the vectors of coefficients of the components in a principal component analysis carried out on the $\bar{\underline{y}}_j$ if the i^{th} element in each $\bar{\underline{y}}_j$ is divided by $\sigma_{ii}^{\frac{1}{2}}$ $(i = 1, \ldots, p)$. From (4.17) the function to be minimized with respect to the σ_{ii} is

$$F_3 = \sum_{i=1}^{p} (n\,m \log \sigma_{ii} + w_{ii}/\sigma_{ii}) + n \sum_{i=1}^{r} \log\{\delta_i(\underline{B})/n\} + \sum_{i=r+1}^{p} \delta_i(\underline{B}) , \tag{4.44}$$

where $\underline{W} = (w_{hi})$, and the maximum of the log-likelihood is

$$-\tfrac{1}{2}\,(F_{3,\min} + n\,r)\,. \tag{4.45}$$

The efficient minimization of (4.44) requires at least the first derivatives of F_3; we therefore note that the derivatives of the $\delta_i(\underline{B})$ may be obtained from standard perturbation theory for symmetric matrices (Bellman, 1970, pp. 61-63).

There is little prospect of finding an explicit solution for the ML estimators of the σ_{ii}; none appears to have been found for the factor analysis model — corresponding to $m = 1$ — despite extensive study. The connexion with factor analysis also suggests that the parameters of the model are identifiable when there is no replication provided that condition (2.33) is fulfilled.

4.2.4 Relationships through the origin

So far in this Section we have assumed that the elements of $\underline{\mu}$ are unconstrained; we now see what progress can be made under the assumption that the $p - r$ relationships pass through the origin. We have shown in Section 2.1 that this is equivalent to assuming that $\underline{\mu}$ lies in the range space of $\underline{\Theta}$; since $\underline{N}_{\underline{\Theta}1}$ is of full rank and has the same range space as $\underline{\Theta}$ we may write $\underline{\mu} = \underline{N}_{\underline{\Theta}1}\,\underline{\nu}$ with $\underline{\nu}$ an r-vector of independent parameters. From (4.4) we have to minimize with respect to $\underline{\nu}$

$$(\bar{\underline{y}} - \underline{N}_{\underline{\Theta}1}\,\underline{\nu})'(\underline{\Sigma} + m\,\underline{\Theta})^{-1}\,(\bar{\underline{y}} - \underline{N}_{\underline{\Theta}1}\,\underline{\nu})\,. \tag{4.46}$$

Equating to zero the derivative with respect to $\underline{\nu}$ and using (4.7), we obtain

$$\hat{\underline{\nu}} = \underline{N}'_{\underline{\Theta}1}\,\underline{\Sigma}^{-1}\,\bar{\underline{y}},$$

and the minimum with respect to $\underline{\nu}$ of (4.46) is

$$\bar{\underline{y}}'\,\underline{\Sigma}^{-1}\,\underline{N}_{\underline{\Theta}2}\,\underline{N}'_{\underline{\Theta}2}\,\underline{\Sigma}^{-1}\,\bar{\underline{y}}\,.$$

But is does not appear possible to proceed beyond this point in the manner of (4.9).

4.3 Tests of the number of relationships

A test of the hypothesis of $p-r$ structural relationships may be based on the difference between the maxima of the log-likelihood when there are no relationships and when there are $p-r$. These differences are, from (4.17), (4.20) and (4.43): for known $\underline{\Sigma}$,

$$\frac{1}{2}\sum_{i=r+1}^{p}\left[\delta_i(\underline{B}) - n\log\{\delta_i(\underline{B})/n\} - n\right];$$

for $\underline{\Sigma}$ known up to a constant factor,

$$\frac{1}{2}n\{(mp-r)\log\hat{\tau} - (m-1)p\left[\log\{\mathrm{tr}(\underline{W}\underline{T}^{-1}) - \log\{n(m-1)p\}\right] - \sum_{i=r+1}^{p}\log(\phi_i/n)\};$$

and for a general positive-definite $\underline{\Sigma}$,

$$\frac{1}{2}n\left[\sum_{i=r+1}^{p}\{m\log(1+\psi_i) - \log\psi_i\} + (p-r)\{(m-1)\log(m-1) - m\log m\}\right].$$

When $\underline{\Sigma}$ is diagonal and there are no structural relationships, σ_{ii} is estimated by $\{n(m-1)\}^{-1}w_{ii}$ and the difference between the maxima of the likelihoods is

$$\frac{1}{2}\left[n(m-1)\sum_{i=1}^{p}\log\hat{\sigma}_{ii} + n\log|n^{-1}\underline{B}| + n(mp-r) - F_{3,\min}\right].$$

For this test to be possible we must have $m>1$ whatever the value of r.

It follows from standard ML theory that twice each of these differences has asymptotically a χ^2 distribution as n tends to infinity under the hypothesis that $p-r$ relationships exist. The degrees of freedom are $\frac{1}{2}(p-r)(p-r+1)$, since $\underline{\Theta}$ has $\frac{1}{2}p(p+1)$ independent parameters when its rank is p and $\{pr - \frac{1}{2}r(r-1)\}$ when its rank is r.

4.4 Asymptotic behaviour

If there is replication then the total number of observations may tend to infinity in more than one way; we consider here two possibilities: (i) that n remains fixed and m tends to infinity, and

(ii) that m is fixed while n tends to infinity. In neither case can we apply standard ML theory directly to the $\underline{y}_{je}$ since they are correlated, but in case (ii) we can take the vectors of observations to be the $\underline{T}_j$ defined in (4.1); since they are independent and identically distributed the usual asymptotic properties hold for the estimators of $\underline{\mu}$, $\underline{\Theta}$, $\underline{\Sigma}$ and τ. This also applies when $m = 1$ in those cases in which the parameters are identifiable.

In case (i) we cannot expect $\hat{\underline{\mu}}$ and $\hat{\underline{\Theta}}$ to be consistent since we have (indirect) observations on only a finite number of values of the underlying variate $\underline{X}$. However, the overall mean $\bar{\underline{Y}}$ is distributed as $N_p(\underline{\mu}, n^{-1}(m^{-1}\underline{\Sigma} + \underline{\Theta}))$, so that any fixed linear function $\underline{c}'\underline{\mu}$ for which $\underline{c}'\underline{\Theta} = \underline{0}$ is consistently estimated. Also, given that $\underline{X}$ takes values $\underline{x}_j$ $(j = 1, \ldots, n)$, $\bar{\underline{Y}}_j \to \underline{x}_j$ as m tends to infinity (we use the symbol $\to$ exclusively to denote convergence in probability). So the unconditional distribution of each $\bar{\underline{Y}}_j$ tends to become concentrated on the hyperplane containing the $\underline{x}_j$, suggesting, correctly, that $\underline{\Gamma}$ is estimated consistently. The matrix $\underline{\Sigma}$ is also consistently estimated in all cases.

To prove these assertions we first note that $\underline{W}$ and $\underline{B}$ are distributed as $W_p(n(m-1), \underline{\Sigma})$ and $W_p(n-1, \underline{\Sigma} + m\underline{\Theta})$ respectively. Let $\underline{D}$ be $(\underline{N}'_{\underline{\Theta}})^{-1}$ and partition it as $(\underline{D}_1\ \underline{D}_2)$ with $\underline{D}_1$ $p \times r$ so that, from (4.7),

$$\underline{D}'\underline{\Sigma}\,\underline{D} = \underline{I}\ , \quad \underline{D}_1'\,\underline{\Theta}\,\underline{D}_1 = \underline{\Omega}_1\ , \quad \underline{D}_2'\,\underline{\Theta}\,\underline{D}_2 = \underline{0}\ . \tag{4.47}$$

Then $m^{-1}\,\underline{D}_1'\,\underline{B}\,\underline{D}_1 \sim W_r(n-1, m^{-1}\underline{I} + \underline{\Omega}_1)$ and $m^{-1}\,\underline{D}_2'\,\underline{B}\,\underline{D}_2 \sim W_{p-r}(n-1, m^{-1}\underline{I})$, so that, in the limit, $m^{-1}\,\underline{D}_1'\,\underline{B}\,\underline{D}_1$ and $m^{-1}\,\underline{\Delta}_{\underline{B}1}$ remain finite and positive definite almost surely, while

$$m^{-1}\,\underline{D}_2'\,\underline{B}\,\underline{D}_2 \to \underline{0}, \quad m^{-1}\,\underline{\Delta}_{\underline{B}2} \to \underline{0}\ . \tag{4.48}$$

Now define $\underline{P}$ as $(\underline{K}^{-1}\,\underline{D})'$ and partition it as

$$\underline{P} = (\underline{P}_1\ \underline{P}_2) = \begin{pmatrix} \underline{P}_{11} & \underline{P}_{12} \\ \underline{P}_{21} & \underline{P}_{22} \end{pmatrix}$$

where $\underline{P}_1$ is $p \times r$ and $\underline{P}_{11}$ is $r \times r$; $\underline{P}$ is orthogonal since $\underline{K}'\underline{\Sigma}\,\underline{K} = \underline{D}'\underline{\Sigma}\,\underline{D} = \underline{I}$, so $\underline{K} = \underline{D}\,\underline{P}$ and $\underline{K}_2 = \underline{D}\,\underline{P}_2$. Also, since $\underline{\Delta}_{\underline{B}} = \underline{K}'\underline{B}\,\underline{K}$, we have

$$m^{-1}\,\underline{\Delta}_{\underline{B}2} = m^{-1}\,\underline{K}_2'\,\underline{B}\,\underline{K}_2 = m^{-1}\,\underline{P}_2'\underline{D}'\underline{B}\,\underline{D}\,\underline{P}_2 .$$

In the limit the left-hand side of this equation is zero, so the right-hand side cannot depend on $m^{-1}\,\underline{D}_1'\,\underline{B}\,\underline{D}_1$; we must have $\underline{P}_{12} \rightarrow \underline{O}$. Therefore, in the limit,

$$\underline{K}_2'\,\underline{\Theta}\,\underline{K}_2 = \underline{P}_{22}'\,\underline{D}_2'\,\underline{\Theta}\,\underline{D}_2\,\underline{P}_{22} = \underline{O} ,$$

because of (4.47). Thus, when $\underline{\Sigma}$ is given, $\hat{\underline{\Gamma}}$ is consistent in the sense that, for any solution of (4.16) which remains finite in the limit,

$$\hat{\underline{\Gamma}}'\underline{\Theta} \rightarrow \underline{O} . \qquad (4.49)$$

When $\underline{\Sigma}$ has to be estimated, the consistency of $\hat{\underline{\Gamma}}$ will follow if we can show that $\hat{\underline{\Sigma}}$ is consistent. We have

$$(n\,m)^{-1}\,\underline{W} \rightarrow \underline{\Sigma} \quad \text{and} \quad \underline{\Psi}_2 \rightarrow \underline{O} , \qquad (4.50)$$

the latter because $n\,m\,\underline{\Psi}_2$ must have the same limiting behaviour as $\underline{\Delta}_{\underline{B}2}$ (equation 4.48). For $\underline{\Sigma}$ known up to a constant factor we have from (4.19) that

$$\hat{\tau} = (n\,m\,p)^{-1}\,\tau\{\mathrm{tr}(\underline{W}\,\underline{\Sigma}^{-1}) + \mathrm{tr}(\underline{\Delta}_{\underline{B}2})\} + o(1) ,$$

which tends in probability to τ because of (4.50) and (4.48). When $\underline{\Sigma}$ is a general positive-definite matrix, (4.41) and (4.50) give

$$\hat{\underline{\Sigma}} = (n\,m)^{-1}\,(\underline{L}_1\underline{L}_1' + \underline{L}_2\underline{L}_2') + o(1) = (n\,m)^{-1}\,\underline{W} + o(1) ,$$

so that $\hat{\underline{\Sigma}}$ is consistent. Finally, when $\underline{\Sigma}$ is diagonal we use the fact that if $\eta_i = m^{-1}\,\delta_i(\underline{B})$ $(i = 1, \ldots, r)$ then the η_i are almost surely finite and positive. Minimizing (4.44) is equivalent to minimizing

$$\sum_{i=1}^{p}\{\log \sigma_{ii} + (n\,m\,\sigma_{ii})^{-1} w_{ii}\} + m^{-1}\sum_{i=1}^{r}\log(n^{-1}\,m\,\eta_i) + (n\,m)^{-1}\sum_{i=r+1}^{p}\delta_i(\underline{B}) .$$

The second and third terms tend to zero in probability as m tends to infinity (the second because $a \log a$ tends to zero with a) while the

first becomes

$$\sum_{i=1}^{p} (\log \sigma_{ii} + \sigma_{ii}^{-1} \sigma_{oii}) , \tag{4.51}$$

where σ_{oii} is the true value of σ_{ii} (i = 1, ..., p). The minimum of (4.51) occurs at $\hat{\sigma}_{ii} = \sigma_{oii}$ (i = 1, ..., p), so $\hat{\underline{\Sigma}}$ is again consistent.

5. ESTIMATING FUNCTIONAL RELATIONSHIPS

We consider in this Chapter the estimation of systems of linear functional relationships by the methods of maximum likelihood and generalized least-squares. Our assumptions are that in each of n 'groups' there are m_j vectors of observations $\underline{Y}_{je}$, uncorrelated and with a common positive-definite dispersion matrix $\underline{\Sigma}$ ($e=1,\ldots,m_j$; $j=1,\ldots,n$), that observations in different groups are also uncorrelated, and that the mean vectors $\underline{\xi}_j$ are all subject to the same set of p-r independent linear relationships. When considering ML estimation we also suppose that the $\underline{Y}_{je}$ are jointly normally distributed. Our assumptions about departures from the 'true' values are then similar to those of Chapter 4, but we find here that it is straightforward to incorporate both unequal numbers of replicates and relationships through the origin. We again maximize the likelihood given the value of $\underline{\Sigma}$ before considering separately different parametric forms for the dispersion matrix. We also extend Sprent's generalized least-squares procedure to p-r relationships and show its connexion with maximum likelihood under the assumption of normality. As in Chapter 4 we indicate when it is necessary to have replicate observations, give tests for the number of relationships, and examine the asymptotic behaviour of the estimators. There are many similarities with Chapter 4; in particular the ML estimators of the relationships are the same as in the structural case when $\underline{\Sigma}$ is known up to a constant factor and when $\underline{\Sigma}$ is a general, but unknown, positive-definite matrix.

5.1 Maximum likelihood estimation

The mean vectors $\underline{\xi}_j$ may be viewed as being subject to p-r linear constraints or, equivalently, as lying in a hyperplane defined by a

vector in the hyperplane and r vectors parallel to it. We shall derive the ML estimators of the $\underline{\xi}_j$ using the first (Type I) representation since it facilitates comparisons with the work on estimation reviewed in Section 2.5 and with the extension of the GLS procedure developed in Section 5, but we shall also consider briefly the use of the Type II representation.

We take the vectors of observations $\underline{Y}_{je}$ to be distributed independently as $N_p(\underline{\xi}_j, \underline{\Sigma})$ $(e = 1,\ldots, m_j;\ j=1,\ldots,n)$, so if $M = \Sigma m_j$ and $\underline{W}$ is given by (2.49) the log-likelihood is

$$-\frac{1}{2}\{M \log|\underline{\Sigma}| + \sum_{j=1}^{n}\sum_{e=1}^{m_j}(\underline{y}_{je}-\underline{\xi}_j)'\,\underline{\Sigma}^{-1}(\underline{y}_{je}-\underline{\xi}_j)\}$$

$$= -\frac{1}{2}\{M \log|\underline{\Sigma}| + \mathrm{tr}(\underline{W}\underline{\Sigma}^{-1}) + \sum_j m_j(\bar{\underline{y}}_j-\underline{\xi}_j)'\underline{\Sigma}^{-1}(\bar{\underline{y}}_j-\underline{\xi}_j)\} \qquad (5.1)$$

For given $\underline{\Sigma}$, the maximization of (5.1) with respect to the $\underline{\xi}_j$ is equivalent to the minimization of

$$\sum_j m_j(\bar{\underline{y}}_j-\underline{\xi}_j)'\underline{\Sigma}^{-1}(\bar{\underline{y}}_j-\underline{\xi}_j). \qquad (5.2)$$

Following Tintner (1952) we minimize the jth term in (5.2) subject to the constraint

$$\underline{\Gamma}'(\underline{\xi}_j-\underline{\xi}_o) = \underline{0} \qquad (j=1,\ldots,n) \qquad (5.3)$$

by introducing a (p-r)- vector of Lagrange multipliers $\underline{\pi}_j$; the function to be differentiated is

$$(\bar{\underline{y}}_j-\underline{\xi}_j)'\,\underline{\Sigma}^{-1}(\bar{\underline{y}}_j-\underline{\xi}_j) + 2\underline{\pi}_j'\,\underline{\Gamma}'(\underline{\xi}_j-\underline{\xi}_o)\ ,$$

and the vector of derivatives is

$$-2\,\underline{\Sigma}^{-1}(\bar{\underline{y}}_j-\underline{\xi}_j) + 2\underline{\Gamma}\underline{\pi}_j.$$

Equating this to zero we have

$$\bar{\underline{y}}_j-\underline{\xi}_j = \underline{\Sigma}\underline{\Gamma}\underline{\pi}_j \quad ;$$

premultiplication by $\underline{\Gamma}'$ gives

$$\underline{\Gamma}'(\bar{\underline{y}}_j - \underline{\xi}_o) = \underline{\Gamma}'\underline{\Sigma}\underline{\Gamma}\underline{\pi}_j .$$

Eliminating $\underline{\pi}_j$ between the last two equations we have

$$\bar{\underline{y}}_j - \hat{\underline{\xi}}_j = \underline{\Sigma}\ \underline{\Gamma}(\underline{\Gamma}'\underline{\Sigma}\ \underline{\Gamma})^{-1}\ \underline{\Gamma}'\ (\bar{\underline{y}}_j - \underline{\xi}_o), \tag{5.4}$$

and substitution of (5.4) into (5.2) gives

$$\mathrm{tr}\{\underline{\Gamma}'\underline{B}\underline{\Gamma}(\underline{\Gamma}'\underline{\Sigma}\underline{\Gamma})^{-1}\} + \sum_j m_j(\bar{\underline{y}} - \underline{\xi}_o)'\underline{\Gamma}(\underline{\Gamma}'\underline{\Sigma}\underline{\Gamma})^{-1}\ \underline{\Gamma}'(\bar{\underline{y}} - \underline{\xi}_o),$$

where $\underline{B}$ is defined by (2.22). The second term is nonnegative definite in $(\bar{\underline{y}} - \underline{\xi}_o)$, so we have

$$\hat{\underline{\xi}}_o = \bar{\underline{y}} \tag{5.5}$$

unless the hyperplane is constrained to pass through the origin. Thus we have to minimize with respect to $\underline{\Gamma}$

$$\mathrm{tr}\{\underline{\Gamma}'\underline{B}\underline{\Gamma}(\underline{\Gamma}'\underline{\Sigma}\underline{\Gamma})^{-1}\}. \tag{5.6}$$

For a set of relationships through the origin, $\underline{\xi}_o$ is zero and $\underline{B}$ in (5.6) is replaced by the raw between-groups SSP matrix $\underline{B}^*$; unlike the structural case there is no special difficulty in estimating homogeneous relationships. Guessing that the ML solution may be similar to (2.44) or (4.16), we define the p x (p-r) matrix $\underline{V}$ by $\underline{\Gamma} = \underline{K}\underline{V}$ with $\underline{K}$ given by (2.42). Equation (5.6) then becomes

$$\mathrm{tr}\{\underline{\Delta}_{\underline{B}}\ \underline{V}(\underline{V}'\underline{V})^{-1}\underline{V}'\}\ . \tag{5.7}$$

The matrix $\underline{V}(\underline{V}'\underline{V})^{-1}\underline{V}'$ is symmetric, idempotent, and of rank p-r and so has trace equal to p-r and nonnegative diagonal elements ≤ 1, so that the coefficients of the $\delta_i(\underline{B})$ in (5.7) are between 0 and 1 and sum to p-r. The minimum of (5.7) is therefore $\delta_{r+1}(\underline{B})+\ldots+\delta_p(\underline{B})$ or $\mathrm{tr}(\underline{\Delta}_{\underline{B}2})$, and if $\delta_r(\underline{B})>\delta_{r+1}(\underline{B})$ (as it almost surely is if $n > p$), it is attained if and only if

$$\underline{V}(\underline{V}'\underline{V})^{-1}\underline{V}' = \begin{pmatrix} \underline{0} & \underline{0} \\ \underline{0} & \underline{I}_{p-r} \end{pmatrix}\ . \tag{5.8}$$

Postmultiplying by $\underline{V}$ we have

$$\underline{V} = \begin{pmatrix} \underline{0} & \underline{0} \\ \underline{0} & \underline{I}_{p-r} \end{pmatrix} \underline{V} = \begin{pmatrix} \underline{0} \\ \underline{R} \end{pmatrix}, \tag{5.9}$$

say, where $\underline{R}$ is (p-r) x (p-r) and nonsingular. It is easily verified that (5.9) implies (5.8) for any nonsingular $\underline{R}$. It follows from (5.9) that, as expected, the ML estimator of $\underline{\Gamma}$ is

$$\hat{\underline{\Gamma}} = \underline{K}_2 \underline{R} \tag{5.10}$$

with $\underline{R}$ any nonsingular (p-r) x (p-r) matrix. Substituting (5.10) and (5.5) into (5.4) we have

$$\hat{\underline{\xi}}_j = \bar{\underline{y}} + \underline{\Sigma} \underline{K}_1 \underline{K}_1' (\bar{\underline{y}}_j - \bar{\underline{y}}), \tag{5.11}$$

or, for relationships through the origin,

$$\hat{\underline{\xi}}_j = \underline{\Sigma} \underline{K}_1 \underline{K}_1' \bar{\underline{y}}_j$$

with $\underline{K}$ defined relative to $\underline{B}^*$ rather than to $\underline{B}$. Thus we have extended Tintner's (1952) result to the case of replicate observations without his rather dubious use of constraints on the Lagrange multipliers.

The maximum of (5.1) for given $\underline{\Sigma}$ is

$$-\tfrac{1}{2}\{M \log|\underline{\Sigma}| + \mathrm{tr}(\underline{W}\underline{\Sigma}^{-1}) + \mathrm{tr}\,(\underline{\Delta}_{\underline{B}2})\}\ . \tag{5.12}$$

The same results may be obtained using the Type II representation: as we stated in Section 2.2 it is convenient to transform $\bar{\underline{y}}_j$ and $\underline{\xi}_j$ to $\underline{u}_j = \underline{N}_{\underline{B}}^{-1} \bar{\underline{y}}_j$ and $\underline{\zeta}_j = \underline{N}_{\underline{B}}^{-1} \underline{\xi}_j$; this gives (5.2) the simpler form $\Sigma m_j (\underline{u}_j - \underline{\zeta}_j)' (\underline{u}_j - \underline{\zeta}_j)$, and makes the between-groups SSP matrix of the $\underline{u}_j$ equal to the diagonal matrix $\underline{\Delta}_{\underline{B}}$. This procedure is used by Theobald (1975b) to minimize (5.2) and obtain (5.11). The hyperplane on which the $\underline{\xi}_j$ lie is estimated by $\underline{\xi} = \bar{\underline{y}} + \underline{\Sigma} \underline{K}_1 \underline{K}_1' \underline{c}$, where $\underline{c}$ is any p-vector, or, equivalently by $\underline{\xi} = \bar{\underline{y}} + \underline{\Sigma} \underline{K}_1 \underline{a}$ with $\underline{a}$ an arbitrary r-vector. An ML estimator of $\underline{\Gamma}$ therefore satisfies $\hat{\underline{\Gamma}}' \underline{\Sigma} \underline{K}_1 = \hat{\underline{\Gamma}}' \underline{N}_{\underline{B}1} = \underline{0}$, the general

solution of which we have shown in Section 4.2 to be given by (5.10).

5.1.1. $\underline{\Sigma}$ known up to a constant factor

As in Section 4.2.1, we suppose that $\underline{\Sigma}$ has the form $\tau\underline{\Upsilon}$, with τ unknown, and write ϕ_i for the ith root of $|\underline{B}-\phi\underline{\Upsilon}| = 0$ so that $\phi_i = \tau\,\delta_i(\underline{B})$ (i=1,...,p). Apart from terms not involving τ, (5.12) becomes

$$-\tfrac{1}{2}\left[Mp\,\log\tau + \tau^{-1}\{\mathrm{tr}(\underline{W}\underline{\Upsilon}^{-1}) + \mathrm{tr}(\underline{\Phi}_2)\}\right],$$

and this is maximized with respect to τ at

$$\hat{\tau} = (Mp)^{-1}\{\mathrm{tr}(\underline{W}\underline{\Upsilon}^{-1}) + \mathrm{tr}(\underline{\Phi}_2)\}. \tag{5.13}$$

It is immaterial whether we substitute $\hat{\tau}\underline{\Upsilon}$ for $\underline{\Sigma}$ in (5.10) and (5.11) or define $\underline{K}$ with $\underline{\Sigma}$ and $\underline{\Delta}_{\underline{B}}$ replaced by $\underline{\Upsilon}$ and $\underline{\Phi}$ respectively. Replication is not essential in this case. The maximum of the log-likelihood is

$$-\tfrac{1}{2}\,M(\log|\underline{\Upsilon}| + p\,\log\hat{\tau} + p). \tag{5.14}$$

5.1.2 $\underline{\Sigma}$ a general positive-definite matrix

To find the stationary points of (5.1) with respect to variation in $\underline{\Sigma}$ we equate to zero the matrix of derivatives, which equals

$$-\tfrac{1}{2}\left[M\,\underline{\Sigma}^{-1} - \underline{\Sigma}^{-1}\{\underline{W} + \sum_j m_j(\bar{\underline{y}}_j - \underline{\xi}_j)(\bar{\underline{y}}_j - \underline{\xi}_j)'\}\,\underline{\Sigma}^{-1}\right],$$

to obtain

$$M\,\hat{\underline{\Sigma}} = \underline{W} + \sum_j m_j(\bar{\underline{y}}_j - \hat{\underline{\xi}}_j)(\bar{\underline{y}}_j - \hat{\underline{\xi}}_j)'\,. \tag{5.15}$$

It follows from (5.11) that

$$\bar{\underline{y}}_j - \hat{\underline{\xi}}_j = (\underline{I} - \underline{\Sigma}\,\underline{K}_1\,\underline{K}_1')(\bar{\underline{y}}_j - \bar{\underline{y}}) = \underline{\Sigma}\,\underline{K}_2\,\underline{K}_2'\,(\bar{\underline{y}}_j - \bar{\underline{y}}),$$

so the second term on the right-hand side of (5.15) is

$\underline{\Sigma}\,\underline{K}_2\,\underline{K}_2'\,\underline{B}\,\underline{K}_2\,\underline{K}_2'\,\underline{\Sigma}$ or

$$(\underline{K}')^{-1}\begin{pmatrix}\underline{0} & \underline{0}\\ \underline{0} & \underline{\Delta}_{B2}\end{pmatrix}\underline{K}^{-1}\,.$$

If we substitute this into (5.15), premultiply by $\hat{\underline{K}}'$, and post-multiply by $\hat{\underline{K}}$, we obtain

$$\hat{\underline{K}}' \underline{W} \hat{\underline{K}} = M \underline{I} - \begin{pmatrix} \underline{0} & \underline{0} \\ \underline{0} & \hat{\underline{\Delta}}_{\underline{B}2} \end{pmatrix} . \tag{5.16}$$

As in Section 4.2.2, $\hat{\underline{K}}$ diagonalizes $\underline{W}$ as well as $\hat{\underline{\Sigma}}$ and $\underline{B}$, so we try to express $\hat{\underline{K}}$ and $\hat{\underline{\Delta}}_{\underline{B}}$ in terms of the simultaneous reduction of $\underline{W}$ and $\underline{B}$ to diagonal form. It follows from (5.16) that the ψ_i defined in (4.24) are in this case the roots of

$$|\hat{\underline{\Delta}}_{\underline{B}} - \psi\{M \underline{I} - \begin{pmatrix} \underline{0} & \underline{0} \\ \underline{0} & \hat{\underline{\Delta}}_{\underline{B}2} \end{pmatrix}\}| = 0$$

or

$$\prod_{i=1}^{r} \{\delta_i(\underline{B}) - M\psi\} . \prod_{i=r+1}^{p} \{\delta_i(\underline{B}) - M\psi + \delta_i(\underline{B})\psi\} = 0,$$

so that for some permutation ρ of $(1,\ldots,p)$ we have

$$\left.\begin{aligned} \hat{\delta}_i(\underline{B}) &= M\psi_{\rho(i)} \quad (i=1,\ldots,r), \\ \hat{\delta}_i(\underline{B}) &= M\psi_{\rho(i)}(1+\psi_{\rho(i)})^{-1} \quad (i=r+1,\ldots,p) . \end{aligned}\right\} \tag{5.17}$$

We cannot immediately put $\rho(i) = i$ since it may happen that $\psi_s > \psi_r(1+\psi_r)^{-1}$ even though $s>r$. But substituting (5.15) into (5.1) gives $-\frac{1}{2} M(\log|\hat{\underline{\Sigma}}| + p)$, so that we have to minimize $|\hat{\underline{\Sigma}}|$ or, since $|\underline{B}| = |\hat{\underline{\Sigma}}||\hat{\underline{\Delta}}_{\underline{B}}|$, to maximize

$$|\hat{\underline{\Delta}}_{\underline{B}}| = M^p |\underline{\Psi}| \Big/ \prod_{i=r+1}^{p} (1 + \psi_{\rho(i)}) .$$

If $\psi_r > \psi_{r+1}$ then this is achieved if and only if ρ maps the set $(r+1,\ldots,p)$ onto itself; we are then able to write

$$\hat{\underline{\Delta}}_{\underline{B}1} = M \underline{\Psi}_1 , \qquad \hat{\underline{\Delta}}_{\underline{B}2} = M \underline{\Psi}_2(\underline{I} + \underline{\Psi}_2)^{-1} . \tag{5.18}$$

To obtain $\hat{\underline{K}}$ in terms of $\underline{\Psi}$ and $\underline{G}$ we define $\underline{H}$ as $\underline{G}^{-1}\,\hat{\underline{K}}$ and rewrite (4.26) as

$$\underline{G}'\,\underline{W}\,\underline{G} = \underline{I} \quad , \qquad \underline{G}'\,\underline{B}\,\underline{G} = \underline{\Psi} \quad . \tag{5.19}$$

Substitution of the first part of (5.19) and the second part of (5.18) into (5.16) gives

$$\underline{H}'\,\underline{H} = M\begin{pmatrix} \underline{I} & \underline{0} \\ \underline{0} & (\underline{I}+\underline{\Psi}_2)^{-1} \end{pmatrix} \quad , \tag{5.20}$$

and from the second parts of (2.42) and (5.19) we have

$$\underline{H}'\,\underline{\Psi}\,\underline{H} = \hat{\underline{\Delta}}_B = M\begin{pmatrix} \underline{\Psi}_1 & \underline{0} \\ \underline{0} & \underline{\Psi}_2(\underline{I}+\underline{\Psi}_2)^{-1} \end{pmatrix} . \tag{5.21}$$

Our argument then follows the same lines as that of Section 4.2.2: the general solution of (5.20) and (5.21) is

$$\underline{H} = M^{\frac{1}{2}}\begin{pmatrix} \underline{J}_{11} & \underline{0} \\ \underline{0} & \underline{J}_{22}(\underline{I}+\underline{\Psi}_2)^{-\frac{1}{2}} \end{pmatrix} \quad , \tag{5.22}$$

where $\underline{J}_{11}$ and $\underline{J}_{22}$ are orthogonal matrices which commute with $\underline{\Psi}_1$ and $\underline{\Psi}_2$ respectively; since $\hat{\underline{K}} = \underline{G}\,\underline{H}$ we have

$$\hat{\underline{K}}_1 = M^{\frac{1}{2}}\,\underline{G}_1\,\underline{J}_{11} \quad , \qquad \hat{\underline{K}}_2 = M^{\frac{1}{2}}\,\underline{G}_2\,\underline{J}_{22}(\underline{I}+\underline{\Psi}_2)^{-\frac{1}{2}} \quad . \tag{5.23}$$

The ML estimator of $\underline{\Gamma}$ is, following (5.10), $\hat{\underline{\Gamma}} = \hat{\underline{K}}_2\,\underline{R}$, which becomes

$$\hat{\underline{\Gamma}} = \underline{G}_2\,\underline{R} \tag{5.24}$$

when the remaining terms in the second part of (5.23) are subsumed into $\underline{R}$. As in the case of known $\underline{\Sigma}$, the estimators of $\underline{\Gamma}$ are the same for functional as for structural relationships when there are equal numbers of replicates. For the estimator of $\underline{\Sigma}$ we have from (5.22) and (4.38)

$$\hat{\underline{\Sigma}} = (\hat{\underline{K}}\,\hat{\underline{K}}')^{-1} = (\underline{G}\,\underline{H}\,\underline{H}'\,\underline{G}')^{-1} = M^{-1}\{\underline{W} + \underline{L}_2\,\underline{\Psi}_2\,\underline{L}_2'\} \quad . \tag{5.25}$$

The estimator of $\underline{\xi}_j$ is, following (5.11), $\bar{\underline{y}} + \hat{\underline{\Sigma}}\,\hat{\underline{K}}_1\,\hat{\underline{K}}_1'\,(\bar{\underline{y}}_j - \bar{\underline{y}})$,

which becomes

$$\hat{\underline{\xi}}_j = \bar{\underline{y}} + \underline{W}\,\underline{G}_1\,\underline{G}_1'\,(\bar{\underline{y}}_j - \bar{\underline{y}}) \qquad (5.26)$$

or

$$\hat{\underline{\xi}}_j = \bar{\underline{y}}_j - \underline{W}\,\underline{G}_2\,\underline{G}_2'\,(\bar{\underline{y}}_j - \bar{\underline{y}}) \quad . \qquad (5.27)$$

For relationships through the origin we delete $\bar{\underline{y}}$ in these two equations and define $\underline{G}$ relative to $\underline{B}^*$ rather than $\underline{B}$.

The maximum of the log-likelihood is

$$-\tfrac{1}{2}\,M\{\log|\underline{W}| + \sum_{i=r+1}^{p} \log(1+\psi_i) + p - p\log M\} \; . \qquad (5.28)$$

It is essential to have replicate observations in this case.

5.1.3 $\underline{\Sigma}$ a diagonal matrix

Most of Section 4.2.3 on the corresponding structural relationship problem applies equally here. The only exceptions are that — as we showed in Section 2.4(a) — it is essential to have replication for all values of r, and that the log-likelihood is, from (5.12), $-\tfrac{1}{2}\,F_4$, where

$$F_4 = \sum_{i=1}^{p} (M \log \sigma_{ii} + w_{ii}/\sigma_{ii}) + \sum_{i=r+1}^{p} \delta_i(\underline{B}) \; . \qquad (5.29)$$

5.2 Tests of the number of relationships

As in Section 4.3, we may base a test of the hypothesis of p-r functional relationships on the difference between the maxima of the log-likelihood when there are no relationships and when there are p-r. In this case the differences are, from (5.12), (5.14), (5.28) and (5.29):

for known $\underline{\Sigma}$,

$$\tfrac{1}{2} \sum_{i=r+1}^{p} \delta_i(\underline{B}) \quad ; \qquad (5.30)$$

for $\underline{\Sigma}$ known up to a constant factor

$$\tfrac{1}{2} Mp \log\{1 + \sum_{i=r+1}^{p} \phi_i / \mathrm{tr}(\underline{W}\underline{\Upsilon}^{-1})\} ;$$

for a general positive-definite $\underline{\Sigma}$,

$$\tfrac{1}{2} M \sum_{i=r+1}^{p} \log(1+\psi_i) \quad ; \qquad (5.31)$$

and for $\underline{\Sigma}$ diagonal,

$$\tfrac{1}{2}\{M(\sum_{i=1}^{p} \log \hat{\sigma}_{ii} + p) - F_4, \min\}.$$

Twice each of these differences is distributed asymptotically as χ^2 under the null hypothesis of p-r relationships. Since the asymptotic results of ML theory do not apply in the presence of incidental parameters, we mean by 'asymptotically' that n is fixed and each m_j tends to infinity. For relationships through the origin the degrees of freedom are $(p-r)(n-r)$, since the number of conditions applied to the $\underline{\xi}_j$ by writing $\underline{\Gamma}'\underline{\xi}_j = \underline{0}$ $(j=1,\ldots,n)$ is reduced from $n(p-r)$ to zero when there are no relationships, while the number of independent parameters in $\underline{\Gamma}$ is reduced from $r(p-r)$ to zero. Relationships not constrained in this way may be written with only p-r extra parameters as $\underline{\Gamma}'\underline{\xi}_j + \underline{\beta} = \underline{0}$ $(j=1,\ldots,n)$, so the degrees of freedom are $(p-r)(n-r-1)$. The test based on (5.30) is due to Fisher (1938), and a refinement of (5.31) is Bartlett's (1947, p.179) test of the significance of the last p-r roots in a canonical analysis; M in (5.31) is replaced by $\{M-1-\frac{1}{2}(p+n)\}$ to give a better approximation to the χ^2 distribution.

5.3 Generalized least-squares estimation

We have seen in Section 2.5 that Sprent's generalized least-squares (GLS) procedure for estimating a single linear functional relationship may be extended from two to p variates and, when p is even, from one

relationship to $\frac{1}{2}p$, and that it coincides with the ML procedure for estimating $\underline{\Gamma}$ in these cases if normality is assumed and $\underline{\Sigma}$ is known. We give below a formal justification for Sprent's intuitive extension of the method to any number of relationships, and show its equivalence to ML under the assumption of normality when $\underline{\Sigma}$ is either known to within a constant factor or is a general positive-definite matrix estimated from replicate observations. In Section 5.5 we give, inter alia, sufficient conditions, not including normality, for the method to be consistent.

Instead of the scalar null variate $z_j = \underline{\gamma}' \bar{\underline{y}}_j$ of Section 1.5 we define a (p-r)- vector of null variates by $\underline{z}_j = \underline{\Gamma}' \bar{\underline{y}}_j (j+1,\ldots,n)$, the $\bar{\underline{y}}_j$ being measured from $\bar{\underline{y}}$ unless the relationships are constrained to pass through the origin. The $\underline{z}_j$ are uncorrelated with zero mean and dispersion matrix $m_j^{-1}\ \underline{\Gamma}'\underline{\Sigma}\underline{\Gamma}$. Instead of (2.47) we have to minimize

$$U = \sum_j \underline{z}_j' \{D(\underline{z}_j)\}^{-1} \underline{z}_j$$

$$= \mathrm{tr}\{\underline{\Gamma}'\underline{B}\underline{\Gamma}(\underline{\Gamma}'\underline{\Sigma}\underline{\Gamma})^{-1}\} \quad .$$

Since this is the same as (5.6), the GLS estimator of $\underline{\Gamma}$ when $\underline{\Sigma}$ is known up to a constant factor is given by (5.10). When $\underline{\Sigma}$ is a general positive-definite matrix and has to be estimated from replicate observations its GLS estimator is $\underline{W}/(M-n)$, so that $\underline{K}_2$ in (5.10) is replaced by $\underline{G}_2$ and we have (5.24). In both cases the relationships are estimated by equating to zero the last p-r canonical variates in a canonical analysis using 'within' and 'between' SSP matrices.

5.4 Asymptotic behaviour

The presence of incidental parameters prevents us from appealing to the usual asymptotic properties of ML. As in Chapter 4, we consider

two ways in which the total number of observations may tend to infinity; in this case they are (i) that M/n tends to infinity (which includes the possibility that n remains finite), and (ii) that M/n tends to a finite limit m>1. We have also to make some assumption about the behaviour of the $\underline{\xi}_j$: in the case of two variates and no replication Kendall and Stuart (1967, p.387) assume that the ξ_{1j} have a finite 'variance', that is that $n^{-1}\Sigma(\xi_{1j}-\bar{\xi}_1)^2$ has a finite positive limit. To extend this idea to p variates, p-r functional relationships, and replicate observations, we suppose that the matrix $M^{-1}\Sigma m_j(\underline{\xi}_j-\bar{\underline{\xi}})(\underline{\xi}_j-\bar{\underline{\xi}})'$ has a finite limit $\underline{E}$ of rank r, and that each $(\underline{\xi}_j-\bar{\underline{\xi}})$ is in the range space of $\underline{E}$; this is bound to be true in case (i) if n remains finite and the m_j are equal. For relationships through the origin $\bar{\underline{\xi}}$ is omitted. We shall not require the $\underline{Y}_{je}$ to be normal but, as in Section 5.3, take them to be uncorrelated with a common dispersion matrix $\underline{\Sigma}$. If we then assume that either the $\underline{Y}_{je}$ are independent or that the variates $Y_{ije}\ Y_{hje}$ (e=1,...,m_j ; j=1,...,n; i,h=1,...,p) have finite variance independent of j and e and zero covariance between replicates and between groups, then in case (i)

$$M^{-1}\ \underline{W}\to \underline{\Sigma}_o \qquad \text{and} \qquad M^{-1}\ \underline{B}\to \underline{E}, \tag{5.32}$$

and in case (ii)

$$M^{-1}\ \underline{W}\to (1-m^{-1})\underline{\Sigma}_o \quad \text{and} \quad M^{-1}\ \underline{B}\to \underline{E}+m^{-1}\ \underline{\Sigma}_o\ , \tag{5.33}$$

where $\underline{\Sigma}_o$ denotes the true (or known) value of $\underline{\Sigma}$. Let $\omega_1(\underline{E}),\ldots,\omega_r(\underline{E})$ be the nonzero roots in descending order of $|\underline{E}-\omega\ \underline{\Sigma}_o| = 0$, $\underline{\Omega}_{\underline{E}}$ the p×p matrix diag $\{\omega_1(\underline{E}),\ldots,\omega_r(\underline{E}),0,\ldots,0\}$, $\underline{\Omega}_{\underline{E}1}$ the matrix diag$\{\omega_1(\underline{E}),\ldots,\omega_r(\underline{E})\}$, $\underline{F}$ a p×p matrix satisfying

$$\underline{F}'\underline{\Sigma}_o\ \underline{F} = \underline{I}\ , \qquad \underline{F}'\underline{E}\underline{F} = \underline{\Omega}_{\underline{E}}\ , \tag{5.34}$$

and $\underline{P}$ the p×p matrix $(\underline{K}^{-1}\ \underline{F})'$; $\underline{P}$ is orthogonal since $\underline{K}'\ \underline{\Sigma}\ \underline{K} = \underline{F}'\ \underline{\Sigma}\ \underline{F} = \underline{I}$. If we partition $\underline{F}$ as $(\underline{F}_1\ \underline{F}_2)$ with $\underline{F}_1$ p×p then $\underline{F}_2'\ \underline{E}\ \underline{F}_2 = \underline{O}$ and

$$\underline{F}_2'\ (\underline{\xi}_j - \bar{\underline{\xi}}) = \underline{O} \qquad (j=1,\ldots,n)\ . \tag{5.35}$$

To demonstrate the consistency of $\hat{\underline{\Gamma}}$ when $\underline{\Sigma}$ is known we have to show that as M tends to infinity $\underline{K}_2'\ \underline{E}\ \underline{K}_2 \to \underline{O}$. In case (i) it follows from the second parts of (5.32) and (2.42) that $M^{-1}\ \underline{F}'\ \underline{B}\ \underline{F} \to \underline{\Omega}_{\underline{E}}$ and

$$M^{-1}\ \underline{P}\ \underline{\Delta}_{\underline{B}}\ \underline{P}' \to \underline{\Omega}_{\underline{E}}\ . \tag{5.36}$$

Since the latent roots of a matrix are continuous functions of its elements we also have

$$M^{-1}\ \underline{\Delta}_{\underline{B}} \to \underline{\Omega}_{\underline{E}}\ . \tag{5.37}$$

Thus if we write

$$M^{-1}\ \underline{P}\ \underline{\Delta}_{\underline{B}}\ \underline{P}' = \underline{P}\ \underline{\Omega}_{\underline{E}}\ \underline{P}' + \underline{P}(M^{-1}\ \underline{\Delta}_{\underline{B}} - \underline{\Omega}_{\underline{E}})\underline{P}'\ , \tag{5.38}$$

then the second term on the right-hand side tends to zero. Combining (5.36) and (5.38) we have $\underline{P}\underline{\Omega}_{\underline{E}}\underline{P}' \to \underline{\Omega}_{\underline{E}}$ or

$$\underline{P}\underline{\Omega}_{\underline{E}} - \underline{\Omega}_{\underline{E}}\underline{P} \to \underline{O}\ . \tag{5.39}$$

So if $\underline{P}$ is partitioned as

$$\underline{P} = (\underline{P}_1\ \underline{P}_2) = \begin{pmatrix} \underline{P}_{11} & \underline{P}_{12} \\ \underline{P}_{21} & \underline{P}_{22} \end{pmatrix}\ , \tag{5.40}$$

with $\underline{P}_1$ p × r and $\underline{P}_{11}$ r × r, we must have $\underline{P}_{12} \to 0$ and $\underline{P}_{21} \to 0$. Since $\underline{K} = \underline{F}\ \underline{P}$, we have $\underline{K}_2 = \underline{F}\ \underline{P}_2$ and, using the second part of (5.34),

$$\underline{K}_2'\ \underline{E}\ \underline{K}_2 = \underline{P}_2'\ \underline{\Omega}_{\underline{E}}\ \underline{P}_2 = \underline{P}_{12}'\ \underline{\Omega}_{\underline{E}1}\ \underline{P}_{12} \to \underline{O} \tag{5.41}$$

as M tends to infinity. In case (ii) we have $M^{-1}\ \underline{P}\ \underline{\Delta}_B\ \underline{P}' \to \underline{\Omega}_E + m^{-1}\ \underline{I}$ and

$$M^{-1}\ \underline{\Delta}_B \to \underline{\Omega}_E + m^{-1}\ \underline{I}, \tag{5.42}$$

but (5.39) and (5.41) are unaffected.

Only minor alterations are required to show that $\hat{\underline{\Gamma}}$ is also consistent when $\underline{\Sigma}$ is known up to a constant factor. In case (i) we have from (5.32) and (5.37) that $M^{-1}\ \mathrm{tr}(\underline{W}\underline{\Upsilon}^{-1}) \to \tau p$ and $M^{-1}\ \mathrm{tr}(\underline{\Phi}_2) \to 0$; substitution into (5.13) shows that $\hat{\tau}$ is consistent. However, in case (ii), (5.33) and (5.42) give $M^{-1}\mathrm{tr}(\underline{W}\underline{\Upsilon}^{-1}) \to \tau p(1-m^{-1})$ and $M^{-1}\mathrm{tr}(\underline{\Phi}_2) \to \tau(p-r)/m$, so that

$$\hat{\tau} \to \tau\{1-r(mp)^{-1}\}\ . \tag{5.43}$$

The special case of this result with $m_j = 1$ was noted in Section 2.5.

When $\underline{\Sigma}$ is a general, unknown, positive-definite matrix its GLS estimator $\underline{W}/(M-n)$ is consistent in both cases, so the estimators of $\underline{\Gamma}$ are also consistent. The ML estimators of $\underline{\Sigma}$ and $\underline{\Gamma}$ are consistent in case (i) since (5.32) and (5.25) imply that $\underline{\Psi}_2 \to \underline{0}$ and $\hat{\underline{\Sigma}} \to \underline{\Sigma}_o$. In case (ii) the ψ_i tend in probability to the roots of

$$|\underline{E} + m^{-1}\ \underline{\Sigma}_o - \psi(1 - m^{-1})\underline{\Sigma}_o| = 0,$$

that is, using (5.34), to the roots of

$$|\underline{\Omega}_E + m^{-1}\ \underline{I} - \psi(1 - m^{-1})\ \underline{I}| = 0,$$

so we have

$$\underline{\Psi} \to (m-1)^{-1}\ (m\underline{\Omega}_E + \underline{I}). \tag{5.44}$$

If $\underline{Q} = \{(1 - m^{-1})\ M\}^{-\frac{1}{2}}\ \underline{G}^{-1}\ \underline{F}$ then the first parts of each of (5.19), (5.33) and (5.34) show that in the limit $\underline{Q}$ is orthogonal. Similarly, equations (5.34) and (5.44) and the second parts of (5.19) and (5.33) show that in the limit $\underline{Q}$ commutes with $m\underline{\Omega}_E + \underline{I}$, so if $\underline{Q}$ is partitioned

in the same way as $\underline{P}$ (equation 5.40) then $\underline{Q}_{12} \to \underline{0}$ and $\underline{Q}_{21} \to \underline{0}$. Since $\underline{G}_2$ is given by $\{(1-m^{-1})M\}^{-\frac{1}{2}}$ $(\underline{F}_1 \underline{Q}'_{21} + \underline{F}_2 \underline{Q}'_{22})$, and $\underline{F}'_2 \underline{E} \underline{F}_2 = \underline{0}$, it follows that $M \underline{G}'_2 \underline{E} \underline{G}_2 \to \underline{0}$ and therefore, using (5.23), that $\hat{\underline{K}}'_2 \underline{E} \hat{\underline{K}}_2 \to \underline{0}$, so that $\hat{\underline{\Gamma}}$ is consistent. Also, since (5.44) implies that $\underline{\Psi}_2 \to (m-1)^{-1} \underline{I}$, the limit of $M \underline{G}_2 \underline{\Psi}_2 \underline{G}'_2$ is $m^{-1} \underline{F}_2 \underline{F}'_2$. Thus from (5.25) and (5.33)

$$\hat{\underline{\Sigma}} = M^{-1}\{\underline{W} + \underline{W} \underline{G}_2 \underline{\Psi}_2 \underline{G}'_2 \underline{W}\} \to (1-m^{-1})\underline{\Sigma}_o + m^{-1}(1-m^{-1})^2 \underline{\Sigma}_o \underline{F}_2 \underline{F}'_2 \underline{\Sigma}_o , \qquad (5.45)$$

so that in this case $\hat{\underline{\Sigma}}$ is not consistent.

When $\underline{\Sigma}$ is diagonal we have to minimize with respect to the σ_{ii}

$$\sum_{i=1}^{p} \{\log \sigma_{ii} + w_{ii}/(M \sigma_{ii})\} + M^{-1} \operatorname{tr}(\underline{\Delta}_{B2}). \qquad (5.46)$$

In case (i) (5.32) and (5.37) show that $M^{-1} w_{ii}$ tends to the true value σ_{oii} and that $M^{-1} \underline{\Delta}_{B2} \to \underline{0}$; in the limit (5.46) becomes

$$\sum_{i=1}^{p} (\log \sigma_{ii} + \sigma_{oii} \sigma_{ii}^{-1}) ,$$

and, since this is minimized at $\hat{\sigma}_{ii} = \sigma_{oii} (i=1,\ldots,p)$, $\hat{\underline{\Sigma}}$ is again consistent. In case (ii) (5.33) and (5.42) show that (5.46) tends to

$$\sum_{i=1}^{p} \{\log \sigma_{ii} + (1 - m^{-1}) \sigma_{oii} \sigma_{ii}^{-1}\} + (p-r)/m ,$$

which is minimized at $\hat{\sigma}_{ii} = (1 - m^{-1}) \sigma_{oii}$. Thus $\hat{\underline{\Sigma}}$ tends in probability to $(1 - m^{-1})\underline{\Sigma}_o$. Since in both cases the limit of $\hat{\underline{\Sigma}}$ is a multiple of $\underline{\Sigma}_o$ the estimator of $\underline{\Gamma}$ is also consistent.

Finally, we examine the consistency of $\hat{\underline{\xi}}_j$, which is estimated by $\bar{\underline{y}}_j - \underline{\Sigma} \underline{K}_2 \underline{K}'_2 (\bar{\underline{y}}_j - \bar{\underline{y}})$ when $\underline{\Sigma}$ is known and by $\bar{\underline{y}}_j - \hat{\underline{\Sigma}} \hat{\underline{K}}_2 \hat{\underline{K}}'_2 (\bar{\underline{y}}_j - \bar{\underline{y}})$ when $\underline{\Sigma}$ is unknown. Since $\hat{\underline{\Gamma}}$ is consistent, we have $\underline{K}'_2 \underline{E} \underline{K}_2 \to \underline{0}$ or $\hat{\underline{K}}'_2 \underline{E} \hat{\underline{K}}_2 \to \underline{0}$

and therefore, since each $(\underline{\xi}_j - \bar{\underline{\xi}})$ is in the range space of $\underline{E}$, $\underline{K}_2'(\underline{\xi}_j - \bar{\underline{\xi}}) \to \underline{0}$ or $\hat{\underline{K}}_2'(\underline{\xi}_j - \bar{\underline{\xi}}) \to 0$ $(j=1,\ldots,n)$. The mean vector and dispersion matrix of $\bar{\underline{Y}}$ are $\bar{\underline{\xi}}$ and $M^{-1}\underline{\Sigma}$, while those of $\underline{Y}_j$ are $\underline{\xi}_j$ and $m_j^{-1}\underline{\Sigma}$, so $\bar{\underline{Y}}$ tends to $\bar{\underline{\xi}}$ and, if m_j tends to infinity, $\bar{\underline{Y}}_j$ and $\hat{\underline{\xi}}_j$ both tend to $\underline{\xi}_j$. For relationships through the origin we just delete $\bar{\underline{y}}$ and $\bar{\underline{\xi}}$ in the above.

In summary, we have shown that the GLS estimators of $\underline{\Gamma}$ and $\underline{\Sigma}$ are consistent in cases (i) and (ii). The ML estimators of $\underline{\Gamma}$ are consistent in both cases, but those of $\underline{\Sigma}$ have this property only in case (i). For those groups in which m_j tends to infinity, $\hat{\underline{\xi}}_j$ is also consistent.

6. COMPARATIVE CALIBRATION

In this Chapter we consider in detail the structural-relationship model for the comparative calibration of p measuring instruments which was introduced in Section 2.3, and re-analyse the data on the measurement of human lung capacity given by Barnett (1969). Some residual analyses and probability plots are suggested for indicating departures from the assumptions of the model, and are illustrated on these data.

6.1 The calibration model

Our model for the comparative calibration of p instruments used to measure the same property is that on the j^{th} specimen to be measured, the p-vector $\underline{Y}_j$ of observations is the sum of a vector $\underline{X}_j$ of true values, distributed normally over the population of possible specimens, and an independent vector of errors $\underline{\varepsilon}_j$, the elements of which are independently distributed as $N(0, \sigma_i^2)$ $(i = 1, \ldots, p)$, the σ_i not depending on j. The n vectors $\underline{Y}_j$ are independent, and the distribution of the $\underline{X}_j$ is confined to a line. The degeneracy is conveniently expressed as

$$\underline{X} = \underline{\mu} + \underline{\lambda} f , \qquad (6.1)$$

where $\underline{\mu} = E(\underline{X})$, $\underline{\lambda}$ is a p-vector of unknown 'calibration factors' and f is a hypothetical standard measurement distributed as N(0, 1) over the population. The $\underline{Y}_j$ are thus distributed as $N_p(\underline{\mu}, \underline{\Sigma} + \underline{\lambda}\underline{\lambda}')$, where $\underline{\Sigma}$ is $\mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$.

The elimination of f from (6.1) shows that the true values of any two instruments h and i are related by

$$X_h = \mu_h + (X_i - \mu_i)\, \lambda_h/\lambda_i ,$$

and therefore vary in the ratio $\lambda_h:\lambda_i$. When $\lambda_h = \lambda_i$ the true measurements differ by a simple additive bias; if also $\mu_h = \mu_i$ the true measurements are the same. If a linear calibration model of the above type is thought appropriate and the instruments are to be compared

for their precision as well as their relative calibration, then an obvious index of this property is the ratio of the variances of the i^{th} observed measurement and the corresponding error term, that is, $1 + \lambda_i^2/\sigma_i^2$. So we formally define the *precision* π_i of the i^{th} instrument to be λ_i/σ_i ; the vector $\underline{\pi} = (\pi_1, \ldots, \pi_p)'$ is given by

$$\underline{\pi} = \underline{\Sigma}^{-\frac{1}{2}} \underline{\lambda} .$$

We may therefore be interested in testing whether the π_i are equal, or the λ_i, or the λ_i and the μ_i.

6.2 Maximum likelihood estimation

Our model is equivalent to a factor model with a single common factor; the criterion (2.33) for such a model to be identifiable becomes, with r equal to one, $p \geq 3$. Thus the parameters of the model may be estimated using a program for carrying out factor analysis by ML, for example that of Jöreskog (1966). When $p = 3$ the numbers of minimal-sufficient statistics and independent parameters are equal, and equating sample and population moments gives the ML solution

$$\hat{\mu}_1 = \bar{y}_1 , \quad \hat{\lambda}_1^2 = s_{12}\, s_{13}/s_{23} , \quad \hat{\sigma}_1^2 = s_{11} - \hat{\lambda}_1^2$$

etc., where

$$\underline{S} = (s_{hi}) = n^{-1} \underline{B} = n^{-1} \sum_j (\underline{y}_j - \bar{\underline{y}})(\underline{y}_j - \bar{\underline{y}})' .$$

As Moran (1971) has pointed out, equating moments may not always give a solution; in any sensible application of the method all the elements of $\underline{S}$ will be positive and so therefore will be the expressions for $\hat{\lambda}_i^2$, but if one of the σ_i^2 is very small there is a danger that its estimator will be negative.

As we noted in Section 2.4(b), our model leads to a particular case of Jöreskog's 'covariance structures'. The special cases in which the π_i, the λ_i or the μ_i are taken to be equal are also included in this general formulation, so we could carry out estimation in all these cases

using the program for ML estimation given by Jöreskog et al. (1971). Nevertheless, to illustrate the procedure derived in Chapter 4 we consider the estimation of the basic model, with no constraints, in some detail.

We define the matrix $\underline{S}_\sigma$ as $\underline{\Sigma}^{-\frac{1}{2}} \underline{S} \underline{\Sigma}^{-\frac{1}{2}}$ and write χ_i for $n^{-1} \delta_i(\underline{B})$; since $\delta_i(\underline{B})$ is the i^{th} root of $|\underline{B} - \delta\underline{\Sigma}| = 0$, χ_i is the i^{th} latent root of $\underline{S}_\sigma$. We have shown in Section 4.2.3 that if $\underline{C}$ is the $p \times p$ matrix whose columns are the latent vectors of $\underline{S}_\sigma$ then, for given $\underline{\Sigma}$, $\underline{N}_{\underline{B}} = \underline{\Sigma}^{\frac{1}{2}} \underline{C}$. Since $m = r = 1$, we have from (4.14)

$$\hat{\underline{\lambda}} = (\hat{\chi}_1 - 1)^{\frac{1}{2}} \hat{\underline{n}}_1 ,$$

where $\hat{\underline{n}}_1$ is the first column of $\hat{\underline{N}}_{\underline{B}} = \hat{\underline{\Sigma}}^{\frac{1}{2}} \underline{C}$. Thus

$$\hat{\underline{\lambda}} = (\hat{\chi}_1 - 1)^{\frac{1}{2}} \hat{\underline{\Sigma}}^{\frac{1}{2}} \hat{\underline{c}}_1 \quad \text{and} \quad \hat{\underline{\pi}} = (\hat{\chi}_1 - 1)^{\frac{1}{2}} \hat{\underline{c}}_1 ,$$

while the $\hat{\mu}_i$ are the sample means. The diagonal elements of $\hat{\underline{\Sigma}}^{\frac{1}{2}}$ are the values of σ_i $(i = 1, \ldots, p)$ minimizing (4.17), which in this case equals

$$F_4 = n \left\{ \sum_{i=1}^{p} (\log \sigma_i^2 + s_{ii} \sigma_i^{-2}) + \log \chi_1 - \chi_1 + 1 \right\}. \qquad (6.2)$$

In order to use a gradient method for this minimization we require the first derivatives of F_4. Bellman (1970, pp.61-63) shows that if a symmetric $p \times p$ matrix $\underline{A} = (a_{st})$ has a simple latent root α with corresponding latent vector $\underline{x} = (x_1, \ldots, x_p)'$ then $\partial\alpha/\partial a_{ss} = x_s^2$ and $\partial\alpha/\partial a_{st} = 2 x_s x_t$ $(s,t = 1, \ldots, p ; s \neq t)$. Applying this result to $\underline{S}_\sigma$ and χ_1 we have

$$\frac{\partial \chi_1}{\partial \sigma_\ell^2} = - \frac{c_{\ell 1}}{\sigma_\ell^3} \sum_{i=1}^{p} \frac{s_{\ell i} c_{i1}}{\sigma_i} = - \frac{\chi_1 c_{\ell 1}^2}{\sigma_\ell^2} \qquad (\ell = 1, \ldots, p)$$

where $\underline{c}_1 = (c_{11}, \ldots, c_{p1})'$, so the derivatives of (6.2) are given by

$$\frac{\partial F_4}{\partial \sigma_\ell^2} = \frac{n}{\sigma_\ell^2} \left\{ (\chi_1 - 1) c_{\ell 1}^2 + 1 - \frac{s_{\ell\ell}}{\sigma_\ell^2} \right\} \quad (\ell = 1, \ldots, p) . \qquad (6.3)$$

The maximum of the log-likelihood is $-\frac{1}{2} F_{4,\min}$.

As we have seen already in the case of three instruments, the derivatives (6.3) may vanish at a point outside the region in which all the σ_i^2 are positive, so that the likelihood attains its maximum at a boundary point of the parameter space, that is, where one of the σ_i is zero. Lawley and Maxwell (1971, p.32) report that such improper solutions occur quite frequently in factor analysis. Only one of the σ_i may be estimated as zero since any greater number would make the estimated dispersion matrix of $\underline{Y}$ singular. If it is, say, σ_q then the ML solution amounts to regressing the readings of the remaining instruments on those of the q^{th}.

For the asymptotic dispersion matrix of the ML estimators we require the second derivatives of the log-likelihood with respect to the parameters μ_i, λ_i and σ_i $(i = 1, \ldots, p)$. Let η_s, η_t denote single parameters or vectors of parameters, and write $d(\eta_s, \eta_t)$ for the value of $E\{-\partial/\partial\eta_t(\partial L/\partial\eta_s)\}$ at the likelihood maximum. If h and i are in the range $(1, \ldots, p)$, $h \neq i$, $b_{hi} = n(\hat{\chi}_1^2 \hat{\sigma}_h \hat{\sigma}_i)^{-1}$ and $b_{ii} = n(\hat{\chi}_1^2 \hat{\sigma}_i^2)^{-1}$, then we have

$$d(\underline{\mu}, \underline{\mu}) = n(\hat{\underline{\Sigma}} + \hat{\underline{\lambda}}\hat{\underline{\lambda}}')^{-1} , \tag{6.4}$$

$$\left.\begin{aligned}
d(\underline{\mu}, \eta) &= \underline{0} \quad \text{if } \eta \text{ is not in } \underline{\mu} , \\
d(\lambda_i, \lambda_i) &= b_{ii}\{\hat{\chi}_1(\hat{\chi}_1 - 1) + (2 - \hat{\chi}_1)\hat{\pi}_i^2\} , \\
d(\lambda_h, \lambda_i) &= b_{hi}\hat{\pi}_h\hat{\pi}_i(2 - \hat{\chi}_1) , \\
d(\lambda_i, \sigma_i) &= 2 b_{ii}\hat{\pi}_i(\hat{\chi}_1 - \hat{\pi}_i^2) , \\
d(\lambda_h, \sigma_i) &= -2 b_{hi}\hat{\pi}_h\hat{\pi}_i , \\
d(\sigma_i, \sigma_i) &= 2 b_{ii}(\hat{\chi}_1 - \hat{\pi}_i^2)^2 , \\
d(\sigma_h, \sigma_i) &= 2 b_{hi}\hat{\pi}_h^2\hat{\pi}_i^2 .
\end{aligned}\right\} \tag{6.5}$$

We next consider, in less detail, ML estimation in some situations in which parameters are constrained to be equal. If the λ_i have a

common value λ then $\underline{\mu}$ is estimated by the sample mean and we have

$$\hat{\lambda}^2 = (\phi - 1)/\sum_i \sigma_i^{-2} , \qquad (6.6)$$

where

$$\phi = \sum_h \sum_i s_{hi}\, \sigma_h^{-2} \sigma_i^{-2} / \sum_i \sigma_i^{-2} ,$$

and the maximum of the log-likelihood for given σ_i $(i = 1, \ldots, p)$ is

$$- \tfrac{1}{2} n \{\sum_i (\log \sigma_i^2 + s_{ii}\, \sigma_i^{-2}) + \log \phi - \phi + 1\} .$$

If the μ_i also have a common value μ then

$$\hat{\mu} = \sum_i \bar{y}_i\, \sigma_i^{-2} / \sum_i \sigma_i^{-2} ,$$

while $\hat{\lambda}$ is again given by (6.6). The conditional maximum of the log-likelihood is

$$- \tfrac{1}{2} n \{\sum_i (\log \sigma_i^2 + s_{ii}\, \sigma_i^{-2} + \bar{y}_i^2 \sigma_i^{-2}) - (\sum_i \bar{y}_i\, \sigma_i^{-2})^2 / \sum_i \sigma_i^{-2} + \log \phi - \phi + 1\} .$$

Under the constraint that the π_i have a common value π the ML estimators of $\underline{\mu}$ and π are given by

$$\hat{\underline{\mu}} = \bar{\underline{y}} \quad \text{and} \quad \hat{\pi}^2 = (s_\sigma - 1)/p ,$$

where

$$s_\sigma = \sum_h \sum_i s_{hi}\, \sigma_h^{-1} \sigma_i^{-1} / p \ ;$$

the conditional maximum of the log-likelihood is

$$- \tfrac{1}{2} n \{\sum_i (\log \sigma_i^2 + s_{ii}\, \sigma_i^{-2}) + \log s_\sigma - s_\sigma + 1\} .$$

6.3 Barnett's data on human lung capacity

Barnett (1969) reports the results of a medical study of the relative merits of two instruments used for measuring human lung capacity, the first being a standard type and the second a newer design and more portable, easier to operate, and cheaper. Since it was claimed that the standard instrument required some skill to operate, the two instruments were used by both a skilled and an unskilled operator on the common group of 72 patients. Barnett assumes a linear structural model relating the

four instrument-operative combinations which is equivalent to the model of Section 6.1 with p equal to 4. He estimates the parameters of his model using some asymptotically unbiased moment estimators based on the exact ML solution in the case $p = 3$. The data are given in an Appendix, the order being : standard instrument/skilled operative, standard/unskilled, new/skilled, new/unskilled. We shall refer loosely to 'instruments' 1 to 4 for these instrument-operative combinations but use the terms 'standard' and 'new' instrument when necessary.

The sample mean vector is

$$\bar{\underline{y}} = (2246 \quad 2176 \quad 2149 \quad 2102)' ,$$

and the sample dispersion matrix (with divisor 72) is

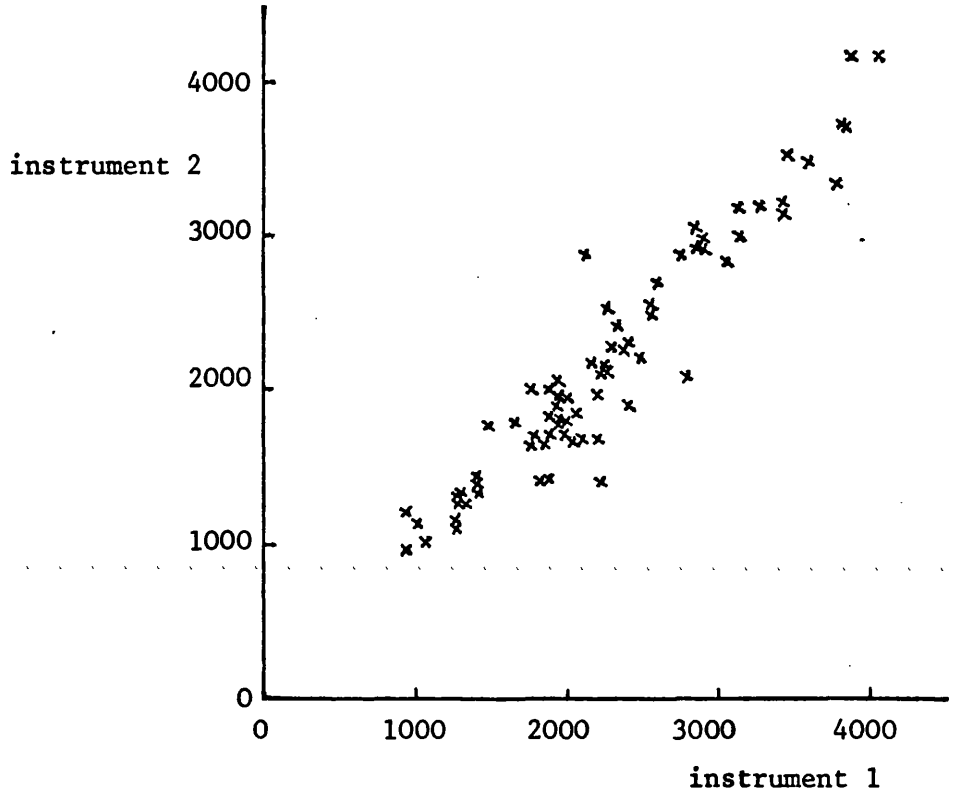
$$\underline{S} = \begin{pmatrix} 584290 & 573373 & 631330 & 596894 \\ & 618838 & 672753 & 637204 \\ & & 787937 & 726598 \\ & & & 721495 \end{pmatrix} .$$

The high degree of linearity in the bivariate relationships is shown by the correlation matrix

$$\underline{R} = \begin{pmatrix} 1 & .954 & .931 & .919 \\ & 1 & .963 & .954 \\ & & 1 & .964 \\ & & & 1 \end{pmatrix} . \qquad (6.7)$$

Figure 6.1 gives scatter diagrams for two pairs of instruments. They appear not to contradict the hypotheses of linearity and homoscedasticity, but do suggest that the distribution of the hypothetical 'standard measurement' f may be positively skewed.

The ML estimates are given in Table 6.2 ; the estimates of the calibration factors are in the ratios 1 : 1.06 : 1.19 : 1.13 . These ratios in particular, and most of the other estimates, are similar to those that would be obtained by transferring Barnett's estimates to our



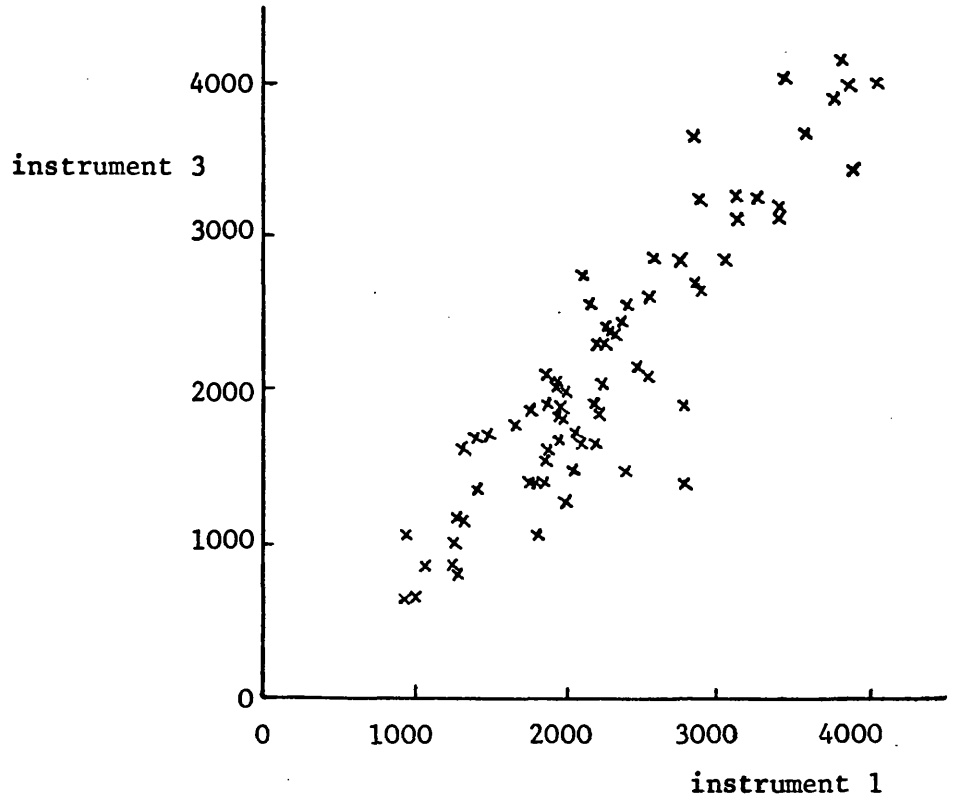


Figure 6.1. Scatter diagram for two pairs of instruments

Table 6.2 Maximum likelihood estimates for Barnett's data

Instrument i	Mean μ_i	Calibration factor λ_i	Standard deviation σ_i	Precision π_i
1	2246	730.8	224.2	3.260
2	2176	774.4	138.4	5.596
3	2149	871.0	171.0	5.094
4	2102	826.2	197.1	4.192

representation of the model. The major exception is that his estimate of σ_2 is 85.5, a strange result in that it relates to the *unskilled* operative using the standard instrument, the one which is supposed to require a certain skill in its operation. The anomaly is less marked with the ML estimates, but because of it the estimates of precision in Table 6.2 are of limited usefulness to the choice of instrument beyond showing that, at any rate, the new instrument is not much worse than the standard in this respect. Another slightly curious aspect of these results is that the ratios $\lambda_1 : \lambda_2$ and $\lambda_3 : \lambda_4$ are respectively greater than and less than 1. But we shall show below that the difference is not significant in the sense that the data are consistent with the hypothesis that $\lambda_1 = \lambda_2$ and $\lambda_3 = \lambda_4$.

The estimate of the asymptotic dispersion matrix of the $\hat{\mu}_i$ is, from (6.4),

$$\frac{1}{72}(\hat{\Sigma} + \hat{\lambda}\,\hat{\lambda}') = \begin{pmatrix} 8115 & 7860 & 8841 & 8386 \\ & 8595 & 9368 & 8886 \\ & & 10943 & 9995 \\ & & & 10021 \end{pmatrix},$$

so that the standard errors are respectively 90, 93, 105 and 100. The estimate for the remaining parameters in the order $\lambda_1, \ldots, \lambda_4, \sigma_1, \ldots, \sigma_4$ is found by inverting the matrix defined in (6.5) and is

$$\begin{pmatrix} 4417 & 3929 & 4420 & 4193 & -34 & 8 & 6 & 3 \\ & 4438 & 4681 & 4441 & 5 & -41 & 17 & 9 \\ & & 5685 & 4996 & 4 & 19 & -49 & 7 \\ & & & 5290 & 2 & 12 & 8 & -40 \\ & & & & 459 & -27 & -19 & -10 \\ & & & & & 361 & -95 & -50 \\ & & & & & & 451 & -34 \\ & & & & & & & 440 \end{pmatrix}.$$

One advantage of our choice of parameters is that the correlations between the estimates of the σ_i are fairly low and those between the $\hat{\sigma}_i$ and the $\hat{\lambda}_i$ are very low. The standard errors are respectively 66, 67, 75, 73, 21.4, 19.0, 21.2 and 21.0.

We next consider whether the data are consistent with certain restricted models in which equality constraints are imposed on some of the parameters. The first has $\lambda_1 = \lambda_2$ and $\lambda_3 = \lambda_4$, which implies that the 'true' values for the skilled and unskilled operators on each of the standard and new instruments differ by a simple additive bias. The second, and more interesting, hypothesis adds the further restrictions that $\mu_1 = \mu_2$ and $\mu_3 = \mu_4$ so that on each of the two instruments the true values for the operatives are the same. In other words we have two replicate observations on each of the two instruments, but the variances of the replicates are unequal. The ML estimators in these two cases are given in Tables 6.3 and 6.4 respectively.

Table 6.3 Maximum likelihood estimates with $\lambda_1 = \lambda_2$ and $\lambda_3 = \lambda_4$

Instrument i	Mean μ_i	Calibration factor λ_i	Standard deviation σ_i	Precision π_i
1	2246	762.3	225.2	3.384
2	2176	762.3	139.7	5.455
3	2149	851.2	174.5	4.877
4	2102	851.2	197.9	4.300

Table 6.4 Maximum likelihood estimates with $\lambda_1 = \lambda_2$, $\lambda_3 = \lambda_4$, $\mu_1 = \mu_2$ and $\mu_3 = \mu_4$

Instrument i	Mean μ_i	Calibration factor λ_i	Standard deviation σ_i	Precision π_i
1	2195	762.0	233.0	3.271
2	2195	762.0	143.2	5.320
3	2128	851.5	174.7	4.875
4	2128	851.5	198.9	4.281

In the first case they were calculated by maximizing the likelihood numerically as a function of λ_1, λ_3 and the σ_i ; in the second case μ_1 and μ_3 were also included. The maxima of the log-likelihoods are in the first case -2.004 and in the second -5.575 ; here and elsewhere the zero point of the log-likelihood is taken as its maximum in the basic, unconstrained model. To test the hypotheses expressed by these two sets of constraints we could take minus twice each of these maxima as being distributed approximately as χ^2 on 2 and 4 degrees of freedom under the respective null hypotheses. With this approximation, the first null hypothesis is accepted and the second is rejected at the .05 level but not at the .01 level. In an alternative 'likelihood' approach, preferred by the author, the reduction in the maximum of the log-likelihood resulting from restrictions on the parameters is compared directly with the number of constraints ; if we adopt Edwards' (1972, p.200) tentative suggestion of accepting a null hypothesis unless the reduction in the log-likelihood exceeds twice the number of constraints, then we shall be ready to accept the second of our null hypotheses and assume that the true measurements of the two operatives on each of the two instruments are the same. A comparison of Tables 6.2 and 6.4. shows that the anomaly of the unskilled operative taking the most precise measurements persists under this assumption, although it is slightly less marked.

The estimates of μ_1 and μ_3 in Table 6.4 have asymptotic dispersion matrix estimated by

$$\begin{pmatrix} 8270 & 9011 \\ & 10310 \end{pmatrix};$$

the remaining estimates, in the order $\hat{\lambda}_1, \hat{\lambda}_3, \hat{\sigma}_1, \ldots, \hat{\sigma}_4$, are asymptotically independent of the $\hat{\mu}_i$ with dispersion matrix estimated by

$$\begin{pmatrix} 4243 & 4503 & -7 & -27 & 13 & 9 \\ & 5279 & 4 & 15 & -23 & -16 \\ & & 503 & -26 & -21 & -14 \\ & & & 359 & -82 & -51 \\ & & & & 447 & -38 \\ & & & & & 471 \end{pmatrix}.$$

As before, there are low correlations between estimates apart from that between the $\hat{\lambda}_i$.

Next we test whether the data are consistent with any of the three sets of restrictions suggested at the end of Section 6.2. If we constrain the calibration factors λ_i to be equal then the maximum of the log-likelihood is -9.603 ; if also the means are equal, so that all four instruments have the same true measurements, the maximum is -17.295. Under the hypothesis that the four instruments are equally precise the maximum is -7.069. All three hypotheses are rejected in approximate χ^2 tests, the first two at the .001 level, the third at the .01 level ; an enthusiast for the likelihood approach might be prepared to accept the third hypothesis, particularly in view of the anomalous nature of the differences in precision. The estimates in this case are given in Table 6.5.

Table 6.5 Maximum likelihood estimates with common precision

Instrument i	Mean μ_i	Calibration factor λ_i	Standard deviation σ_i	Precision π_i
1	2246	750.9	176.9	4.244
2	2176	760.6	179.2	4.244
3	2149	860.8	202.8	4.244
4	2102	827.8	195.1	4.244

Finally, we return an observation made earlier in this Section that the scatter diagrams in Figure 6.1 suggest that the distribution of the true measurements appears to be positively skewed, and investigate whether a better fit to the data might be obtained by applying a common transformation to the four observed measurements. The transformation we consider assumes that the model of Section 6.1 is appropriate for some value of ψ to the variates $Y_i(\psi)$ defined by

$$Y_i(\psi) = Y_i^{\psi} \qquad (i = 1, \ldots, 4 \; ; \; 0 < \psi < 1).$$

In order to compare the maxima of the log-likelihoods for different values of ψ we require the Jacobian, J, of the transformation $y_{ij} \to y_{ij}(\psi)$ $(i = 1, \ldots, p \; ; \; j = 1, \ldots, n)$ (we give it in terms of p and n for the sake of generality). We have

$$J = \prod_{i,j} \left| dy_{ij}(\psi)/dy_{ij} \right| = \prod_{i,j} (\psi y_{ij}^{\psi-1}) = \psi^{np} \{ \prod_{i,j} y_{ij} \}^{\psi-1},$$

so for the purpose of comparison we have to add

$$\log J = np \log \psi + (\psi - 1) \sum_{i,j} \log y_{ij}$$

to the maximum of the log-likelihood obtained using the $y_{ij}(\psi)$. The maximized log-likelihoods for a series of values of ψ are given in Table 6.6. A plot of these maxima shows that they are roughly quadratic in ψ with a maximum near 0.8 ; a parabola fitted at the points 0.7, 0.8, 0.9 takes its maximum value of 1.747 at $\psi = .79$. The improvement in the fit

Table 6.6 Likelihood maxima for five power transformations

Transformation ψ	1	0.9	0.8	0.7	0.6	0.5
Likelihood maximum	0	1.251	1.740	1.462	0.415	-1.398

of the model is therefore not significant. Nevertheless, we give in Table 6.7 estimates of the parameters for the transformation $y^{0.8}$ since we shall use this transformation for illustrative purposes in the next Section. A comparison of the correlation matrix for the transformed data,

$$\underline{R}_{0.8} = \begin{pmatrix} 1 & .952 & .929 & .915 \\ & 1 & .962 & .953 \\ & & 1 & .961 \\ & & & 1 \end{pmatrix},$$

with that of the raw observations (equation 6.7) shows that each of the off-diagonal terms is reduced. Thus the transformation intended to make

Table 6.7 Maximum likelihood estimates for the transformation $y^{0.8}$

Instrument i	Mean μ_i	Calibration factor λ_i	Standard deviation σ_i	Precision π_i
1	475.5	124.2	38.5	3.226
2	463.1	131.8	23.2	5.681
3	456.8	149.3	30.8	4.843
4	449.2	142.1	35.2	4.039

the true measurements more nearly normal tends to produce non-linearity in the calibration relationships. This leads us to relax the hypothesis of normality for the true measurements while tentatively retaining the assumptions of linearity, homoscedasticity and normal errors. We can argue that our estimators are still consistent as follows: the ML estimators are functions of the data through $\bar{y}$ and S only ; if these statistics have the same limiting behaviour under some relaxed set of assumptions

then the estimators are still consistent. Sufficient conditions for $\bar{\underline{y}}$ and $\underline{S}$ to tend in probability to $\underline{\mu}$ and $\underline{\Sigma} + \underline{\lambda}\underline{\lambda}'$ in a linear calibration model with homoscedastic errors are (i) that the $\underline{Y}_j$ are independent or (ii) that the $\underline{Y}_j$ are uncorrelated and the quantities $Y_{ij}\ Y_{hj}$ $(i,h = 1, \ldots, p\ ;\ j = 1, \ldots, n)$ have finite variance independent of j and are uncorrelated between specimens. By the same argument, either of these conditions is sufficient for the ML estimators derived for the structural case to be consistent for the structural parameters $\underline{\xi}_o$, $\underline{\Sigma}$ and $\underline{\lambda}$ in the functional case, provided that the matrix $n^{-1} \Sigma \underline{\xi}_j \underline{\xi}_j'$ has a limit with rank one, which we write as $\underline{\lambda}\underline{\lambda}'$. We shall consider in the next Section techniques for detecting departures from the assumptions of the model.

The computer programs required for this Section were written and run by Dr. J.R. Mallinson. The method of function minimization used was that of Davidon, Fletcher and Powell (Fletcher and Powell, 1963).

6.4 Residual analyses and probability plots

In the belief that one should check against the data the assumptions made in a statistical analysis, we give in this Section some suggestions for residual analyses and probability plots for models of comparative calibration. The intention here is to provide graphical techniques to indicate failure of the assumptions, rather than tests for specific types of departure, although evidence of failure may lead us to formulate hypotheses which require to be tested, as in the case of the apparently skew distribution of true measurements in the previous Section. The techniques are illustrated on the data presented in that Section.

We begin with procedures for detecting departures from multivariate normality of the vector of observations $\underline{Y}$. Healy (1968) suggests a plotting procedure based on the fact that if a random p-vector $\underline{U}$ follows the distribution $N_p(\underline{\mu},\underline{D})$ its covariance form $(\underline{U} - \underline{\mu})' \underline{D}^{-1} (\underline{U} - \underline{\mu})$ has a χ^2 distribution on p degrees of freedom. The quadratic form

$$d_j = (\underline{y}_j - \bar{\underline{y}})' \underline{S}^{-1} (\underline{y}_j - \bar{\underline{y}}) \qquad (6.8)$$

therefore has approximately the same distribution, and plotting the ordered d_j against the quantiles of χ^2_p provides us with an overall indication of departure from multivariate normality and a method for detecting outliers before estimating the parameters of the model.

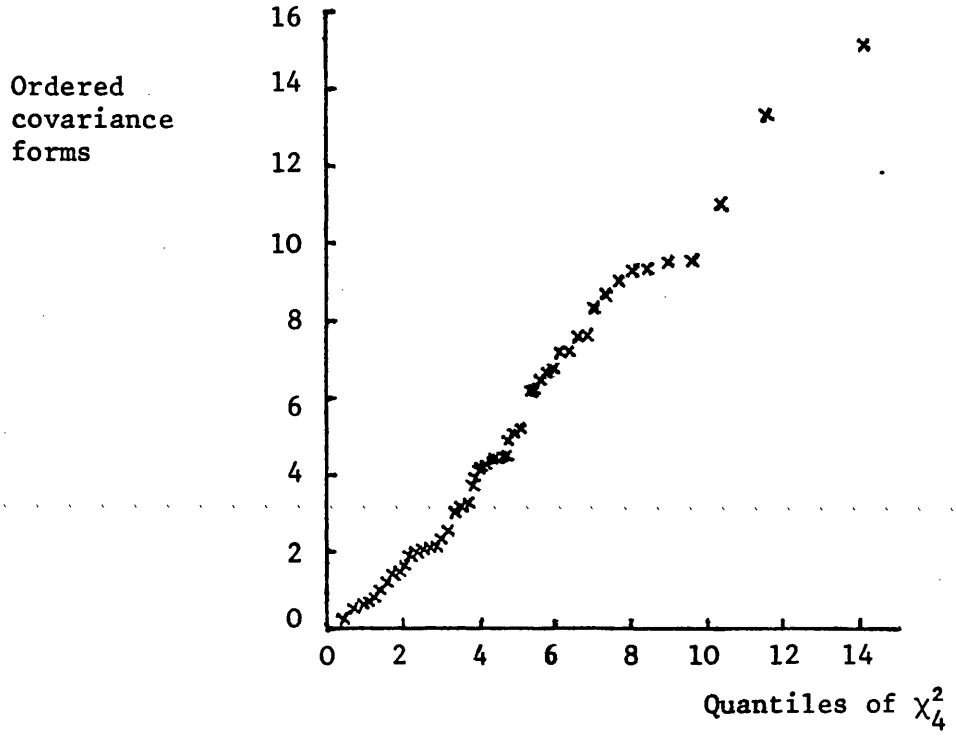
Figure 6.8 gives the plots for the raw data of the previous Section and for the transformation $y^{0.8}$ (alternate points only are shown up to the 40th). The first does give some indication of non-normality, though it is difficult to say in what direction ; the second suggests that the transformation does give some improvement in the fit of the distribution.

Having estimated the parameters of the model, we may replace $\underline{S}$ in (6.8) by the estimated dispersion matrix $\hat{\underline{\Sigma}} + \hat{\underline{\lambda}}\hat{\underline{\lambda}}'$; the resulting plot should give a better indication of departures from the model, since it assumes not a general dispersion matrix for $\underline{Y}$ but one which is the sum of a diagonal and a rank-one matrix.

The plots for the raw and transformed data are given in Figure 6.9. The curvature in these plots is more pronounced than in Figure 6.8; the kink at $d = 2$ suggests that there are too many observations close to the mean. But generally we conjecture that this type of plot is more useful for detecting that a roughly symmetric distribution for the true measurements has long tails than the skewness we suspect here.

Other univariate views of the data are obtainable from the measurements y_i on individual instruments and from linear combinations of the y_i. If the π_i are reasonably large, say greater than 3, then the marginal distributions should be very similar, since all but a proportion $(1 + \pi_i^2)^{-1}$ of the variance of y_i is accounted for by variation in the true measurement. It therefore seems worthwhile to investigate the distribution of the true measurement more directly by estimating the values corresponding to the n specimens. Since they are not parameters in our model we need some ad hoc method of estimation; the obvious one

(a) raw data



(b) transformed data : $y^{0.8}$

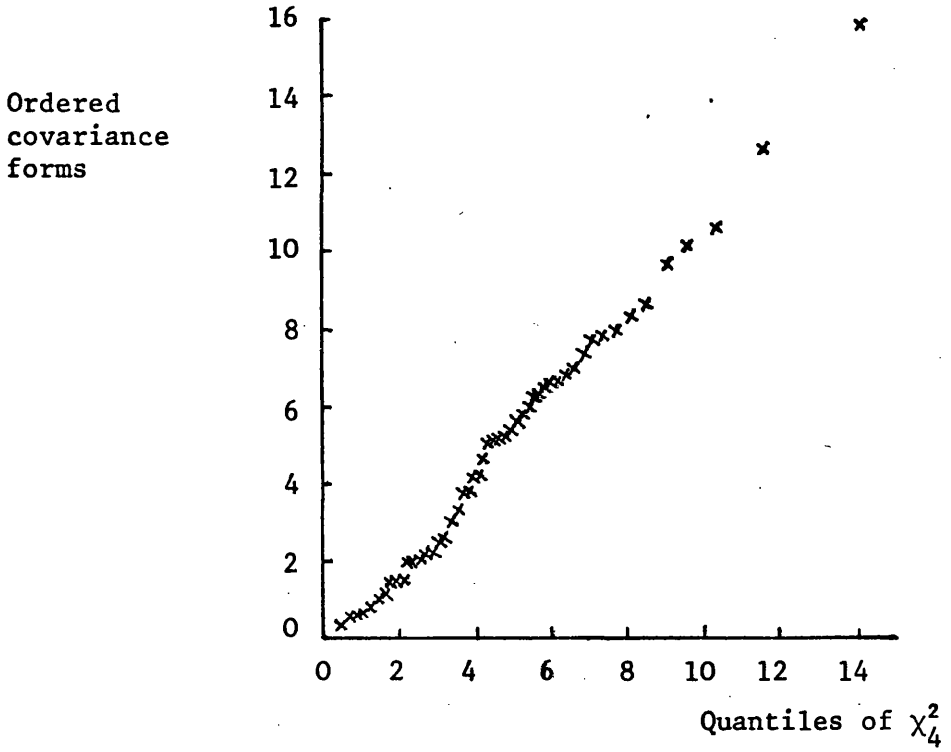
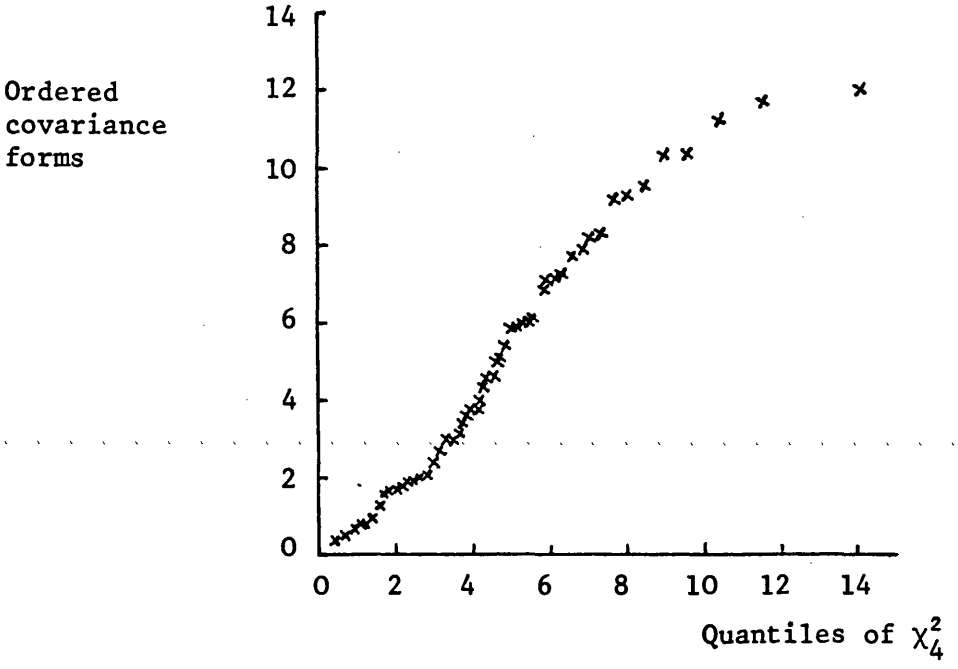


Figure 6.8. Chi-square probability plots of the covariance form d_j using sample dispersion matrix

(a) raw data



(b) transformed data : $y^{0.8}$

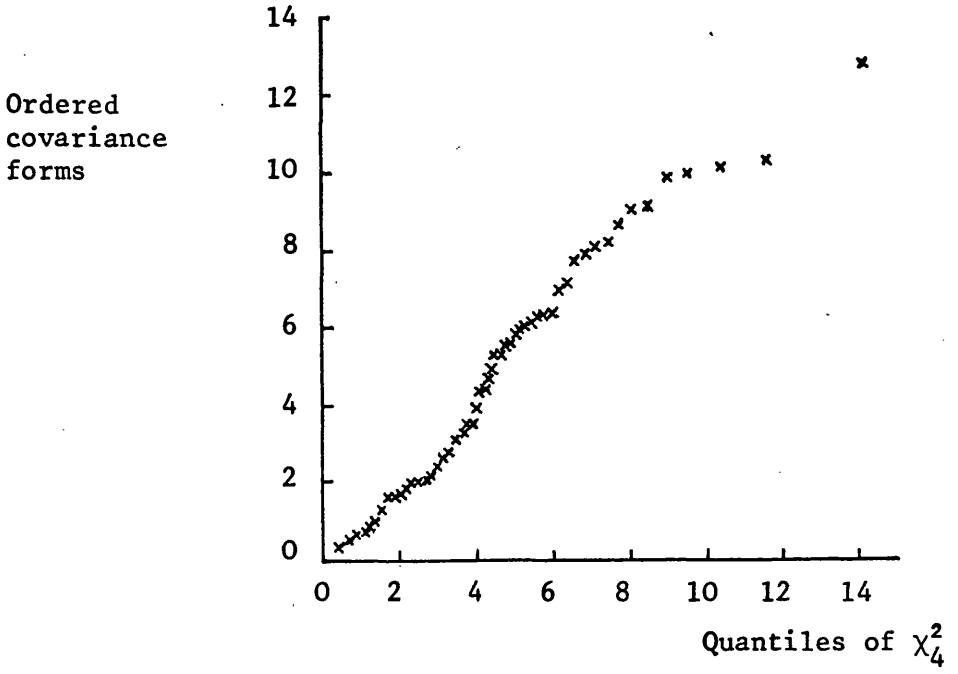


Figure 6.9. Chi-square probability plots of the covariance form d_j using estimated dispersion matrix

seems to be the following:

divide each y_i by the standard error of the corresponding error term. If we write $\underline{X}_\sigma$ for $(\sigma_1^{-1} X_1, \ldots, \sigma_p^{-1} X_p)'$ and define $\underline{\mu}_\sigma$ and $\underline{y}_\sigma$ similarly then our model (6.1) becomes

$$\underline{X}_\sigma = \underline{\mu}_\sigma + \underline{\pi}\, f$$

with a spherically symmetric distribution for the errors. So estimate the vector of standardized true measurements $\underline{X}_{\sigma j}$ corresponding to the vector of observations $\underline{y}_j$ by the orthogonal projection of the standardized vector $\underline{y}_{\sigma j}$ onto the line $\underline{\xi} = \underline{\mu}_\sigma + \underline{\pi}\phi$ (ϕ real). This gives

$$\tilde{\underline{X}}_{\sigma j} = \underline{\mu}_\sigma + \underline{\pi}\, \tilde{f}_j \quad \text{or} \quad \tilde{\underline{X}}_j = \underline{\mu} + \underline{\lambda}\, \tilde{f}_j ,$$

where the tilde denotes an ad hoc estimate and the fitted value $\tilde{f}_j$ of the standard measurement is given by

$$\tilde{f}_j = \underline{\pi}'(\underline{y}_{\sigma j} - \underline{\mu}_\sigma)/\underline{\pi}'\underline{\pi} \quad (j = 1, \ldots, n). \qquad (6.9)$$

Essentially the same problem arises in factor analysis under the heading of the 'estimation of factor scores'; Lawley and Maxwell (1971, ch.8) give two methods. The above procedure corresponds to that of Bartlett (1937, 1938) whose principle of estimation is to minimize with respect to f_j the sum of squares of the standardized 'residuals', that is $(\underline{y}_{\sigma j} - \underline{\mu}_\sigma - \underline{\pi}\, f_j)'(\underline{y}_{\sigma j} - \underline{\mu}_\sigma - \underline{\pi}\, f_j)$. Another method, due to Thomson (1951), is to choose the linear combination $\tilde{f}_j$ of the $y_{ij} - \mu_i$ which minimizes the variance of $\tilde{f}_j - f_j$; this is the same as (6.9) apart from having $1 + \underline{\pi}'\underline{\pi}$ as its denominator, a change which makes no difference when examining for departures from normality in f and very little otherwise if the π_i are at all large. The μ_i, π_i and σ_i in (6.9) have to be estimated, so we shall use as our fitted value

$$g_j = \hat{\underline{\pi}}'\hat{\underline{\Sigma}}^{-\frac{1}{2}}(\underline{y}_j - \bar{\underline{y}})/\hat{\underline{\pi}}'\hat{\underline{\pi}} \quad j = 1, \ldots, n).$$

If the assumptions of our model are correct we should expect the g_j to have roughly the properties of the $\tilde{f}_j$, that is to be distributed

independently as $N(0,\ 1 + (\underline{\pi}'\underline{\pi})^{-1})$. So we may use a normal probability plot of the g_j to examine for departures from the assumption of normality for the true measurements.

For the data of the previous Section, normal probability plots on the four measurements show similar patterns indicating positive skewness. Plots using various power transformations show that better fits are obtained by taking cube roots of each measurement; to take logarithms would be an over-transformation. Our estimates of the π_i indicate that more than 90% of the variance of each measurement is accounted for by the true measurement, so the four plots of the marginal distributions are similar to that of the fitted values of the standard measurement g_j given in Figure 6.10. The evidence of skewness is only slightly reduced in the plot for the best-fitting transformation $y^{0.8}$.

Now that we have a vector of standardized fitted values $\underset{\sim}{\hat{X}}_{\sigma j}$ we can define a vector of standardized residuals as

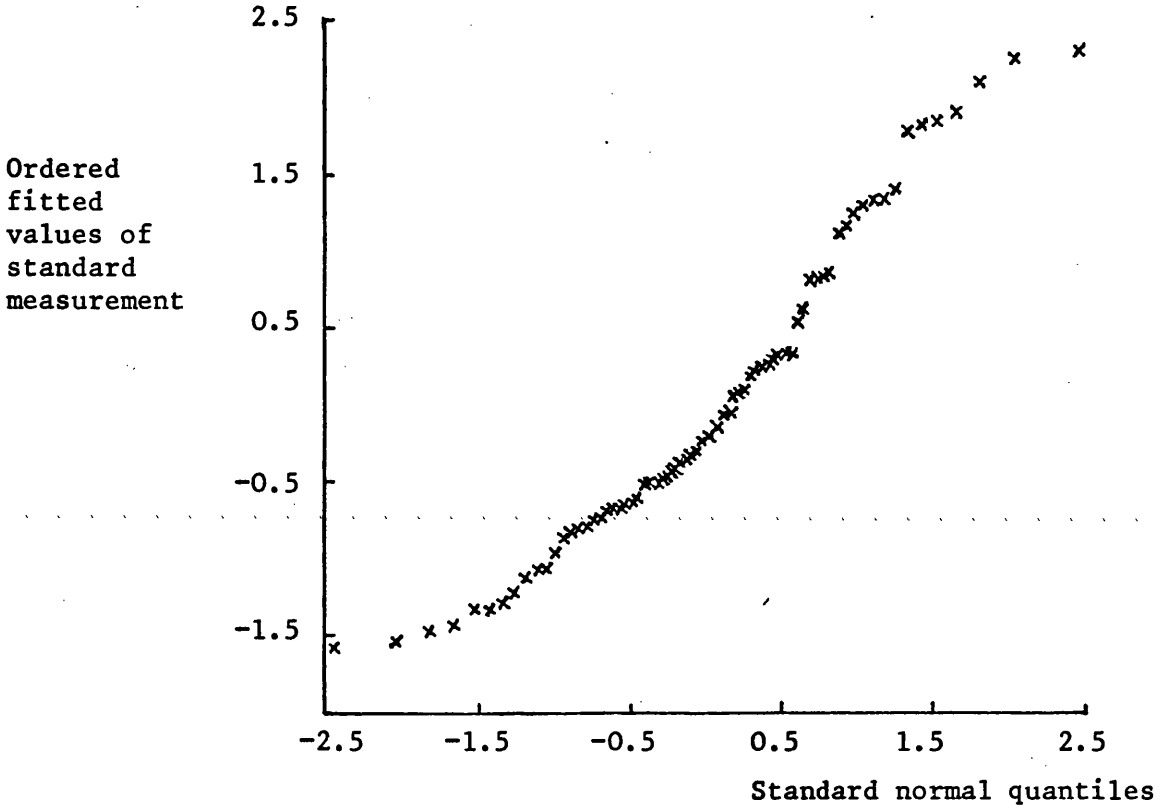
$$\underline{r}_j = \hat{\underline{\Sigma}}^{-\frac{1}{2}}(\underline{y}_j - \bar{\underline{y}}) - \hat{\underline{\pi}}\, g_j = \{\underline{I} - (\hat{\underline{\pi}}'\hat{\underline{\pi}})^{-1}\hat{\underline{\pi}}\,\hat{\underline{\pi}}'\}\,\hat{\underline{\Sigma}}^{-\frac{1}{2}}(\underline{y}_j - \bar{\underline{y}}). \qquad (6.10)$$

Under the assumptions of our model the $\underline{r}_j$ should be roughly independent with the degenerate distribution $N_p(\underline{0}, \underline{I} - (\underline{\pi}'\underline{\pi})^{-1}\,\underline{\pi}\,\underline{\pi}')$. If the true measurements are not normal this should have little effect on the distribution of the $\underline{r}_j$, since, in the limit as n tends to infinity, (6.10) becomes

$$\{\underline{I} - (\underline{\pi}'\underline{\pi})^{-1}\,\underline{\pi}\,\underline{\pi}'\}(\underline{\pi}\, f_j + \underline{\Sigma}^{-\frac{1}{2}}\,\underline{\varepsilon}_j)\ ,$$

from which f_j disappears. So a normal probability plot of the i^{th} elements of the $\underline{r}_j$ should give some indication of departure from normality of the i^{th} error term. Also, plotting the i^{th} elements of the $\underline{r}_j$ against g_j can indicate heteroscedasticity and non-linearity in the dependence of the y_i on f. Of course scatter diagrams of the y_i in pairs provide evidence of departures from linearity and homoscedasticity, but their number increases as the square of p. If we suspect the same

(a) raw data



(b) transformed data : $y^{0.8}$

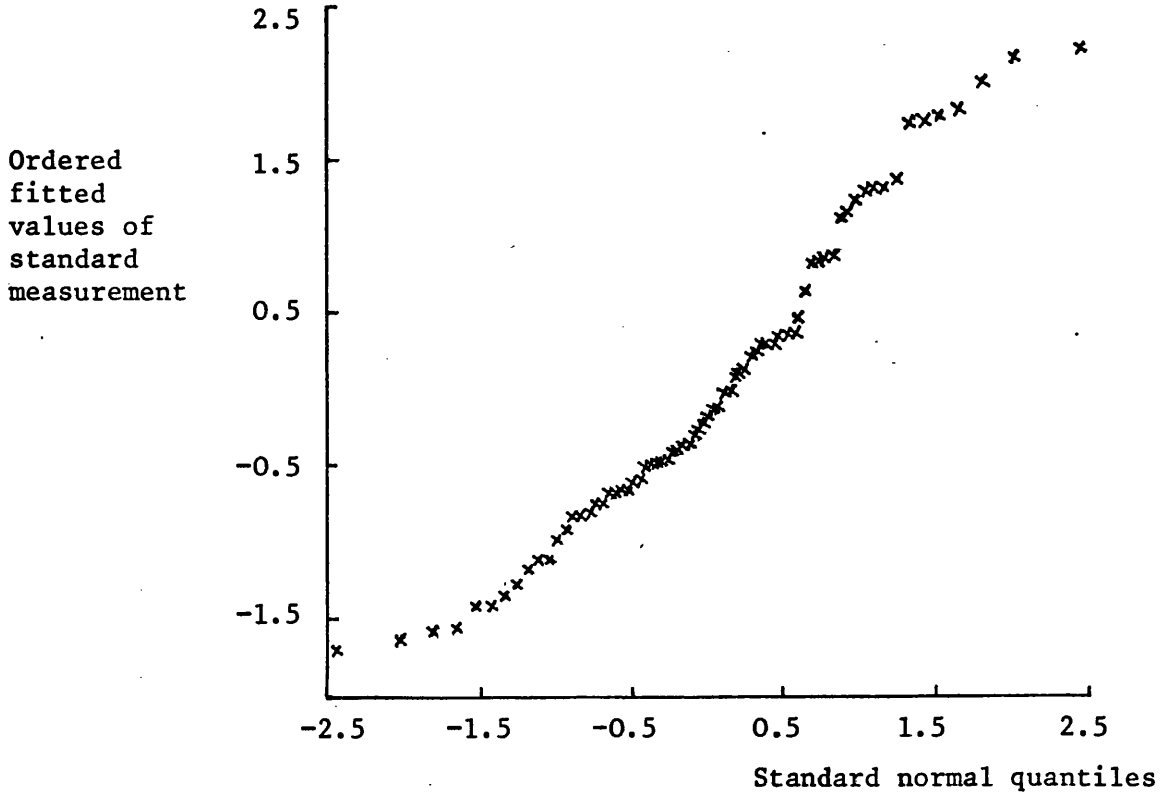


Figure 6.10. Normal probability plot of fitted values of standard measurement

type of departure from homoscedasticity in each of the instruments, for example error variance increasing with true measurement, we may combine the elements of $\underline{r}_j$ by calculating the length e_j of the residual vector defined by

$$e_j^2 = \underline{r}_j'\underline{r}_j = (\underline{y}_j - \bar{\underline{y}})'\hat{\underline{\Sigma}}^{-1}(\underline{y}_j - \bar{\underline{y}}) - (\hat{\underline{\pi}}'\hat{\underline{\pi}})^{-1}\{\hat{\underline{\pi}}'\hat{\underline{\Sigma}}^{-\frac{1}{2}}(\underline{y}_j - \bar{\underline{y}})\}^2 ,$$

and plot e_j against g_j. Another possible use of the residual vector is in detecting outliers where they have not already been thrown up in the chi-square plots of the covariance form d_j.

Normal probability plots of the four elements of the residual vector $\underline{r}_j$ are given in Figure 6.11. There is no consistent pattern of departure from normality and only the fourth plot is markedly non-linear. In search of an explanation for this departure from linearity, we plot the residuals r_{4j} against the fitted values of the standard measurement (Figure 6.12) but it indicates nothing as simple as, say, increasing or decreasing variance or non-linear dependence of y_4 on f. A plot of the length of the residual vector against the fitted measurement (Figure 6.13) also gives no indication of changing variance.

We conclude that there is little evidence against the hypotheses of linearity and homoscedasticity, but that there is a strong suggestion of non-normality and positive skewness in the distribution of the standard measurement f which is not removed when a power transformation is applied to the data. There is also some evidence of non-normality in the error variates. In principle we should perhaps attempt to fit some other distribution to the standard measurement; this would presumably require numerical maximization of the likelihood with respect to all 12 parameters of the model. But we prefer to take comfort from the facts that the consistency of our estimators does not depend on the normality of the true measurements, and that the calibration equations do not depend on the distribution of f. We have to admit, though, that our tests of the fit of the various restricted models are not valid.

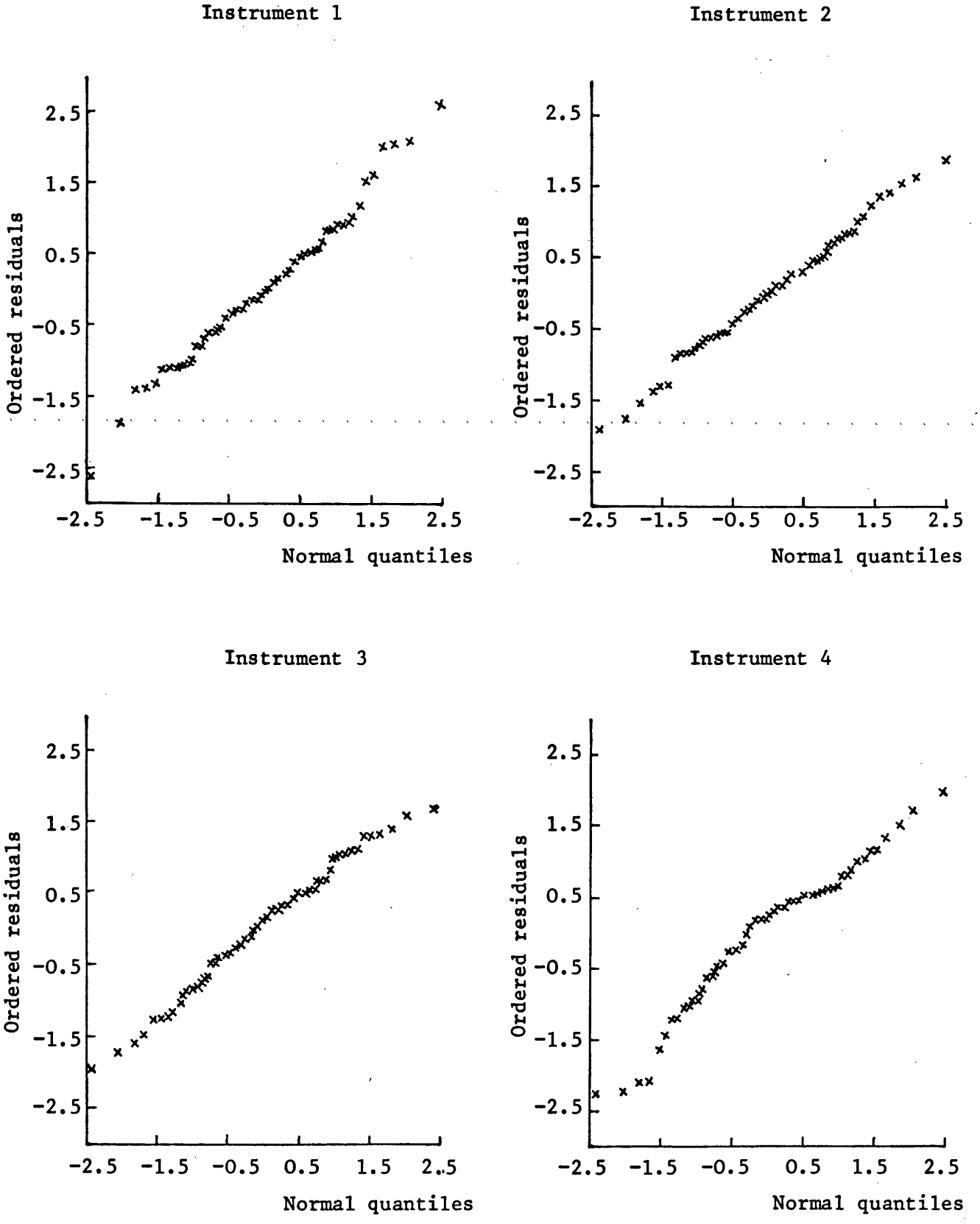


Figure 6.11. Normal probability plots of the residuals r_{ij}

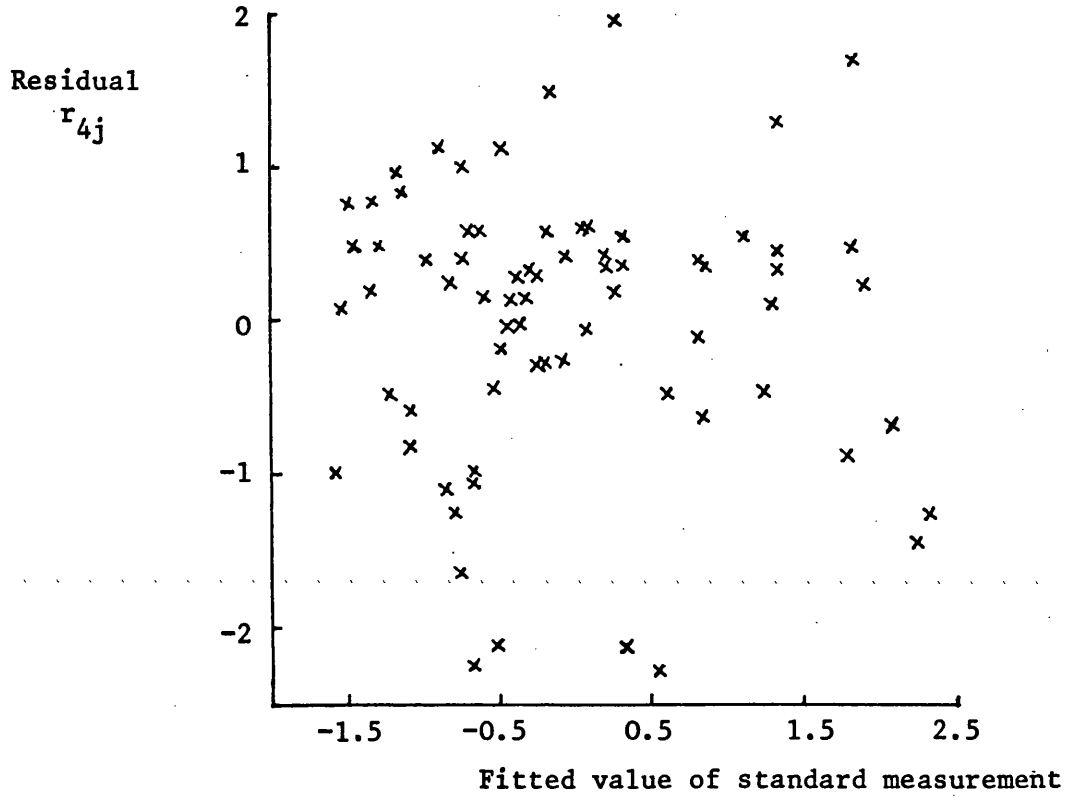


Figure 6.12. Plot of residual r_{4j} against fitted value g_j of standard measurement

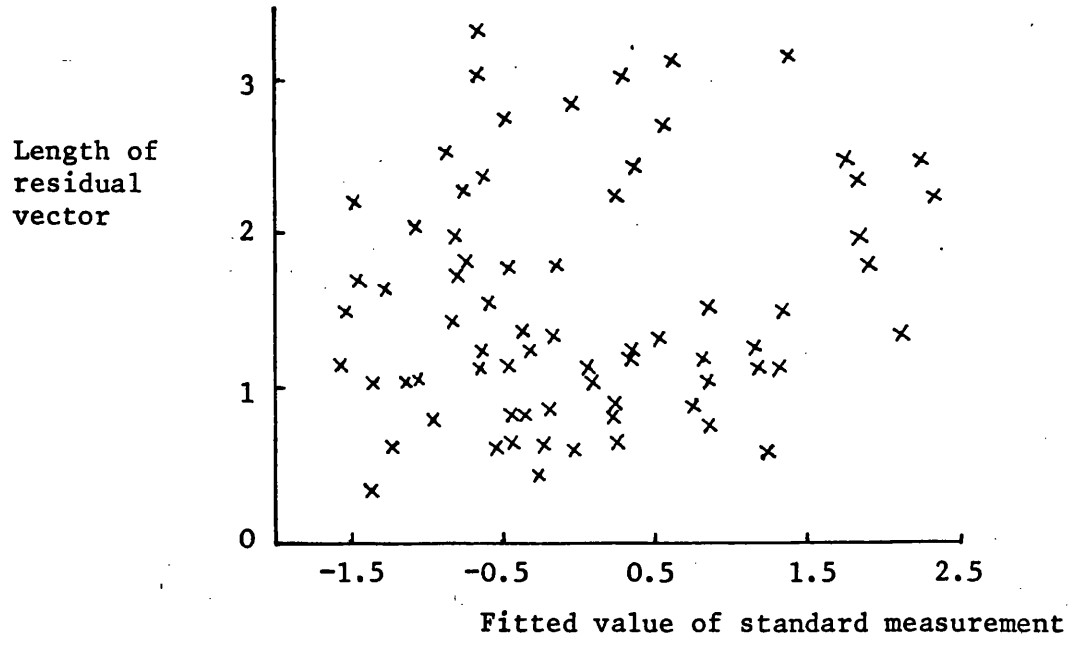


Figure 6.13. Plot of length e_j of residual vector against fitted value g_j of standard measurement

7. CONCLUSIONS

We have been concerned in this dissertation mainly with the representation and estimation of multiple functional and structural relationships between an arbitrary number p of random variables; the examples in Section 2.3 demonstrate that it is necessary to consider more than two variates and more than one relationship.

On the question of representing the relationships, there are clear advantages in methods which are symmetrical in the p variates rather than the regression-like methods usually employed for two variates. The symmetry emphasises the distinction between functional and structural relationships and regression relationships and also simplifies mathematical derivations. We have had to adopt a thoroughgoing matrix approach to the subject and this has shown the importance to our study of the idea of the simultaneous reduction to diagonal form of a pair of symmetric matrices. This has provided the clue to obtaining ML estimators for general p and r in cases for which no such solution had been found for two variates, for example in a structural-relationship model where the departures have a general unknown positive-definite dispersion matrix and there are equal numbers of replicates in each group.

It is of interest to examine the extent to which the well-known problems of estimating one relationship between two variates recur in higher dimensions. To begin with unreplicated cases, we noted in Sections 1.3 and 1.4 that knowledge of the first and second moments is not sufficient for estimating a relationship between two variates. The parameters of $p-r$ structural relationships are identifiable for sufficiently small r if the departure terms corresponding to the p elements of $\underline{Y}$ are independent and $p > 2$; such models are equivalent to factor models and include the important case (considered in Chapter 6) of true values lying on a line in p dimensions.

Our analysis of unreplicated functional relationships is inadequate; we have shown only that if we ignore the inevitable grouping of the data then the unboundedness of the likelihood noted by Solari in the case $p = 2$ occurs for all values of p and r. This is enough to show why Lawley's Method II of factor analysis did not work and why Williams' suggested solution for the functional-relationship version of the problem of comparative calibration with unknown variances could not work. But a proper study of the problem would take account of the grouping of the data and thus extend the results of Copas (1972) to relationships between more than two variates. For values of r for which the corresponding structural model is identifiable, we may extend the argument at the end of Section 6.3 to a set of functional relationships with uncorrelated departures and assert that the parameters of the relationships and the variances are estimated consistently by the estimators appropriate to the corresponding structural model, provided that either of the conditions on the y_{ij} given there applies and that the 'sample dispersion' matrix of the $\underline{\xi}_j$ has a finite limit of rank r.

Another difficulty with unreplicated functional relationships between two variates is that when the dispersion matrix of the departures is known up to a constant factor the proportional constant is not estimated consistently by ML. This lack of consistency recurs in higher dimensions, but may be removed by multiplying the estimator by the ratio of the number of variates to the number of relationships.

When there are replicate observations, both functional and structural relationships are estimated consistently by ML and GLS, but in the functional case the presence of incidental parameters still leads to ML estimators of $\underline{\Sigma}$ which are not consistent if the number of replicates in each group is bounded. A consistent estimator may always be found; for example, if $\underline{\Sigma}$ is a general positive-definite matrix we may use the

within-groups CSSP matrix divided by its degrees of freedom, which is also the GLS estimator. Since the corresponding estimators of the relationships coincide we may view the GLS procedure in this case as a necessary correction to the ML solution. For structural models we cannot expect to obtain consistent estimators of the mean and dispersion matrix of the vector of true variates if the number of groups is bounded, although the relationships and the dispersion matrix of the departures are consistently estimated by ML.

The striking features of the ML and GLS estimators obtained in Chapters 4 and 5 are their close similarity and their connexion with canonical analysis: when $\underline{\Sigma}$ is known to within a constant factor the $p-r$ linear relationships are estimated by equating to zero the last $p-r$ canonical variates in a canonical analysis involving $\underline{\Sigma}$ and the between-groups CSSP matrix; when $\underline{\Sigma}$ is a general positive-definite matrix which has to be estimated from replication the relationships are estimated using a similar analysis in which $\underline{\Sigma}$ is replaced by the within-groups CSSP matrix. Thus we equate to zero the linear functions of $\underline{X}$ or $\underline{\xi}$ corresponding to the $p-r$ linear combinations of the y_i which have the smallest ratios of between-group variation to within-group variance subject to their within-group covariances being zero.

The main point to be made on the design of experiments for estimating functional and structural relationships is the need for replication, since, apart from those structural models which are equivalent to factor models, we require for estimation either replicate observations or some further information such as the values taken by an instrumental variate or the knowledge that the dispersion matrix is proportional to a given matrix. If replication is possible then it is best in the structural case to have equal numbers in each group, as it appears to be much more difficult to maximize the likelihood if the numbers are unequal; this is not an important consideration for functional relationships.

In Chapter 6 we have applied the ML procedure for estimating unreplicated structural relationships to a problem which concerns the comparative calibration of measuring instruments and involves $p-1$ structural relationships between p variates. As well as testing certain hypotheses about the values of parameters in the model, we have been able to develop and illustrate procedures for testing the assumptions of linearity, homoscedasticity and normality in the model. The methods as proposed apply only when the true measurements are confined to a line in p dimensions, and their generalizations to other cases are not immediate. The suggested overall tests of normality are not affected, but the other procedures involve the fitted true value g_j which would in general become an r-vector. The plotting procedures for testing the normality of the true variates and the homoscedasticity of the departures would inevitably become more complicated.

APPENDIX

Barnett's (1969) data, giving readings of a measure of human lung capacity for 72 patients on 4 instrument-operative combinations : these data are re-analysed in Section 6.3.

y_1	y_2	y_3	y_4
3450	3530	4030	3720
1310	1320	1610	1600
3820	3720	4150	3700
2110	2880	2740	2520
1860	1420	1540	1690
1940	1780	2020	1800
2360	2260	2430	2350
2880	2920	2650	2860
1980	1720	1800	1660
3120	3180	3250	3040
1760	1630	1390	1200
1480	1760	1700	1640
1840	1660	1400	1650
3580	3480	3680	3960
1880	2000	2090	2070
2400	2320	2550	2480
2220	2120	2290	2270
2540	2500	2620	1960
920	1200	640	1030
2240	2160	2300	2300
2240	2130	2030	2140
2260	2510	2400	2450
3860	4180	3980	3680
2780	2100	1890	2000
2220	1400	1840	1360
1880	1820	1900	1840
940	960	1060	1000
2480	2220	2150	2150
1660	1780	1760	1800
4040	4180	4000	3770
2540	2560	2080	2250
1780	1700	1390	1200
1280	1300	800	1130
1940	2060	2030	1880
1760	2000	1860	1860
2040	1660	1470	1160

y_1	y_2	y_3	y_4
1060	1000	850	600
2000	1800	1270	1700
2280	2280	2380	2350
1940	1800	1670	1580
2580	2700	2850	2110
1400	1440	1680	1480
1260	1100	1000	1030
2320	2420	2360	2360
2000	1940	1980	1980
2400	1900	1470	1740
2880	2980	3240	3140
3420	3150	3200	3200
1000	1130	650	840
1400	1400	1350	1380
1880	1710	1600	1350
1280	1260	1160	1330
3120	3000	3110	3250
3770	3340	3900	3700
3420	3220	3120	3290
2740	2880	2850	2880
2840	2920	2710	2750
3800	3740	3440	3400
2100	1680	1650	1930
1820	1400	1060	1050
1400	1320	1350	1100
2200	1680	1640	1110
1940	1900	1820	1270
3260	3200	3250	3270
1960	1940	1890	1920
1320	1260	1140	1000
2840	3060	3650	3510
2060	1840	1720	1780
2200	1970	1900	2270
1260	1150	860	1150
3040	2840	2850	2670
2140	2180	2560	2720

REFERENCES

ANDERSON, T.W. and RUBIN, H. (1956). Statistical inference in factor analysis. *Proc. Third Berkeley Symp. math. Statist. Probab.*, 5, 111-150.

BARNETT, V.D. (1966). Discussion on the paper by Sprent. *J.R. Statist. Soc.* B, 28, 291-292.

——— (1969). Simultaneous pairwise linear structural relationships. *Biometrics*, 25, 129-142.

BARTLETT, M.S. (1937). The statistical conception of mental factors. *Br. J. Psychol.*, 28, 97-104.

——— (1938). Methods of estimating mental factors. *Nature, Lond.*, 141, 609-610.

——— (1947). Multivariate analysis. *J.R. Statist. Soc., Suppl.*, 9, 176-197.

——— (1948). A note on the statistical estimation of supply and demand relations from time series. *Econometrica*, 16, 323-329.

——— (1949). Fitting a straight line when both variables are subject to error. *Biometrics*, 5, 207-212.

——— (1950). Discussion on the papers by Kendall and Babington Smith. *J.R. Statist. Soc.* B, 12, 85-87.

——— (1966). Discussion on the paper by Sprent. *J.R. Statist. Soc.* B, 28, 292-293.

BELLMAN, R. (1970). *Introduction to Matrix Analysis.* New York : McGraw-Hill.

COPAS, J.B. (1972). The likelihood surface in the linear functional relationship problem. *J.R. Statist. Soc.* B, 34, 274-278.

DOLBY, G.R. (1972). Generalized least-squares and maximum likelihood estimation of non-linear functional relationships. *J.R. Statist. Soc.* B, 34, 393-400.

DWYER, P.S. (1967). Some applications of matrix derivatives in multivariate analysis. *J. Amer. Statist. Ass.*, 62, 607-625.

EDWARDS, A.W.F. (1972). *Likelihood.* Cambridge University Press.

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

FLETCHER, R. and POWELL, M.J.D. (1963). A rapidly convergent descent method for minimization. *Comput. J.*, 6, 163-168.

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

——— (1943). Relations between statistics : the general and the sampling problem when the samples are large. *Proc. Roy. Irish Acad.* A, 49, 177-196.

——— (1949). Determination of linear relations between systematic parts of variables with errors of observation the variances of which are unknown. *Econometrica*, 17, 30-59.

GLESER, L.J. and WATSON, G.S. (1973). Estimation of a linear transformation. *Biometrika*, 60, 525-534.

HARDY, G.H., LITTLEWOOD, J.E. and POLYA, G. (1952). *Inequalities* (2nd ed.). Cambridge University Press.

HEALY, M.J.R. (1968). Multivariate normal plotting. *Appl. Statist.*, 17, 157-161.

HOPKINS, J.W. (1966). Some considerations in multivariate allometry. *Biometrics*, 22, 747-760.

HUXLEY, J.S. (1924). Constant differential growth ratios and their significance. *Nature, Lond.*, 114, 895-896.

JÖRESKOG, K.G. (1966). UMLFA - A computer program for unrestricted maximum likelihood factor analysis. Research Memorandum 66-20. Princeton, New Jersey : Educational Testing Service.

——— (1970). A general method for analysis of covariance structures. *Biometrika*, 57, 239-251.

——— (1973). Analysis of covariance structures. In *Multivariate Analysis III* (ed. P.R. Krishnaiah), pp. 263-285. New York : Academic Press.

JÖRESKOG, K.G., van THILLO, M. and GRUVAEUS, G.T. (1971). ACOVSM - A general computer program for analysis of covariance structures including generalized MANOVA. Research Bulletin 71-01. Princeton, New Jersey : Educational Testing Service.

KEMPTHORNE, O. (1966). Some aspects of experimental inference. *J. Amer. Statist. Ass.*, 61, 11-34.

KENDALL, M.G. (1950). Factor analysis as a statistical technique. *J.R. Statist. Soc.* B, 12, 60-73.

——— (1961). *A Course in the Geometry of n Dimensions*. London : Griffin.

KENDALL, M.G. and STUART, A. (1967). *The Advanced Theory of Statistics*, Vol. 2 (2nd ed.). London : Griffin.

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

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

LAWLEY, D.N. and MAXWELL, A.E. (1971). *Factor Analysis as a Statistical Method* (2nd ed.). London : Butterworths.

——— (1973). Regression and factor analysis. *Biometrika*, 60, 331-338.

LINDLEY, D.V. (1947). Regression lines and the linear functional relationship. *J.R. Statist. Soc., Suppl.*, 9, 218-244.

LOMAX, K.S. (1948). Analysis of demand and supply in textiles. *Manchester School*, 16, 46.

LORD, F.M. and NOVICK, M.R. (1968). *Statistical Theories of Mental Test Scores*. Reading, Mass. : Addison-Wesley.

MADANSKY, A. (1959). The fitting of straight lines when both variables are subject to error. *J. Amer. Statist. Ass.*, 54, 173-205.

——— (1964). Instrumental variables in factor analysis. *Psychometrika*, 29, 105-113.

MARCUS, M. (1956). An eigenvalue inequality for the product of normal matrices. *Amer. Math. Monthly*, 63, 173-174.

MIRSKY, L. (1959). On the trace of matrix products. *Math. Nachr.*, 20, 171-174.

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

NEYMAN, J. and SCOTT, E.L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16, 1-32.

RAO, C.R. (1973). *Linear Statistical Inference and its Applications*. New York : Wiley.

REIERSØL, O. (1950). Identifiability of a linear relation between variables which are subject to error. *Econometrica*, 18, 375-389.

RICHTER, H. (1958). Zur Abschätzung von Matrizennormen. *Math. Nachr.*, 18, 178-187.

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

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

——— (1968). Linear relationships in growth and size studies. *Biometrics*, 24, 639-656.

——— (1969). *Models in Regression and Related Topics*. London : Methuen.

——— (1972). The mathematics of size and shape. *Biometrics*, 28, 23-37.

TAYLOR, J. (1973). A method of fitting several linear functional relations and of testing for differences between them. *Appl. Statist.*, 22, 239-248.

TEISSIER, G. (1955). Sur la détermination de l'axe d'un nuage rectiligne de points. *Biometrics*, 11, 344-357.

THEOBALD, C.M. (1975a). An inequality for the trace of the product of two symmetric matrices. *Math. Proc. Camb. Phil. Soc.*, 77, 265-267.

——— (1975b). An inequality with application to multivariate analysis. *Biometrika*, 62, 461-466.

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

——— (1946). Multiple regression for systems of equations. *Econometrica*, 14, 5-36.

——— (1952). *Econometrics*. New York : Wiley.

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

WALD, A. (1940). The fitting of straight lines if both variables are subject to error. *Ann. Math. Statist.*, 11, 284-300.

WILLIAMS, E.J. (1955). Significance tests for discriminant functions and linear functional relationships. *Biometrika*, 42, 360-381.

——— (1959). *Regression Analysis*. New York : Wiley.

——— (1969). Regression methods in calibration problems. *Bull. Int. Statist. Inst.*, 43, 17-28.

——— (1973). Tests of correlation in multivariate analysis. *Bull. Int. Statist. Inst.*, 45, 219-232.

PUBLISHED PAPERS

1. An inequality for the trace of the product of two symmetric matrices.
   Mathematical Proceedings of the Cambridge Philosophical Society (1975), 77, 265-267.

2. An inequality with application to multivariate analysis.
   Biometrika (1975), 62, 461-466.

An inequality for the trace of the product of two symmetric matrices

By C. M. THEOBALD

Math. Proc. Camb. Phil. Soc. (1975), **77**, 265
MPCPS 77-22
Printed in Great Britain

An inequality for the trace of the product of two symmetric matrices

By C. M. THEOBALD
University of Bath

(*Received* 1 *June* 1974)

Let A, B be real symmetric $n \times n$ matrices having orthogonal reductions

$$P^{\top}AP = \Lambda_A, \quad Q^{\top}BQ = \Lambda_B,$$

where $\Lambda_A = \operatorname{diag}(\lambda_1(A), \ldots, \lambda_n(A))$ and $\Lambda_B = \operatorname{diag}(\lambda_1(B), \ldots, \lambda_n(B))$, and the sets of latent roots $\{\lambda_i(A)\}_{i=1}^n$, $\{\lambda_i(B)\}_{i=1}^n$ are each in descending order. (Thus we may say that the latent vectors of A, B are *ordered* with respect to $\{\lambda_i(A)\}_{i=1}^n$, $\{\lambda_i(B)\}_{i=1}^n$ respectively.) Let the multiplicities of the roots $\{\lambda_i(A)\}_{i=1}^n$ and $\{\lambda_i(B)\}_{i=1}^n$ in order of occurrence be respectively $\{m_g(A)\}_{g=1}^a$ and $\{m_h(B)\}_{h=1}^b$, so that

$$\sum_{g=1}^{a} m_g(A) = \sum_{h=1}^{b} m_h(B) = n.$$

THEOREM. *With the above notation,*

$$\operatorname{tr} AB \leqslant \operatorname{tr} \Lambda_A \Lambda_B$$

or, equivalently,

$$\sum_{i=1}^{n} \lambda_i(AB) \leqslant \sum_{i=1}^{n} \lambda_i(A)\,\lambda_i(B),$$

with equality if and only if

$$P^{\top}Q = G^{\top}H,$$

where G, H take the form

$$G = \begin{pmatrix} G_1 & & & \\ & G_2 & & \\ & & \ddots & \\ & & & G_a \end{pmatrix}, \quad H = \begin{pmatrix} H_1 & & & \\ & H_2 & & \\ & & \ddots & \\ & & & H_b \end{pmatrix} \tag{1}$$

and $\{G_g\}_{g=1}^a$, $\{H_h\}_{h=1}^b$ are orthogonal matrices of orders $\{m_g(A)\}_{g=1}^a$, $\{m_h(B)\}_{h=1}^b$ respectively.

An equivalent condition is that there exists a common set of latent vectors for A and B which are ordered with respect to both $\{\lambda_i(A)\}_{i=1}^n$ and $\{\lambda_i(B)\}_{i=1}^n$.

Proof. Let $R = P^{\top}Q$. Then R is orthogonal and

$$\operatorname{tr} AB = \operatorname{tr} \Lambda_A R \Lambda_B R^{\top} = \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i(A)\, r_{ij}^2 \lambda_j(B),$$

where $R = (r_{ij})$. The matrix whose ijth element is r_{ij}^2 is doubly stochastic (that is, each element is non-negative and each row and column sum is unity), so it is sufficient to show that, for any doubly stochastic matrix $S = (s_{ij})$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i(A)\, s_{ij} \lambda_j(B) \leqslant \sum_{i=1}^{n} \lambda_i(A)\,\lambda_i(B). \tag{2}$$

For any pair (s_{ue}, s_{ev}) in which $u, v > e$ and both elements are positive, let

$$t_{uv}^{(e)} = \min(s_{ue}, s_{ev})$$

and consider the operation of adding $t_{uv}^{(e)}$ to s_{ee} and s_{uv} and subtracting the same amount from s_{ue} and s_{ev}. This reduces to zero at least one off-diagonal element, preserves the doubly stochastic character of S, and increases the left-hand side of (2) by the non-negative amount

$$t_{uv}^{(e)}\{\lambda_e(A) - \lambda_u(A)\}\{\lambda_e(B) - \lambda_v(B)\}.$$

If we clear the off-diagonal elements of the first row and column, then the second, and so on, it requires at most $(n-1)^2$ such operations to reduce S to the identity matrix. (There are $n(n-1)$ off-diagonal elements, but the operation on the final non-zero element in each row also clears the corresponding column.)

Equality is attained when

$$\operatorname{tr} \Lambda_A R \Lambda_B R^\top = \operatorname{tr} \Lambda_A \Lambda_B. \tag{3}$$

It follows from equation (2) that the only products of roots which may appear on the left-hand side of (3) are those of the pairs $\{(\lambda_i(A), \lambda_i(B))\}_{i=1}^n$. Thus R represents the combined effect of a rotation G of the columns of P which permutes the $\{\lambda_i(A)\}_{i=1}^n$ while preserving their order and a similar rotation H of the columns of Q. So $R = G^\top H$, where G and H satisfy

$$G\Lambda_A G^\top = \Lambda_A, \quad H\Lambda_B H^\top = \Lambda_B.$$

The general forms for orthogonal matrices satisfying these conditions are given by (1).

Let N be an orthogonal matrix whose columns form a common set of latent vectors for A and B and are ordered with respect to the roots of both A and B. Then

$$\operatorname{tr} AB = \operatorname{tr} N\Lambda_A N^\top N \Lambda_B N^\top = \operatorname{tr} \Lambda_A \Lambda_B.$$

Conversely, if there exist G, H satisfying (1), then the columns of $PG^\top$ form the required set of latent vectors of A and B. This completes the proof.

If the roots of A are distinct then the condition for equality is that the (unique) ordered set of latent vectors for A also forms an ordered set of latent vectors for B. If B also has distinct roots the condition is that the ordered sets of vectors are identical.

We obtain theorem 368 of Hardy *et al.* ((1), p. 261) if, in (2), S has the form of a permutation matrix.

A special case of this inequality with n equal to 3 has been given by Hill (2) as a property of the inner product of two symmetric tensors. The method of proof, he indicates, involves showing that the stationary points of $\operatorname{tr} AB$ occur where the latent vectors coincide, that is, where

$$\operatorname{tr} AB = \sum_{i=1}^{3} \lambda_i(A)\lambda_{\pi(i)}(B)$$

and $\pi(i)$ is a permutation of the numbers 1 to 3. He then inspects the six possibilities.

One application of the above result is to be found in the theory of multivariate statistical analysis: Maximum likelihood estimation of the dispersion matrix of a p-variate normal distribution involves the minimization with respect to the positive definite matrix Σ of the function

$$\log|\Sigma| + \operatorname{tr} S\Sigma^{-1},$$

where S is a given positive definite matrix. The standard method of carrying out this minimization is to differentiate with respect to each element of Σ ((3), p. 47). The theorem shows that the minimum value of the function, for fixed $\{\lambda_i(\Sigma)\}_{i=1}^p$, is

$$\sum_{i=1}^{p} \{\log \lambda_i(\Sigma) + \lambda_i(S)/\lambda_i(\Sigma)\},$$

and this occurs when S and Σ have a common set of ordered latent vectors. The ith term in this sum takes its smallest value when $\lambda_i(\Sigma) = \lambda_i(S)$, so the minimum is attained at $\Sigma = S$.

Note added in proof 21 *October* 1974. Since this paper was accepted for publication it has been found that the above inequality is but a special case of a similar result for hermitian matrices proved by Richter(4) and Mirsky(5). It may also be derived from an inequality of Marcus(6), using theorem 368 of Hardy *et al.* ((1), p. 261). However, no condition for equality is given in any of these references. Also the proof given above is somewhat simpler than those appearing previously and may be generalised to apply to hermitian matrices with only minor alterations, for example, writing 'unitary' for 'orthogonal'.

REFERENCES

(1) Hardy, G. H., Littlewood, J. E. and Pólya, G. *Inequalities*, 2nd edition (Cambridge University Press, 1952).
(2) Hill, R. On constitutive inequalities for simple materials, 1. *J. Mech. Phys. Solids*, **16** (1968), 229–242.
(3) Anderson, T. W. *An Introduction to Multivariate Statistical Analysis* (Wiley: New York, 1958).
(4) Richter, H. Zur Abschätzung von Matrizennormen. *Math. Nachr.*, **18** (1958), 178–187.
(5) Mirsky, L. On the trace of matrix products. *Math. Nachr.*, **20** (1959), 171–174.
(6) Marcus, M. An eigenvalue inequality for the product of normal matrices. *Amer. Math. Monthly*, **63** (1956), 173–174.

Biometrika (1975), 62, 2, *p.* 461
Printed in Great Britain

An inequality with application to multivariate analysis

BY C. M. THEOBALD*
School of Mathematics, University of Bath, Avon

SUMMARY

An inequality for the trace of the product of two symmetric matrices can be used to simplify the proofs of a number of results in multivariate analysis. This is illustrated on two problems: estimating the parameters of a multivariate normal population, and estimating a system of linear functional relationships between the mean vectors of k multivariate normal populations with the same dispersion matrix. A new result is proved for the corresponding structural relationship problem.

Some key words: Inequality; Linear functional relationship; Linear structural relationship; Matrix trace; Maximum likelihood estimation; Multivariate normal distribution.

1. INTRODUCTION

There are several instances in multivariate normal theory in which the likelihood of a set of observations involves the trace of the product of two symmetric matrices. Richter (1958) has proved an inequality for such matrix traces and Theobald (1975) has given an alternative proof and a necessary and sufficient condition for equality; these results are stated here as Theorem 1. It is shown in the present paper that other inequalities may be deduced which simplify the proofs of some previously known results on maximum likelihood estimation and facilitate the derivation of new results.

For the basic inequality we require the following notation: let C and D be real symmetric $m \times m$ matrices with latent roots $\lambda_i(C)$ and $\lambda_i(D)$ and corresponding latent vectors p_i and q_i, respectively ($i = 1, \ldots, m$). We may assume that the sequences $\{\lambda_i(C)\}$, $\{\lambda_i(D)\}$ are non-increasing, in which case we may say that the latent vectors of C and D are ordered with respect to the latent roots. In what follows we shall use the same notation for the latent roots of any square matrix. In matrix terms we have $C = P\Lambda_C P'$ and $D = Q\Lambda_D Q'$, where $\Lambda_C = \operatorname{diag}\{\lambda_1(C), \ldots, \lambda_m(C)\}$, $P = (p_1, \ldots, p_m)$, etc. Let the multiplicities of the roots $\lambda_i(C)$ and $\lambda_i(D)$ in order of occurrence be, respectively, $m_g(C)$ ($g = 1, \ldots, c$) and $m_h(D)$ ($h = 1, \ldots, d$), so that

$$\sum_{g=1}^{c} m_g(C) = \sum_{h=1}^{d} m_h(D) = m.$$

The following inequality of Richter (1958) has been given a necessary and sufficient condition for equality by Theobald (1975).

THEOREM 1. *With the above notation,*

$$\operatorname{tr}(CD) \leqslant \operatorname{tr}(\Lambda_C \Lambda_D)$$

* Now at Department of Statistics, University of Warwick, Coventry.

or, equivalently,

$$\sum_{i=1}^{m} \lambda_i(CD) \leqslant \sum_{i=1}^{m} \lambda_i(C)\lambda_i(D),$$

with equality if and only if $P'Q = G'H$, *where* G *and* H *take the form*

$$G = \operatorname{diag}(G_1, \ldots, G_c), \quad H = \operatorname{diag}(H_1, \ldots, H_d), \tag{1}$$

with G_g *and* H_h *being orthogonal matrices of orders* $m_g(C)$ *and* $m_h(D)$, *respectively* ($g = 1, \ldots, c$; $h = 1, \ldots, d$). *An equivalent condition is that there exists a common set of latent vectors for* C *and* D *which are ordered with respect to both* $\lambda_i(C)$ *and* $\lambda_i(D)$.

COROLLARY. *If* D *is nonsingular then*

$$\operatorname{tr}(CD^{-1}) \geqslant \operatorname{tr}(\Lambda_C \Lambda_D^{-1}),$$

with the same condition for equality as before.

The proof is immediate if we replace D by $-D^{-1}$.

The next result involves the simultaneous reduction of symmetric matrices. Let D be positive-definite and let U be a symmetric $m \times m$ matrix. Let the roots in descending order of the equations

$$|C - \phi D| = 0, \quad |U - \phi D| = 0 \tag{2}$$

be, respectively,

$$\phi_i(C), \quad \phi_i(U) \quad (i = 1, \ldots, m),$$

and let $\Phi_C = \operatorname{diag}\{\phi_1(C), \ldots, \phi_m(C)\}$, etc. It follows (Rao, 1973, p. 41) that there exist non-singular $m \times m$ matrices N_C and N_U such that

$$D = N_C N_C', \quad C = N_C \Phi_C N_C', \tag{3}$$

$$D = N_U N_U', \quad U = N_U \Phi_U N_U'. \tag{4}$$

Any column of N_C and N_U which corresponds to a simple root of (2) is unique apart from reversal of sign. It is easily verified that $N_U^{-1} N_C$ is orthogonal.

THEOREM 2. *With the above notation*

$$\operatorname{tr}\{C(D+U)^{-1}\} \geqslant \operatorname{tr}\{\Phi_C(I_m + \Phi_U)^{-1}\},$$

with equality if and only if (3) *and* (4) *have a common solution.*

Proof. We have

$$\begin{aligned} \operatorname{tr}\{C(D+U)^{-1}\} &= \operatorname{tr}[\{N_U^{-1} N_C \Phi_C N_C' (N_U')^{-1}\}(I_m + \Phi_U)^{-1}] \\ &\geqslant \operatorname{tr}\{\Phi_C(I_m + \Phi_U)^{-1}\}, \end{aligned}$$

since, if J is any $m \times m$ orthogonal matrix, $J\Phi_C J'$ has the same latent roots as Φ_C. In terms of Theorem 1, $P = N_U^{-1} N_C$, $Q = I$, and the condition for equality is that

$$N_U^{-1} N_C = H_U' G_C, \tag{5}$$

where G_C and H_U have the form of (1), the orders of the submatrices being given by the multiplicities of the roots of (2). Equation (5) may be rewritten

$$N_C G_C' = N_U H_U'. \tag{6}$$

But the orthogonal transformations $N_C \to N_C G_C'$ and $N_U \to N_U H_U'$ are precisely those which leave (3) and (4) unaffected, since the forms of G_C and H_U are such as to satisfy $G_C' \Phi_C G_C = \Phi_C$ and $H_U' \Phi_U H_U = \Phi_U$. Thus (6) amounts to the statement that (3) and (4) have a common solution.

In §§ 2 and 3 we give examples in which the above Theorems are useful for simplifying the derivation of known results in maximum likelihood estimation; in § 4 a new result is obtained.

2. Multivariate normal population

Suppose that n independent observations are taken on a p-variate normal random variable with mean vector μ and dispersion matrix Σ, both unknown. If we ignore a term in μ which reduces to zero at the sample mean, maximization of the likelihood amounts to minimization with respect to Σ of

$$n \log |\Sigma| + \operatorname{tr}(S\Sigma^{-1}), \tag{7}$$

where S is the matrix of corrected sums of squares and products, assumed to be nonsingular.

We may now use the Corollary to Theorem 1 to show that the minimum value of (7) for fixed $\lambda_l(\Sigma)$ $(l = 1, \ldots, p)$ is

$$\sum_{l=1}^{p} \left\{ n \log \lambda_l(\Sigma) + \frac{\lambda_l(S)}{\lambda_l(\Sigma)} \right\}, \tag{8}$$

and this occurs when S and Σ have a common set of ordered latent vectors. The lth term in (8) takes its smallest value when $\lambda_l(\Sigma) = \lambda_l(S)/n$, so that the minimum is attained at $\Sigma = S/n$.

Another algebraic proof of the same result is given by Watson (1964), who applies to $S\Sigma^{-1}$ the result that for any square matrix A, $|A| \leqslant \exp\{\operatorname{tr}(A - I)\}$.

3. Linear functional relationships

Let there be p-variate normal populations $N_p(\mu_1, \Sigma), \ldots, N_p(\mu_k, \Sigma)$ with a common nonsingular dispersion matrix. Rao (1973, pp. 556–60) considers the problem of examining whether the mean vectors are confined to an r-dimensional hyperplane $(r < k-1)$, that is, whether they may be expressed as

$$\mu_s = \mu_0 + E\alpha_s \quad (s = 1, \ldots, k),$$

where μ_0 is an unknown fixed vector in the hyperplane, E is an unknown fixed $p \times r$ matrix of full rank and α_s is an unknown r-vector. Equivalently, if F is a $p \times (p-r)$ matrix of full rank satisfying $F'E = 0$, then the components of μ_s are subject to $(p-r)$ functional relationships expressible as

$$F'(\mu_s - \mu_0) = 0 \quad (s = 1, \ldots, k). \tag{9}$$

Rao calls (9) a structural relationship, but functional relationship corresponds to the more usual terminology of Kendall & Stuart (1967, pp. 375–8).

Suppose that n_s observations are taken from $N_p(\mu_s, \Sigma)$ and that $\bar{x}_s$ is the sample mean $(s = 1, \ldots, k)$. Then, for fixed Σ, the likelihood is $e^{-\frac{1}{2}t}$, where

$$t = \sum_{s=1}^{k} n_s (\bar{x}_s - \mu_s)' \Sigma^{-1} (\bar{x}_s - \mu_s). \tag{10}$$

Let $\bar{x} = \Sigma n_s \bar{x}_s / \Sigma n_s$ be the overall mean,

$$B = \sum_{s=1}^{k} n_s (\bar{x}_s - \bar{x})(\bar{x}_s - \bar{x})'$$

the matrix of between groups corrected sums of squares and products, and ϕ_l $(l = 1, \ldots, p)$ the roots of $|B - \phi\Sigma| = 0$ in descending order. If $k > p$, then with probability one B is non-singular and the roots are distinct. In deriving the likelihood ratio test for the hypothesis H_0 that the $\{\mu_s\}$ lie on an r-dimensional hyperplane, Rao shows that

$$\min_{H_0} \sum_{s=1}^{k} n_s (\bar{x}_s - \mu_s)' \Sigma^{-1} (\bar{x}_s - \mu_s) = \phi_{r+1} + \ldots + \phi_p. \tag{11}$$

He proceeds by transforming $\bar{x}_s$ to $\Sigma^{-\frac{1}{2}} \bar{x}_s$, although it is not clear whether the symbol $\Sigma^{\frac{1}{2}}$ is intended to denote the symmetric square root of Σ or merely a matrix L satisfying $\Sigma = LL'$. We can simplify the proof of (11), and also derive the maximum likelihood estimator of μ_s, if we define this transformation more precisely and then use Theorem 1, as follows.

There exists a nonsingular $p \times p$ matrix M such that

$$\Sigma = MM', \quad B = M\Phi M',$$

where $\Phi = \operatorname{diag}(\phi_1, \ldots, \phi_p)$. If the ϕ_l are distinct then M is unique, apart from multiplication of any subset of its columns by (-1). Let $z_s = M^{-1}\bar{x}_s$ and $\xi_s = M^{-1}\mu_s$, so that the between-groups matrix for the transformed observations is $M^{-1}B(M')^{-1} = \Phi$, and (10) becomes

$$\sum_{s=1}^{k} n_s (z_s - \xi_s)'(z_s - \xi_s). \tag{12}$$

The hypothesis H_0 applies equally to $\{\xi_s\}$ and amounts to the assumption that there exists a point ξ_0 in the hyperplane, a $p \times r$ matrix R with orthonormal columns, and a set of r-vectors c_s $(s = 1, \ldots, k)$ such that

$$\xi_s = \xi_0 + Rc_s \quad (s = 1, \ldots, k). \tag{13}$$

The columns of R are chosen to form an orthonormal basis for the r-dimensional subspace containing $\xi_s - \xi_0$ $(s = 1, \ldots, k)$. Apart from the factor n_s, the sth term in (12) is

$$(z_s - \xi_0 - Rc_s)'(z_s - \xi_0 - Rc_s). \tag{14}$$

Minimization with respect to c_s amounts to regressing $z_s - \xi_0$ on R. The minimum of (14) is the residual sum of squares which, for fixed ξ_0 and R, is

$$(z_s - \xi_0)'(I_p - RR')(z_s - \xi_0), \tag{15}$$

since $R'R = I_r$, and this is attained when

$$c_s = R'(z_s - \xi_0). \tag{16}$$

The expression to be minimized with respect to ξ_0, R is from (12) and (15),

$$\operatorname{tr}\left\{ \sum_{s=1}^{k} n_s (z_s - \xi_0)(z_s - \xi_0)'(I_p - RR') \right\} = \operatorname{tr}\left[\left\{ \Phi + \left(\sum_{s=1}^{k} n_s \right) (\bar{z} - \xi_0)(\bar{z} - \xi_0)' \right\} (I_p - RR') \right].$$

Since $I_p - RR'$ is positive semi-definite, this is a minimum with respect to ξ_0 when $\xi_0 = \bar{z}$. We now use Theorem 1 to minimize $\operatorname{tr}\{\Phi(I_p - RR')\}$, or to maximize $\operatorname{tr}[\Phi\{-(I_p - RR')\}]$, with

respect to R. The roots of $I_p - RR'$ are zero, with multiplicity r, and unity, with multiplicity $(p-r)$, so that the minimum under H_0 of (10) is $\phi_{r+1} + \ldots + \phi_p$, as required. The columns of R form a set of latent vectors corresponding to the zero roots of $-(I_p - RR')$, so the condition for equality is that $R' = (H_1', 0)$, where H_1 is $r \times r$ and orthogonal. Substituting in (16) and (13), we have

$$\xi_s = \bar{z} + \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} (z_s - \bar{z}). \tag{17}$$

In terms of the original variables (17) becomes

$$\hat{\mu}_s = \bar{x} + M_1 (M^{-1})^1 (x_s - \bar{x}), \tag{18}$$

where M_1 consists of the first r columns of M and $(M^{-1})^1$ of the first r rows of M^{-1}. The matrix M is given by $M = \Sigma K$, where the columns of K are solutions of

$$(B - \phi_l \Sigma) k_l = 0,$$

scaled so that $k_l' \Sigma k_l = 1$ $(l = 1, \ldots, p)$. It is not in fact necessary to carry out this scaling in calculating (18). As Rao points out, Σ may be estimated from the matrix of within-groups corrected sums of squares and products, although this is not the maximum likelihood procedure.

4. Linear structural relationships

As in § 3, we assume that there are k, p-variate normal populations with a common nonsingular dispersion matrix Σ. In this case, however, the mean vectors, instead of being fixed but unknown, are independent realizations of a degenerate random p-vector X distributed as $N_p(\mu, \Theta)$, where μ and Θ are unknown. The rank of Θ is taken to be $r < p$, which implies that there are $(p-r)$ linear relationships between the components of X, which may be expressed as $\Gamma'(X - \mu) = 0$, where Γ is any $p \times (p-r)$ matrix of full rank satisfying $\Gamma' \Theta = 0$. We shall consider the problem of estimating μ and Θ given Σ and a single observation x_s on each population.

The k observations are independently distributed as $N_p(\mu, \Sigma + \Theta)$, so that the log likelihood is

$$-\tfrac{1}{2}[k \log |\Sigma + \Theta| + \operatorname{tr}\{T(\Sigma + \Theta)^{-1}\} + k(\bar{x} - \mu)' (\Sigma + \Theta)^{-1} (\bar{x} - \mu)],$$

where T is the matrix of corrected sums of squares and products. As in § 2 the maximum with respect to μ occurs at $\mu = \bar{x}$. We have to minimize with respect to Θ

$$k \log |\Sigma + \Theta| + \operatorname{tr}\{T(\Sigma + \Theta)^{-1}\}. \tag{19}$$

In accordance with the notation introduced before Theorem 2, let $\phi_l(T)$ and $\phi_l(\Theta)$ $(l = 1, \ldots, p)$ be the roots in descending order of

$$|T - \phi \Sigma| = 0, \quad |\Theta - \phi \Sigma| = 0,$$

respectively, $\Phi_T = \operatorname{diag}\{\phi_1(T), \ldots, \phi_p(T)\}$, etc., and N_T and N_Θ be matrices satisfying

$$\Sigma = N_T N_T', \quad T = N_T \Phi_T N_T', \quad \Sigma = N_\Theta N_\Theta', \quad \Theta = N_\Theta \Phi_\Theta N_\Theta'. \tag{20}$$

Since Θ has rank r, $\phi_{r+1}(\Theta) = \ldots = \phi_p(\Theta) = 0$ and (19) becomes

$$k \log |\Sigma| + k \sum_{l=1}^{r} \log\{1 + \phi_l(\Theta)\} + \operatorname{tr}\{T(\Sigma + \Theta)^{-1}\}. \tag{21}$$

Theorem 2 shows that, for fixed Φ_Θ, (21) is minimized by

$$k\log|\Sigma| + k\sum_{l=1}^{r}\log\{1+\phi_l(\Theta)\} + \operatorname{tr}\{\Phi_T(I+\Phi_\Theta)^{-1}\}$$
$$= k\log|\Sigma| + \sum_{l=1}^{r}\left[k\log\{1+\phi_l(\Theta)\} + \frac{\phi_l(T)}{1+\phi_l(\Theta)}\right] + \sum_{l=r+1}^{p}\phi_l(T), \quad (22)$$

the condition for equality being that (20) have a common solution. The lth term in (22) takes its smallest value when $\phi_l(\Theta) = \{\phi_l(T)/k\} - 1$, and this is the maximum likelihood estimator provided that $\phi_r(T) > k$. This condition must hold in order for the estimators of the $\phi_l(\Theta)$ $(l = 1, \ldots, r)$ all to be positive, since the estimator of Θ must be nonnegative-definite. If it does not hold for a particular set of data we may wish to consider a smaller value for r.

If $k > p$ then with probability one the $\phi_l(T)$ are distinct, N_T is unique and the condition for equality is that the first r columns of N_Θ equal the corresponding columns of N_T. Thus if $\phi_r(T) > k$ and $k > p$ we have, following the notation of (18)

$$\hat{\Theta} = (N_T)_1(\Phi_T/k - I_p)(N_T')^1.$$

As in § 3, it is possible to estimate Σ if we have more than one observation on each population. Observations on the same population are correlated, but, if the numbers of replications are equal, the method of estimating $\hat{\Theta}$ is a straightforward generalization of that given above.

REFERENCES

KENDALL, M. G. & STUART, A. (1967). *The Advanced Theory of Statistics*, Vol. 2, 2nd edition. London: Griffin.

RAO, C. R. (1973). *Linear Statistical Inference and its Applications*, 2nd edition. New York: Wiley.

RICHTER, H. (1958). Zur Abschätzung von Matrizennormen. *Math. Nachr.* **18**, 178–87.

THEOBALD, C. M. (1975). An inequality for the trace of the product of two symmetric matrices. *Math. Proc. Camb. Phil. Soc.* **77**, 265–7.

WATSON, G. S. (1964). A note on maximum likelihood. *Sankhyā* A **26**, 303–4.

[*Received July* 1974. *Revised December* 1974]