STATISTICAL MODELS WITH COVARIANCE CONSTRAINTS

by

Paul Arthur Tukey

Thesis submitted for the
Diploma of Membership of Imperial College
and the degree of
Doctor of Philosophy in the University of London

January 1976

STATISTICAL MODELS WITH COVARIANCE CONSTRAINTS

PhD Thesis, November 1975

by

Paul A. Tukey

Department of Mathematics
Imperial College of Science and Technology, London

and

Bell Telephone Laboratories
Murray Hill, New Jersey

ABSTRACT

A wide class of statistical models involving structured covariance matrices is studied. The models are defined in terms of homogeneous constraints imposed on population covariance matrices, with particular attention to models with vanishing conditional covariances. Special cases include factor analysis models, regression path models, linear covariance structures and linear inverse covariance structures. Certain scale related properties of maximum likelihood estimates for this class are derived, and numerical algorithms for computing constrained covariance estimates are developed. The final chapter considers questions of model selection, fitting and assessment using certain sample statistics whose joint distribution is approximately normal, and emphasizes graphical and heuristic rather than formal techniques. To illustrate a number of these points, a numerical example is provided which involves observations on a set of 66 variables.

ACKNOWLEDGEMENTS

I would like to thank my advisor Professor D.R. Cox for his generous encouragement and guidance throughout the course of this work. I am also indebted to R. Gnanadesikan, C.L. Mallows and M.D. Godfrey for valuable comments on earlier drafts, and to E.B. Fowlkes for providing the data for Example 5.1.

This research was carried out under the Doctoral Support Program of Bell Telephone Laboratories.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EXAMPLES

Chapter 1 - INTRODUCTION

A majority of the standard multivariate statistical methodology is based on assumptions of multivariate normality in which population and sample covariance matrices play a central role. Furthermore, with the development of new methods which relax normality assumptions (Dempster, 1971), it is likely that covariance matrices, perhaps with slightly modified definitions, will remain important (Devlin, et al., 1975). While a sample covariance matrix may represent a significant reduction of an extensive set of data, if the number of variables is moderately large then the number of sample covariances, and of covariance parameters that they estimate, is still very large. In order effectively to summarize the relationships among variables and to expose important features of the data (Tukey and Wilk, 1966) further reduction of the covariance matrix will be required. Dempster (1972) offers additional arguments in favor of parsimonious use of parameters in statistical models.

In a sense, most multivariate methods can be regarded as attempts to impose additional structure on covariance matrices. Psychometricians have approached the problem explicitly in these terms, with hypotheses of patterns and symmetries (Guttman, 1955; Bock and Bargmann, 1966; Olkin and Press, 1969: Mukherjee, 1970). Factor analysis and simultaneous-equation regression analysis correspond to lower dimensional factorizations of covariance matrix; and recently some authors have become interested in additive decompositions (Anderson, 1969; Rao, 1972).

A main objective of the present work is to draw together many of these models as special cases of a general structured covariance model, in which structure is defined by requiring certain functions of the variance and covariance parameters to vanish. Having done this we are able to derive certain properties of the general class, and to develop some unified approaches to estimation and model assessment.

Chapter 2 explores in some detail one such property which essentially fixes the overall scale of the maximum likelihood parameter estimates in relation to the observations, regardless of the structural constraints, and leads to a simplification of likelihood ratio statistics. From one viewpoint this is a direct consequence of scale invariance, but it also is shown to hold for certain members of the exponential family which, because of the discreteness of sample spaces, are not scale invariant.

In Chapter 3 the structured Wishart model is formulated in general terms and a number of examples are discussed. A concise expression is derived for asymptotic covariances (under normal theory assumptions) of statistics that can be used to test structural constraints, and the concepts of sufficiency, ancillarity and invariance are employed as a guide for choosing appropriate test statistics.

Numerical maximum likelihood methods for fitting structured Wishart models are considered in Chapter 4. Of two iterative algorithms discussed, the first is essentially an adaptation to the Wishart problem of a Newton-Raphson procedure modified by Aitchison and Silvey (1960) to deal with equality constraints. The second is an apparently new fix-point algorithm based on a special property of the Wishart likelihood function. In order to compare costs and determine practical limitations, we report the results of computer runs in which both algorithms were applied to each of several numerical examples.

Because of the inherent limitations of maximum likelihood methods applied to structured covariance models and in the belief that these models can still be useful for exploring extensive and complex sets of data, we develop in Chapter 5 some methods based on the asymptotic joint normality of sample correlation statistics. One interesting consequence is that such disparate models as those which hypothesize zeros in Σ^{-1} on the one hand, and in Σ on the other, can be treated very similarly in this context. The ideas developed are shown to be relevant in the areas of model selection, fitting, and assessment. Although the approach is based on underlying assumptions of normality, it is recognized that the assumptions are often not satisfied in practice, so greater emphasis is placed on the use of graphical and heuristic rather than formal techniques. A final example illustrates a number of these points in connection with a search for inherent symmetries among a large set of observational variables.

A short appendix is included to provide some matrix results used in the text.

Chapter 2
SOME PROPERTIES OF MAXIMUM LIKELIHOOD ESTIMATES

2.1 Introduction

In several of the following chapters we deal with p-variate normal probability models where the population covariance matrix is unknown but assumed to have some known structure, and we often work with maximum likelihood estimates. The Wishart log-likelihood functions that occur all involve a term proportional to $\mathrm{tr}(\Sigma^{-1}S)$, where Σ is the Wishart matrix parameter and S is a Wishart observation, and in a wide variety of situations the value of this term in the maximized likelihood is found to be exactly p, the dimension of the matrices Σ and S. That is

$$\mathrm{tr}(\hat{\Sigma}^{-1}S) = p, \tag{2.1}$$

where $\hat{\Sigma}$ is a maximum likelihood estimate of Σ.

Since this result occurs repeatedly for the models considered, the present chapter examines it in some detail. We show that it generalizes in two apparently different but connected ways, and give examples of each. Finally we discuss some consequences and interpretations, and consider similar properties of estimates calculated by other methods.

Bock and Bargmann (1966) noted Eq. 2.1 in the context of fitting multivariate normal covariance structures. We shall consider generalizations of it, but the proofs of the theorems in this chapter are essentially like theirs.

2.2 Generalization of $\mathrm{tr}(\hat{\Sigma}^{-1}S) = p$ to exponential families

The equation $\mathrm{tr}(\hat{\Sigma}^{-1}S) = p$ is a special case of a property of certain exponential families of distributions involving only homogeneous functions of the parameters. We shall call a real-valued function $f(x)$ of the vector variable x homogeneous of degree p if $f(\lambda x) = \lambda^p f(x)$ for any positive number λ and for almost every x, and we note that the following three conditions are equivalent to each other (Euler's theorem):

1. $f(\lambda x) = \lambda^p f(x)$, a.e. x,
2. $Df(x) = pf(x)$, a.e. x,
3. $D \log f(x) = p$, a.e. x,

where D is the differential operator

$$D = \Sigma_k x_k \frac{\partial}{\partial x_k} .$$

The main result of this section is given by the following theorem.

Theorem 2.1

Suppose the multivariate probability density function $f(y;\theta)$ belongs to an exponential family indexed by the vector parameter θ,

$$f(y;\theta) = \exp \left\{ \sum_{k=1}^{m} a_k(\theta)b_k(y) + c(\theta) + d(y)\right\}, \quad (2.3)$$

and suppose that the parameter space Ω is closed under non-negative scalar multiplication. Further suppose that the functions $a_k(\theta)$ are all homogeneous of the same degree r in the parameters $\theta_1,\ldots,\theta_q$, and that $c(\theta)$ and $a_k(\theta)$ are almost everywhere differentiable with respect to all arguments. Let $\hat{\theta} = \hat{\theta}(y)$ be a value of θ that maximizes $f(y;\theta)$ over Ω for a given y. Then,

$$\sum_{k=1}^{m} a_k(\hat{\theta})b_k(y) = -\frac{1}{r} Dc(\hat{\theta}), \quad (2.4)$$

so that

$$f(y;\hat{\theta}) = \exp \{c(\hat{\theta}) - Dc(\hat{\theta})/r + d(y)\}, \quad (2.5)$$

where D is the differential operator $\sum_i \theta_i \frac{\partial}{\partial\theta_i}$.

Proof

Since $\hat{\theta}$ is in Ω, so is $\lambda\hat{\theta}$ by assumption for any positive scalar λ. Define $g(\lambda)$ to be $\log f(y;\lambda\hat{\theta})$. Since $\hat{\theta}$ maximizes $f(y;\theta)$ for fixed y, the function $g(\lambda)$ must attain its maximum at $\lambda = 1$, and we have

$$\frac{dg}{d\lambda}\Big|_{\lambda=1} = 0 = \sum_{k=1}^{m} (Da_k(\hat{\theta}))b_k(y) + Dc(\hat{\theta}).$$

By Euler's equation, $Da_k(\hat{\theta}) = ra_k(\hat{\theta})$, since a_k is homogeneous of degree r. Formulas 2.4 and 2.5 follow directly.

With one additional assumption about the form of the exponential family, a further simplification occurs, and we have

Corollary 2.1

If, in addition to the assumptions of the theorem, $c(\theta)$ is the logarithm of a homogeneous function of degree t, then,

$$\sum_{k=1}^{m} a_k(\hat{\theta})b_k(y) = -\frac{t}{r}, \tag{2.6}$$

and

$$f(y;\hat{\theta}) = \exp\{c(\hat{\theta}) - \frac{t}{r} + d(y)\}. \tag{2.7}$$

The proof follows from the fact that $Dc(\hat{\theta}) = t$, which is a result of Euler's theorem given above.

We note that the condition that Ω is closed under multiplication by positive scalars is equivalent to saying that in all the models under consideration the parameters have a free (unknown) overall scale factor.

Example 2.1 - Contingency table with Poisson counts

We consider an r by c contingency table of counts $y = \{y_{ij}\}$ which have independent Poisson distributions with parameters $\lambda = \{\lambda_{ij}\}$, $(i=1, \ldots, r;\ j=1, \ldots, c)$, and we wish to examine the plausibility of various hypotheses concerning the λ_{ij}. (This is closely related to a multinomial model in which $N = \Sigma\, y_{ij}$ is fixed.) The probability density function is

$$f(y;\lambda) = e^{-\Sigma\lambda_{ij}} \prod \lambda_{ij}^{y_{ij}} / \prod (y_{ij}!)$$

$$= \exp\left\{ \sum (\log \lambda_{ij})\, y_{ij} - \sum \lambda_{ij} - \sum \log(y_{ij}!) \right\} ,$$

with the following correspondence to the exponential family, Eq. 2.3,

$$a_k = \log \lambda_{ij}, \qquad b_k = y_{ij}$$

$$c = - \sum \lambda_{ij}, \qquad d = - \sum \log\,(y_{ij}!) .$$

If we reexpress the model in terms of the natural parameters of the exponential family, $\rho_{ij} = \log \lambda_{ij}$, the density becomes

$$f(y;\rho) = \exp\left\{ \sum \rho_{ij}\, y_{ij} - \sum e^{\rho_{ij}} - \sum \log\,(y_{ij}!) \right\} .$$

The new parameter space is $\Omega = \mathbb{R}^{rc}$ which is closed under scalar multiplication. Now suppose that the various hypotheses of interest correspond to subsets of Ω each of which is closed under scalar multiplication. For instance we might assume that $\rho_{ij} = \alpha_i + \beta_j$ for unknown parameters α_i and β_j, which derives from assuming independence of row and column effects. Then the assumptions of Theorem 2.1 are satisfied for the full model and the submodels, so the maximum likelihood estimate $\hat{\rho}$ for any of these models satisfies,

$$\sum \hat{\rho}_{ij}\, y_{ij} = D\left(\sum e^{\hat{\rho}_{ij}}\right)$$

$$= \sum \hat{\rho}_{ij} e^{\hat{\rho}_{ij}}$$

and

$$f(y;\hat{\rho}) = \sum e^{\hat{\rho}_{ij}}\,(\hat{\rho}_{ij}-1),$$

or, in terms of the original λ's,

$$f(y;\hat{\lambda}) = \sum \hat{\lambda}_{ij}\,(\log \hat{\lambda}_{ij}-1).$$

Example 2.2 - Exponential regression

Suppose that the random variable Y_j has an exponential distribution with parameter $\lambda_j > 0$, given by

$$f_{Y_j}(Y_j;\lambda_j) = \frac{1}{\lambda_j} \exp\,(-Y_j/\lambda_j), \qquad (j=1,\ldots,n),$$

that the Y_j are independent, and that the parameter λ_j, which is the expected value of Y_j, is related to a p-vector x_j of concomitant (or design) variables through an unknown vector parameter β by,

$$\lambda_j = | x_j^T \beta |^{q} ,$$

where q is a fixed number, positive or negative. The likelihood of $y = (y_1,\ldots,y_n)^T$ is

$$f_Y(y \mid x;\beta) = \exp - \left\{ \sum_{j=1}^{n} | x_j^T \beta |^{-q} y_j + \log \prod_{j=1}^{n} | x_j^T \beta |^{q} \right\} .$$

Comparing this to the general form of Eq. 2.3 we see that the functions $a_k(\beta) = | x_j^T \beta |^{-q}$ are homogeneous of degree $-q$ in the elements of β, and the function $c(\beta) = \log \prod_{j=1}^{n} | x_j^T \beta |^{q}$ is the log of a homogeneous function of degree qn. Depending on q and x, the feasible parameter space for β may be smaller than $\mathbb{R}^p$ but it is always closed under positive scalar multiplication, and we will only entertain subhypotheses that place homogeneous constraints on β.

All of the assumptions of Theorem 2.1 and Corollary 2.1 are seen to hold, and we conclude that any maximum likelihood estimate $\hat{\beta}$ satisfies,

$$\sum_{j=1}^{n} | x_j^T \hat{\beta} |^{-q} y_i = - \frac{qn}{-q} = n ,$$

and

$$f_Y (y | x;\hat{\beta}) = e^{-n} / \prod_{j=1}^{n} | x_j^T \hat{\beta} |^q .$$

Furthermore, if we denote a fitted y_j by $\hat{y}_j$, we have

$$\hat{y}_j = E_{\hat{\beta}} (y_j | x_j) = \hat{\lambda}_j = | x_j^T \hat{\beta} |^q ,$$

and

$$f_Y (y_j | x;\hat{\beta}) = e^{-n} / \prod_{j=1}^{n} \hat{y}_j .$$

If we use a likelihood ratio to compare two hypotheses H_A and H_B concerning β, we find,

$$LR = \frac{\text{lik}(\hat{\beta}_A;y)}{\text{lik}(\hat{\beta}_B;y)} = \prod_{j=1}^{n} \left(\frac{\hat{y}_{Bj}}{\hat{y}_{Aj}} \right).$$

That is, the likelihood ratio depends only on the ratios of fitted values under the hypotheses.

We note that if the exponential parameters λ_j were linked to the design variables x_j and regression parameters β by

$$\lambda_j = \exp\ (x_j^T\ \beta),$$

then the conditions of Theorem 2.1 would not be fulfilled, and the results would not hold.

Example 2.3 - Structured Wishart models

The most important example for subsequent chapters is a model in which a random p by p matrix S is assumed to have a Wishart distribution $W_p(n, \frac{1}{n}\Sigma)$, where Σ is unknown but required to lie in a subspace Ω_0 of the space Ω of all positive definite matrices, and Ω_0 is assumed to be closed under multiplication by any positive constant. The matrix S, for instance, might be a sample covariance matrix from a multivariate normal distribution with population covariance matrix Σ. The likelihood function is

$$\mathrm{lik}(\Sigma;S) = \frac{|S|^{\frac{1}{2}(n-p-1)}\exp\{-\frac{n}{2}\ \mathrm{tr}(\Sigma^{-1}S)\}}{n^{-\frac{1}{2}p(n+p-1)}2^{\frac{1}{2}np}\pi^{\frac{1}{4}p(p-1)}|\Sigma|^{n/2}\prod_{i=1}^{p}\Gamma\ (\frac{n+1-i}{2})}\ , \tag{2.9}$$

which can be rewritten as

$$\mathrm{lik}(\Sigma;S) = \exp\left\{-\frac{n}{2}\sum_{i,j}\sigma^{ij}\ s_{ji} + \log|\Sigma|^{-\frac{n}{2}} + d(S)\right\}\ . \tag{2.10}$$

Comparing the exponential family density Eq. 2.3 with Eq. 2.10, we see that θ and y correspond to Σ and S, respectively; $a_k(\theta)$ corresponds to σ^{ij}, the (i,j) element of Σ^{-1}, which is a homogeneous function of degree $r = -1$ in the variables $(\sigma_{11}, \sigma_{12}, \ldots, \sigma_{pp})$; $b_k(y)$ corresponds to $-\frac{n}{2} s_{ji}$; and $c(\theta)$ corresponds to $\log |\Sigma|^{-\frac{n}{2}}$ which is the logarithm of a homogeneous function of degree $t = -\frac{np}{2}$. All of the conditions for Corollary 2.1 are met and we conclude that if $\hat{\Sigma}$ maximizes $\text{lik}(\Sigma;S)$ over Ω_0, then

$$\text{tr}(\hat{\Sigma}^{-1}S) = \sum_{i,j} \hat{\sigma}^{ij} s_{ji} = -\frac{2}{n}\left(-\frac{t}{r}\right) = p. \tag{2.11}$$

We note that the exponential family under consideration could equally well have been parameterized by Σ^{-1} instead of Σ. Since the elements of Σ are homogeneous functions of degree -1 in the elements of Σ^{-1}, the functions $a_k(\Sigma^{-1})$ and $\exp(c(\Sigma^{-1}))$ in the reparameterization are again homogeneous and the same results follow anew.

Corollary 2.1 can be applied in this way to any Wishart model in which the parameter space Ω_0 is a subset of the space of positive definite matrices defined by a set of constraints of the form $h_1(\Sigma) = \ldots = h_q(\Sigma) = 0$, where the h_k are all homogeneous functions in the elements of Σ. The condition that Ω_0 is closed under multiplication by a positive constant follows immediately. This is a very large class of models and includes many of the specialized models in common use in a number of areas of application. These include:

1. Models in which Σ has a block structure, i.e. some or all of the elements σ_{ij} are equal to each other in diagonal and in off-diagonal blocks. These models result from assumptions of interchangeability of subsets of the X_i variables.

2. Models in which elements of Σ are constant in diagonal stripes. Covariance matrices of this sort arise for example in the analysis of stationary Gaussian time series. In addition there might be assumptions concerning the ratios of values in consecutive stripes, a constant ratio corresponding to first-order autoregressive process.

3. Models which specify virtually any kind of pattern hypothesis on the elements ρ_{ij} of the correlation matrix. Since the ρ_{ij} are homogeneous functions of degree zero in the elements of Σ, any function of them is also homogeneous. In particular, some or all of the ρ_{ij} can be assumed equal to known constants.

4. Linear covariance structures, as defined by T. W. Anderson (1969), in which either Σ or Σ^{-1} can be represented as a linear combination with unknown weights of a set of known symmetric matrices, say, $\Sigma = \xi_1 B_1 + \ldots + \xi_k B_k$. The set Ω_0 of permissible ξ's clearly has the necessary free scale parameter for the simplication results to hold, and

the representation in terms of homogeneous constraints can be shown explicitly as follows. Let $\underset{\sim}{\xi} = (\xi_1, \ldots, \xi_k)^T$, and let $\underset{\sim}{\sigma}$ and $\underset{\sim}{b}_i$ represent the distinct elements of Σ and B_i rolled out into column vectors of length $q = \frac{1}{2}p(p+1)$. Also, let B be the matrix whose columns are $\underset{\sim}{b}_i$. Then the linear decomposition of Σ can be written as $\underset{\sim}{\sigma} = B\underset{\sim}{\xi}$, which means that $\underset{\sim}{\sigma}$ is constrained to lie entirely in the linear space $\mathcal{L}(B)$ spanned by the columns of B. This in turn means that the projection of $\underset{\sim}{\sigma}$ into the space orthogonal to $\mathcal{L}(B)$ vanishes, a property which can be written as,

$$(I - B(B^TB)^- B^T)\underset{\sim}{\sigma} = 0.$$

Similar results hold when Σ^{-1} has a linear structure.

5. Factor analysis models, in which Σ is decomposed into $\Sigma = \Theta\Delta\Theta^T + \Psi$ where Δ and Ψ are diagonal, Θ is orthogonal, and Δ has rank less than $p = \text{rank}(\Sigma)$. Since the elements of Λ and Ψ are unspecified, it is clear that Σ has the free overall scale parameter required by Theorem 2.1 and Corollary 2.1.

6. Models with certain constraints on the shape or orientation of the distribution of the variables X_i. If $\Sigma = \Theta\Lambda\Theta^T$ is the eigenanalysis of Σ, Θ can be interpreted as describing the orientation and Λ the shape of the distribution, where "shape" means both size and relative spread along the principal axes. The elements of Λ and Θ are

homogeneous functions of degree 1 and 0, respectively, in the elements of Σ. Hence Eq. 2.11 applies to classes of models that restrict Θ in any differentiable way, or that restrict Λ in any homogeneous way.

7. Causal path models, which are characterized by the vanishing of certain conditional covariances among the variables X_i. These models are examined in some detail in chapters below.

8. As an example of a structured Wishart model which does not satisfy the requirements of Theorem 2.1, we assume that some or all of the variances σ_{ii} are equal to known constants. The parameters are then essentially the correlations (which may satisfy further pattern hypotheses), and the functions of them that enter the likelihood do not meet the homogeneity requirements. Let Δ^2 be a diagonal matrix with the known variances as diagonal elements. Although Theorem 2.1 does not apply, one can employ arguments like those in its proof to obtain

$$\mathrm{tr}(\hat{\Sigma}^{-1}S) = \mathrm{tr}(\hat{\underset{\sim}{\rho}}^{-1}\ S^*) = p - \mathrm{tr}\ \hat{\underset{\sim}{\rho}}^{-2}\ (S^* - \hat{\underset{\sim}{\rho}}),$$

where $\underset{\sim}{\rho} = \Delta^{-1}\Sigma\ \Delta^{-1}$ is the population correlation matrix, and S^* is $\Delta^{-1}\ S\ \Delta^{-1}$.

2.3 Generalization to scale invariant families

The property $\mathrm{tr}(\hat{\Sigma}^{-1}S) = p$ can also be viewed as a special case of a result that derives from scale invariance, which we define as follows. Suppose that the m-vector Y has density function $f_Y(y;\theta)$ with parameter space Ω, and $f_Y(y;\theta)$ is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^m$ and differentiable with respect to each y_i and each θ_j. We wish to examine various hypotheses corresponding to subsets Ω_0, $\Omega_1, \ldots,$ of Ω. Define a group $\mathcal{G}$ of scale transformations on the sample space by

$$g_\lambda : Y \rightarrow \lambda Y, \quad \text{for each } \lambda > 0,$$

and assume that $\mathcal{G}$ induces a group $\mathcal{G}^*$ of transformations on Ω,

$$g_\lambda^* : \theta \rightarrow \theta^{(\lambda)},$$

such that every Ω_i is preserved under $\mathcal{G}^*$. We will also assume that each component of $\theta^{(\lambda)}$ is a differentiable function of λ. The model defined by $f_Y(y;\theta)$ and any one of the subsets Ω_k is then said to be scale invariant, and we have the following theorem.

Theorem 2.2

If $\hat{\theta}$ is a maximum likelihood estimate of the vector parameter in a scale-invariant model with parameter space Ω_0 as defined above, and if θ is held fixed at $\hat{\theta}$, then $f_r(y;\hat{\theta})$ satisfies,

$$D_y \log f_Y(y;\hat{\theta}) = \sum_{k=1}^{m} \frac{\partial \log f_Y(y;\hat{\theta})}{\partial y_k} y_k = - m. \qquad (2.14)$$

Proof

Let A be any measurable set in the sample space and let λ be a positive scalar, then

$$pr(\lambda Y \in A;\theta) = pr(Y \in A;\theta^{(\lambda)}).$$

Equivalently, in terms of density functions,

$$f_Y(\lambda y;\theta)d(\lambda y) = f_Y(y;\theta^{(\lambda)})dy.$$

The Jacobian of the transformation $y \to \lambda y$ is λ^m, so

$$d(\lambda y) = \lambda^m dy,$$

and

$$f_Y(\lambda y;\theta)\lambda^m = f_Y(y;\theta^{(\lambda)}).$$

Taking logarithms and differentiating both sides of this equation with respect to λ, we are led to,

$$\sum_k \frac{\partial \log f_Y(\lambda y;\theta)}{\partial y_k} \frac{y_k}{\lambda} + \frac{m}{\lambda} = \frac{\partial \log f_Y(y;\theta^{(\lambda)})}{\partial \lambda} . \qquad (2.15)$$

Now fix θ at the maximum likelihood estimate $\hat{\theta}$ for the hypothesis Ω_0. Since $\hat{\theta}^{(\lambda)} \in \Omega_0$ for each λ by assumption, and $\hat{\theta}^{(1)} = \hat{\theta}$, we must have

$$\frac{\partial f_Y(y;\theta^{(\lambda)})}{\partial\lambda} = 0 \qquad \text{for} \quad \lambda = 1,$$

and Eq. 2.15 becomes

$$\sum_k \frac{\partial \log f_Y(y;\hat{\theta})}{\partial y_k} y_k = - m,$$

which completes the proof.

To help clarify this theorem we note the following points. First, scale invariance as defined here is slightly more general than scale invariance in a family defined by,

$$f_Y(y;\theta,\sigma) = \frac{1}{\sigma^m} \quad g\left(\frac{y}{\sigma} \; ; \; \theta\right),$$

since θ in the expression on the right must in general be allowed to depend on σ. Second, despite the similarity between Eq. 2.14 and Euler's formula, it does not follow that for fixed $\hat{\theta}$, $f(y;\hat{\theta})$ is homogeneous in the arguments $y_1, \ldots, y_m$. The reason is that Eq. 2.14 holds only at the one point y for which $\hat{\theta}$ was calculated. However, if $\hat{\theta}$ is allowed to vary with y, then with one small additional assumption (which is perhaps not strictly necessary)

$f(y;\hat{\theta}(y))$ is homogeneous, as shown in the following theorem.

Theorem 2.3

Under the conditions of Theorem 2.2, if Ω_0 can be reparameterized in a neighborhood of $\hat{\theta}(y)$ in a continuous and differentiable way by a vector φ whose elements are not subject to equality constraints in that neighborhood, then the probability density function $f_Y(y;\hat{\theta}(y))$ is homogeneous of degree $-m$ in $y_1,\ldots,y_m$.

Proof

$\hat{\theta}(y)$ corresponds to $\hat{\varphi}(y)$ in the new parameterization, and $\hat{\theta}(y) = \theta(\hat{\varphi}(y))$. Writing $g(y) = \log f(y;\hat{\theta}(y))$ we have,

$$Dg(y) = \frac{\partial g(\lambda y)}{\partial \lambda}\bigg|_{\lambda=1}$$

$$= \left[\sum_k \frac{y_k}{\lambda}\,\frac{\partial \log f(\lambda y;\hat{\theta}(\lambda y))}{\partial y_k} + \sum_j \frac{\partial \log f\{\lambda y;\theta(\hat{\varphi}(\lambda y))\}}{\partial \hat{\varphi}_j}\,\frac{\partial \hat{\varphi}_j}{\partial \lambda}\right]_{\lambda=1} .$$

Now the derivatives of $\log(f)$ with respect to $\hat{\varphi}_j$ are zero by the maximum likelihood conditions, so the second summation is zero. Also, the first summation is equal to $-m$ for $\lambda=1$ by Theorem 2.2. Hence $Dg(y) = -m$ for almost all y, and it follows that $f(y;\hat{\theta}(y))$ is homogeneous of degree $-m$ in $y_1,\ldots,y_m$.

Example 2.4 - Structured Wishart models (cont'd)

All of the Wishart models discussed in Example 2.3 are scale invariant, since the transformation g: $S \to \lambda S$ induces g*: $\Sigma \to \frac{1}{\lambda} \Sigma$ on Ω. Differentiating the log of the probability density, Eq. 2.9, with respect to S we have

$$\frac{\partial \log f(S;\Sigma)}{\partial S} = \frac{n - (p+1)}{2} S^{-1} - \frac{n}{2} \Sigma^{-1}. \qquad (2.16)$$

Before applying Theorem 2.2 we must take into account the symmetry of S. Although S is a p by p matrix, there are only $\frac{1}{2}p(p+1)$ functionally independent elements, which we can take to be the elements in the upper triangle, including the diagonal. Then each element of S is a homogeneous function of degree 1 in these $\frac{1}{2}p(p+1)$ elements, and we can show

$$\sum_{i\leq j} \frac{\partial \log f(S)}{\partial s_{ij}} s_{ij} = \sum_{i,j} \frac{\partial \log f(S)}{\partial s_{ij}} s_{ij} , \qquad (2.17)$$

where the derivatives on the left hand side take account of symmetry. The correct value of m in Eq. 2.14 is $\frac{1}{2}p(p+1)$ and combining Eqs. 2.14, 216, and 2.17, we have

$$\frac{-p(p+1)}{2} = \operatorname{tr}\left\{ \frac{\partial \log f(S;\hat{\Sigma})}{\partial S} S \right\} =$$

$$= \frac{n - (p+1)}{2} p - \frac{n}{2} \operatorname{tr}(\hat{\Sigma}^{-1} S) .$$

It follows that $\operatorname{tr}(\hat{\Sigma}^{-1}S) = p$.

The structured Wishart models described in Examples 2.3 and 2.4 can be derived from normal models in which the population means are unknown and unrestricted. The Wishart likelihoods can be equivalently thought of as marginal or conditional likelihoods. As an example where Theorem 2.1 does not apply but scale invariance does, we consider the normal model itself in which there may be restrictions on the means as well as covariances.

Example 2.5 - Normal model with structured mean and covariance

Suppose that Y_i $(i=1,\ldots,n)$ are independent normally distributed p-dimensional vectors with common mean vector μ and covariance matrix Σ. The parameter space for the full model is $\Psi = \mathbb{R}^p \times \Omega$ where Ω is the space of positive definite matrices. We want to compare the full model to submodels that impose homogeneous constraints on both μ and Σ. Let Y be the n by p matrix whose rows are Y_i^T, and let M be the n by p matrix whose rows are all equal to μ^T. Then the probability density function is

$$f_Y(y;\mu,\Sigma) = \exp\left[-\tfrac{1}{2}\operatorname{tr}\left\{\Sigma^{-1}(Y-M)^T(Y-M)\right\} - \frac{n}{2}\log|\Sigma| \right.$$

$$\left. - \frac{np}{2}\log(2\pi)\right] . \tag{2.18}$$

This is an exponential family, but the functions corresponding to $a_k(\theta)$ are not homogeneous of the same degree in the elements of μ and Σ, so Theorem 2.1 does not apply.

Nevertheless this model is scale invariant, the transformation g: $Y \rightarrow \lambda Y$ inducing the transformation g^*: $(\lambda, \Sigma) \rightarrow (\lambda^{-1}\mu, \lambda^{-2}\Sigma)$ on the parameter space, so Theorem 2.2 does apply. Let $(\hat{\mu}, \hat{\Sigma})$ be the maximum likelihood estimates over some homogeneous subspace $\Psi_0 \subset \Psi$. Using matrix derivative formulas A7 and A8 from Appendix A, we have

$$\frac{\partial \log f(y;\mu,\Sigma)}{\partial Y} = \hat{M}\,\hat{\Sigma}^{-1} - Y\hat{\Sigma}^{-1},$$

and from Eq. 2.14, noting that the proper value of m is now np,

$$-\ np = tr\left\{ \frac{\partial \log f(Y;\hat{\mu},\hat{\Sigma}\ )}{\partial Y}\ Y^T \right\} = -\ tr(Y-\hat{M})\hat{\Sigma}^{-1}Y^T. \tag{2.19}$$

In order to simplify this result further we can write $\tilde{S} = \frac{1}{n}\,(Y-\hat{M})^T(Y-\hat{M})$ which is a (biased) mean squares and cross-products matrix corrected not for the usual sample means, but for means estimated according to the constraints. Then applying Lemma 2.3 below, Eq. 2.19 becomes,

$$\mathrm{tr}(\hat{\Sigma}^{-1}\tilde{S}) = p, \tag{2.20}$$

which is analogous to Eq. 2.1.

The lemma that is required to complete the derivation is

Lemma 2.3

If Y is an n×p random matrix whose rows are independent observations from a multinormal distribution with unknown mean and covariance,

$$EY = M \quad \text{and} \quad E(Y-M)^T(Y-M) = n\Sigma,$$

where M and Σ may be subject to homogeneous restrictions, and if $\hat{M}$ maximizes the likelihood for some fixed estimate $\hat{\Sigma}$ of Σ, then

$$\mathrm{tr}(\hat{M}\ \hat{\Sigma}^{-1}\ Y^T) = \mathrm{tr}(\hat{M}\ \hat{\Sigma}^{-1}\ \hat{M}^T).$$

Proof

The proof follows the pattern of other proofs in this chapter by writing the likelihood as $f_Y(y;\lambda\hat{M}, \hat{\Sigma})$, differentiating with respect to λ, and setting the result to zero for λ=1.

2.4 Generalization to a class of fitting criteria

The foregoing sections have dealt with a single estimation criterion, maximum likelihood, and several families of probability models. We now fix our attention on the family of structured Wishart models and consider various possible estimation criteria. In particular, we show that maximum likelihood, least squares, and generalized least squares are all formally special cases of a general criterion, and that the condition $\mathrm{tr}(\hat{\Sigma}^{-1}S) = p$ is a special case of a relationship that holds more generally.

For a fixed sample covariance matrix S, the basic problem is to find a matrix Σ in a given homogeneous set Ω_0 that optimizes some objective function, $g(\Sigma;S)$. Maximum (Wishart) likelihood corresponds to minimizing

$$\text{ML: } \log|\Sigma| - \mathrm{tr}\,(\Sigma^{-1}S), \qquad (2.21)$$

unweighted least squares requires minimizing

$$\text{LS: } \tfrac{1}{2}\mathrm{tr}(\Sigma-S)^2 = \mathrm{tr}(\tfrac{1}{2}\Sigma^2-\Sigma S) + \tfrac{1}{2}\mathrm{tr}S^2,$$

and generalized least squares in which the squared deviations of $\hat{\sigma}_{ij}$ from s_{ij} are weighted by the inverse of their asymptotic covariance matrix, leads to minimizing

$$\text{GLS: } \tfrac{1}{2}\mathrm{tr}\{\Sigma^{-1}(\Sigma-S)\}^2 = \mathrm{tr}\{-\Sigma^{-1}S + \tfrac{1}{2}(\Sigma^{-1}S)^2\} + \tfrac{1}{2}p.$$

Apart from terms not involving Σ, these last two are clearly special cases of

$$g(\Sigma;S,a,b) = \mathrm{tr}\left\{ \frac{1}{a+1} \quad \langle\Sigma^{(a+1)},\ S^{(b-1)}\rangle \right.$$

$$\left. - \frac{1}{a} \quad \langle\Sigma^{(a)},\ S^{(b)}\rangle \right\}, \qquad (2.22)$$

where a and b are positive or negative integers, and an expression like $\langle\Sigma^{(2)}, S^{(2)}\rangle$ means $\Sigma S \Sigma S$. In fact the ML criterion also fits the form of $g(\Sigma;S,a,b)$ if the notational conventions in the following paragraph are adopted.

If X is any square symmetric matrix with eigenvalue decomposition $X = H\Lambda H^T$, and f is any analytic function, let f(X) be the matrix-valued function

$$f(X) = H \text{ diag } \{f(\lambda_i)\}H^T.$$

In particular, the matrix logarithm function $\log(\Sigma)$ is well-defined for any positive definite Σ, and its inverse is the matrix exponential function. It follows immediately that

$$\log|\Sigma| = \mathrm{tr}\{\log(\Sigma)\}.$$

Furthermore, as in the case of the scalar logarithm we have

$$\lim_{\alpha\to 0} \frac{\log \ \Sigma^{\alpha} - I}{\alpha} = \log \Sigma ,$$

so that in a formal sense $\frac{1}{0}\Sigma^0$ is log Σ, apart from a "constant" term. Note especially that the formal rule for differentiating a power works properly for the log expressed this way, when $0\frac{1}{0}$ is taken to be 1. For consistency, we define Σ^0 (without the factor $\frac{1}{0}$) to be the identity matrix. (See Eq. A6.)

With these conventions established, Eq. 2.21 becomes

$$\text{ML:}\quad \operatorname{tr}\{\tfrac{1}{0}\Sigma^0 - \Sigma^{-1}S\},$$

which is in the form of $g(\Sigma;S,a,b)$.

Now letting $\hat{\Sigma}$ be a point in the Ω_0 that maximizes g, and observing that the derivative of $g(\lambda\hat{\Sigma};S,a,b)$ with respect to λ must vanish for $\lambda=1$, we have

$$\operatorname{tr}\left\{\langle\hat{\Sigma}^{(a+1)},S^{(b-1)}\rangle - \langle\hat{\Sigma}^{(a)},S^{(b)}\rangle\right\} = 0. \qquad (2.23)$$

This is the generalization to $g(\Sigma;S,a,b)$ of the relationship $\operatorname{tr}(\hat{\Sigma}^{-1}S) = p$.

Table 2.1 summarizes results for various values of a and b. The cases indicated by * require some interpretation. LS* is simple least squares applied to the residuals $\hat{\sigma}^{ij} - s^{ij}$. GLS* is generalized least squares applied to the same residuals, but using the fact that asymptotically the covariances of elements of S^{-1}

TABLE 2.1 - Relationship of $\hat{\Sigma}$ to S for various estimation methods

Method	a	b	Equation 2.23
ML	-1	1	$\mathrm{tr}(I - \hat{\Sigma}^{-1}S) = 0 = \mathrm{tr}\,\hat{\Sigma}^{-1}(\hat{\Sigma} - S)$
LS	1	1	$\mathrm{tr}(\hat{\Sigma}^{2} - \hat{\Sigma}S) = 0 = \mathrm{tr}\,\hat{\Sigma}(\hat{\Sigma} - S)$
GLS	-2	2	$\mathrm{tr}(\hat{\Sigma}^{-1}S - \hat{\Sigma}^{-1}S\hat{\Sigma}^{-1}S) = 0 = \mathrm{tr}\,\hat{\Sigma}^{-1}S(I - \hat{\Sigma}^{-1}S)$
ML*	0	0	$\mathrm{tr}(\hat{\Sigma}S^{-1} - I) = 0 = \mathrm{tr}\,\hat{\Sigma}(S^{-1} - \hat{\Sigma}^{-1})$
LS*	-2	0	$\mathrm{tr}(\hat{\Sigma}^{-1}S^{-1} - \hat{\Sigma}^{-2}) = 0 = \mathrm{tr}\,\hat{\Sigma}^{-1}(S^{-1} - \hat{\Sigma}^{-1})$
GLS*	1	-1	$\mathrm{tr}(\hat{\Sigma}S^{-1}\hat{\Sigma}S^{-1} - \hat{\Sigma}S^{-1}) = 0 = \mathrm{tr}\,\hat{\Sigma}S^{-1}(\hat{\Sigma}S^{-1} - I)$

TABLE 2.2 - Taxonomy of estimation methods

	a: ...	-4	-3	-2	-1	0	1	2	3	...
	4									
	3									
	2			GLS						
b:	1				ML		LS			
	0			LS*		ML*				
	-1						GLS*			
	-2									
	-3									
	...									

bear the same relationship to Σ^{-1} as the covariances of elements of S bear to Σ. ML* can be interpreted as maximum Wishart likelihood if S^{-1} is assumed to have a Wishart distribution (which it does asymptotically).

In order for $\langle\Sigma^{(a)}S^{(b)}\rangle$ and $\langle\Sigma^{(a+1)}S^{(b-1)}\rangle$ both to be expandable as alternating factors of Σ and S, the values of a and b must be restricted to the region outlined in Table 2.2.

Returning to Eq. 2.23 and Table 2.1, since we can interpret tr(XY) as the inner product of two vectors in Euclidean $\mathbb{R}^{p^2}$ space, the last equation in each row of Table 2.1 is really an orthogonality relationship. In each case the vector of residuals (on some scale) is orthogonal to fitted values (on a possibly different scale). When the two scales are the same, we note the following interpretation. If A and B are two points (vectors) in Euclidean space and $A \perp (B-A)$, then the point A must be on the hypersphere whose diameter is the line segment from 0 to B. In the LS case, for instance, $\hat{\Sigma}$ must lie on the hypersphere whose diameter is $\overline{0,S}$.

Finally, we note an interesting relationship between GLS and ML Since $\hat{\Sigma}_{GLS}$ minimizes the function

$$h(\Sigma) = tr(\Sigma^{-1}S-I)^2 \geq 0,$$

we can employ the GLS line of Table 2.1 to obtain

$$h(\hat{\Sigma}_{GLS}) = tr\ \hat{\Sigma}^{-1}_{GLS}\ S(\hat{\Sigma}^{-1}_{GLS}\ S - I) - tr\ (\hat{\Sigma}^{-1}_{GLS}\ S - I)$$

$$= -\ tr(\hat{\Sigma}^{-1}_{GLS}\ S - I).$$

Hence,

$$tr(\hat{\Sigma}^{-1}_{GLS}\ S) = p - h\ (\hat{\Sigma}_{GLS}) \leq p.$$

Thus, $tr(\hat{\Sigma}^{-1}_{GLS}\ S)$ is always less than $tr(\hat{\Sigma}^{-1}_{ML}\ S)$ by an amount $h(\hat{\Sigma}_{GLS})$.

2.5 Consequences and interpretations

As a theoretical point the property developed and generalized above is interesting in itself, but we now consider ways in which it provides useful insight into the structure of models for which it holds.

First, consider the homogeneous exponential models satisfying the conditions of Corollary 2.1, with estimation done by maximum likelihood. Since the quantities $a_k = a_k(\theta)$ are the exponential parameters of the model and $b_k = b_k(y)$ are the sufficient statistics, the condition

$$\Sigma_k \hat{a}_k b_k = \text{constant}\ ,$$

can be interpreted to mean that the length of the projection of the estimated vector $\hat{\underset{\sim}{a}}$ of natural parameters on the vector $\underset{\sim}{b}$ of sufficient statistics is constant.

The main consequence of this result is a simplification of likelihood ratio statistics that occurs. If two competing submodels are defined by subsets Ω_A and Ω_B of the parameter space, and if $\hat{\theta}_A$ and $\hat{\theta}_B$ are the corresponding estimates of the parameters, then under Corollary 2.1 the likelihood ratio statistic for testing A against B becomes

$$LR = \frac{f(y;\hat{\theta}_A)}{f(y;\hat{\theta}_B)} = \exp\{c(\hat{\theta}_A) - c(\hat{\theta}_B)\}.$$

For the exponential regression model in Example 2.2 it was shown that LR is consequently just the product of ratios of fitted values. And for the broad class of structural Wishart models in Examples 2.3 and 2.4 it follows that the likelihood ratio statistic is

$$LR = \left\{\frac{\det(\hat{\Sigma}_A)}{\det(\hat{\Sigma}_B)}\right\}^{n/2} .$$

This means that LR depends only on the ratios of eigenvalues of $\hat{\Sigma}_A$ and $\hat{\Sigma}_B$. Insofar as each normal distribution corresponds to an ellipsoid in p-dimensional space, LR depends only on the "shapes" and relative "sizes" of the fitted distributions, and not on their "orientations" as determined by the eigenvectors. Nor does LR depend on the locations of the fitted distributions, since the property was extended in Example 2.5 to models with structure in the means. This principle has been generally overlooked in the statistical literature, since authors who develop various Wishart models with specialized structure proceed to calculate the likelihood contribution from $\mathrm{tr}(\hat{\Sigma}^{-1}S)$ in each case, always obtaining p.

Finally, we indicate the possibility of using the results of this chapter to compare maximum likelihood estimates of θ with estimates obtained by other methods. If we agree to compare two estimates $\tilde{\theta}_1$ and $\tilde{\theta}_2$ by computing the quantity

$$d(\tilde{\theta}_1,\tilde{\theta}_2) = \Sigma_k\{a_k(\tilde{\theta}_1) - a_k(\tilde{\theta}_2)\}b_k(y),$$

which is essentially the length of the projection of $\underset{\sim}{a}(\hat{\theta}_1) - \underset{\sim}{a}(\hat{\theta}_2)$ on $\underset{\sim}{b}(y)$, and is in the Wishart case

$$d(\tilde{\Sigma}_1,\tilde{\Sigma}_2) = \mathrm{tr}(\tilde{\Sigma}_1^{-1} - \tilde{\Sigma}_2^{-1})S,$$

then the position of $\hat{\theta}_{ML}$ along this coordinate direction is fixed and known, and the distance from any other estimate to $\hat{\theta}_{ML}$ can be obtained without calculating $\hat{\theta}_{ML}$ itself. The principle can be used in reverse as a computational aid. If, in a large and complex problem, an iterative calculation is to be used to obtain a maximum likelihood estimate for Σ so that a good initial value

$\Sigma_{(0)}$ will be helpful, one might first calculate, say, a least squares estimate $\hat{\Sigma}_{LS}$, and then derive $\Sigma_{(0)}$ from $\hat{\Sigma}_{LS}$ by requiring $tr(\Sigma^{-1}_{(0)} S) = p$ to hold. This could be accomplished by rescaling $\hat{\Sigma}_{LS}$ or by projecting $\hat{\Sigma}^{-1}_{LS}$ into the linear space defined by $tr(\Sigma^{-1}S) = p$.

Chapter 3

CONSTRAINED WISHART MODELS

3.1 Introduction

Chapter 2 examined some properties of estimators for certain general models that have either an exponential or a scale invariant structure. We now focus on the subclass of structured Wishart models discussed in Examples 2.3 and 2.4, working primarily with their formulation in terms of functional constraints among the elements of the Wishart parameter matrix. We first describe the general model and a number of special cases. We then develop some asymptotic formulas for use in subsequent chapters for model testing. The chapter concludes with a discussion of the limited extent to which the principles of minimal sufficiency, ancillarity and invariance can be applied to this class of models.

3.2 General formulation

Suppose that the random pxp matrix S has a Wishart distribution $W_p(n, \frac{1}{n}\Sigma)$, and that the parameter matrix Σ is positive definite. The probability density function is

$$f_S(S;\Sigma) = \frac{|S|^{\frac{1}{2}(n-p-1)} \exp\left\{-\frac{n}{2}\operatorname{tr}(\Sigma^{-1}S)\right\}}{n^{-\frac{1}{2}(n+p-1)} 2^{\frac{1}{2}np} \pi^{\frac{1}{4}p(p-1)} |\Sigma|^{\frac{1}{2}n} \prod_{i=1}^{p} \Gamma\left(\frac{n+1-i}{2}\right)} , \tag{3.1}$$

which leads to a likelihood function for Σ proportional to

$$\text{lik}(\Sigma;S) = \exp \frac{n}{2} \{\log |\Sigma^{-1}| - \text{tr}(\Sigma^{-1}S)\}. \qquad (3.2)$$

Following the general approach taken by Silvey (1970, Section 4.7), we shall be interested in a general model corresponding to a parameter space

$$\Omega = \{\Sigma: \ \Sigma \text{ is positive definite}\},$$

and in restricted models in which the elements of Σ satisfy certain constraints of the following form. If $h(\Sigma) = (h_1(\Sigma),\ldots,h_r(\Sigma))$ is a vector-valued function mapping Ω into $\mathbb{R}^r$, then

$$\Omega_0 = \{\Sigma: \ \Sigma \in \Omega, \ h(\Sigma) = 0\}.$$

That is, Ω_0 is the intersection of Ω and the null space of the function h. Further assume that the functions h_i are continuous, differentiable, and homogeneous in the elements of Σ, that they are functionally independent, and that Ω_0 is not empty.

It follows immediately that Ω_0 is closed under multiplication by any positive constant. This model is essentially the one discussed in Examples 2.3 and 2.4 where it was shown to be a scale-invariant exponential family composed of homogeneous functions and it encompasses all of the special cases of Example 2.3. We note that the unconstrained parameter space Ω is an open convex conical region of $\mathbb{R}^q$ where $q = \frac{1}{2}p(p+1)$, and that Ω_0 is in general a q-r dimensional nonlinear manifold which is not convex. (If the equations $h_i(\Sigma) = 0$ are linear or equivalent to a set of linear equations then Ω_0 is linear and convex.) When necessary we shall denote the unknown true Σ by Σ_t to distinguish it from other points in Ω or Ω_0.

The constrained covariance model thus defined can be compared to the model formulated by Browne (1974), who writes Σ as a function of an unknown vector parameter γ. The two approaches are formally equivalent since γ can be taken as an arbitrary parameterization of Ω_0, but from a practical viewpoint, the model appropriate to a particular problem may be easy to write down one way and extremely difficult to formulate the other way. This point was made by Aitchison and Silvey (1960) who refer to formulations in terms of constraint equations and freedom equations. Since many of the constrained conditional covariance models studied in later sections cannot be easily written in terms of a vector γ, we shall work primarily with the constraint formulation.

Although the model described above is defined in terms of Wishart distributions, in most cases the matrix S will in fact be the sample covariance matrix of a set of n observations of a p-variate normally distributed random variable $X = (X_1, \ldots, X_p)^T$ with covariance matrix γ. Furthermore, X may represent either raw observations, or residuals after fitting some preliminary linear model, with an appropriate adjustment to n for the number of parameters fitted.

It must be stressed that to study the covariance structure of the variables X_j and assume joint normality is essentially to study only linear regression relationships among the X_j. Also, although the model as described requires all the X_j to be random, there are many situations in which some nonstochastic and design variables can be accommodated. These are cases in which the constraints h_i and any test statistics depend only on conditional covariances of the stochastic variables given subsets of the fixed variables.

Example 3.1 - Block structure

In certain biological and psychological applications it is not unreasonable to assume that a set of random variables can be partitioned into subsets such that the variables within each subset are interchangeable with regard to their joint statistical behavior (Arnold, 1973; Elston, 1975). For example, X_1, X_2, and X_3 might be measurements of some attribute on each of three brothers, and X_4, X_5 and X_6 the measurements for three sisters in families with three children of each sex. An immediate consequence of interchangeability is that the population covariance matrix of these variables has a constant-in-blocks structure, e.g.

$$\Sigma = \left[\begin{array}{ccc|ccc} a & b & b & e & e & e \\ b & a & b & e & e & e \\ b & b & a & e & e & e \\ \hline e & e & e & c & d & d \\ e & e & e & d & c & d \\ e & e & e & d & d & c \end{array}\right]$$

When sample variances are all standardazed to 1 so that covariances become correlations, this becomes the intra-class correlation model studied by Fisher (1921) and others. The simplest case of course has a single block, i.e., all correlations are equal.

The block structure model can be useful in exploratory data analysis when there are many variables available and it is desired to choose a representative subset of them for further work. One possibility is to find an appropriate partitioning of the variables and select one representative variable from each group. This will be illustrated in an example in Chapter 5. The advantages of proceeding in this way rather than taking "best" linear combinations of the original variables as representatives (as is done in factor analysis, principal component analysis, or canonical correlation analysis) include the fact that the chosen variables retain a direct meaning in terms of the original system, and that whatever further analyses are done may be repeated on other similar sets of data without having to collect and process such a large set of variables.

The two equivalent formulations of this model in terms of constraint equations and freedom equations are obvious. We note, however, that since the number of free parameters is relatively small, the number of constraints is large.

Example 3.2 - Vanishing covariances

One obvious way to reduce the effective number of covariance parameters is to require some subset of the covariances (or equivalently the correlations) to be zero. Constraints of this sort are clearly homogeneous. In a normal theory framework each vanishing covariance corresponds to the fact that a pair of variables are marginally independent. Hills (1969) has used this idea in an exploratory analysis of a set of data summarized by a 10x10 sample correlation matrix. He assumes all population correlations to be zero except those between pairs of variables whose sample correlation differs from zero by some threshold amount that is statistically significant.

Example 3.3 Covariance selection

In contrast to the last example, we mention a class of models proposed by Dempster (1972) in which elements of Σ^{-1} are assumed to vanish. Dempster also discusses data dependent procedures for selecting appropriate subsets of $\{\sigma^{ij}\}$ to set to zero. Motivation for this model derives partly from the fact that the elements of Σ^{-1} are the natural exponential parameters of the probability model, i.e., the functions $a_k(\theta)$ in the notation of Section 2.2. Hence, setting certain σ^{ij} to zero removes parameters in a natural way, and also reduces the dimension of the sufficient statistic $\underset{\sim}{b} = (s_{11}, s_{12}, \ldots, s_{pp})$. Clearly this is also a Wishart model subject to homogeneous restrictions.

Dempster does not give an interpretation in terms of independence of the variables $X_1, \ldots, X_p$, but by observing that

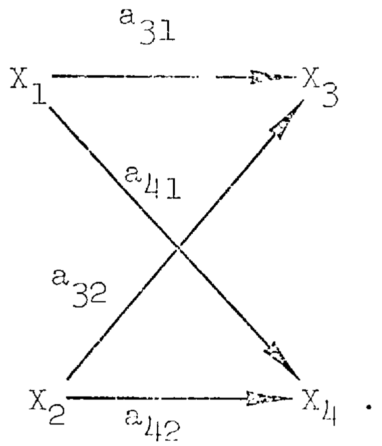
$$\sigma^{12} = \frac{-\sigma_{12.c}}{\sigma_{11.c}\sigma_{22.c} - \sigma^2_{12.c}} , \qquad (3.3)$$

where c is the index set $\{3,4,\ldots,p\}$, we see that each zero in Σ^{-1} corresponds to conditional independence of two variables given all the other variables in the model.

Although Dempster has proposed the covariance selection model partly as a data analytic tool, it suffers from one serious drawback in this regard. When studying complicated sets of data it is not unusual to have a large number of variables available, and one often considers including additional variables derived from the original set by combining and transforming them in various ways. An important initial task is to select a subset of variables to analyze, and having done this one is rarely able to say with certainty that exactly the correct set has been chosen. When dealing with covariances directly as in Example 3.2, the consequences of this selection are minimized by the fact that the marginal covariance matrix of any subset of variables is just the corresponding partition of the overall covariance matrix. But when one takes Dempster's approach and studies the inverse covariance matrix, the whole set of parameters, and in particular patterns of zeros among them, are affected by the inclusion or exclusion of additional variables. The problem becomes acute, for instance, when the sample covariance matrix is singular and thus not invertible, which can occur when the number of potentially interesting variables exceeds the number of observations, or when some variables are exact linear combinations of others. The latter situation occurs, for example, in the data example that Hills presents. A close look at his correlation matrix reveals that among his ten variables, exact linear identities hold between variables 1,2 and 3, and between variables 7,8 and 9. Clearly, two variables must be removed before Dempster's model could be applied, and considerations external to the data must be invoked to decide which variables to exclude.

Example 3.4 - Causal path model

Sewall Wright (1918; 1934) proposed the use of simple diagrams to represent the causal links among a set of random variables, in which nodes represent variables and arrows represent the causal paths. E.g.,



(3.4)

Associated with each path is a number called a path coefficient which measures the strength of causation along that path. The idea is intuitively attractive and has been discussed by several authors subsequently (e.g., Li, 1956; Turner & Stevens, 1959; J. W. Tukey, 1954; Blalock, 1968). Additional assumptions are required in order to define a complete statistical model, a sufficient set being the following:

i) All causal effects are linear,

ii) Each variable is measured from its mean,

iii) Each variable can be written as a linear combination of the variables that point toward it in the diagram, plus a random error term.

iv) The random error terms are uncorrelated and have finite variances which are constant across observations.

The case of correlated error terms can be formally accommodated by introducing additional (unobservable) variables as common causes.

With these assumptions a path diagram translates into a set of simultaneous linear regression equations which can be written, for the diagram 3.4,

$$\left.\begin{array}{lllll} a_{11}X_1 & & & & = \varepsilon_1 \\ & a_{22}X_2 & & & = \varepsilon_2 \\ a_{31}X_1 & + a_{32}X_2 & + a_{33}X_3 & & = \varepsilon_3 \\ a_{41}X_1 & + a_{42}X_2 & & + a_{44}X_4 & = \varepsilon_4 \ , \end{array}\right\} \qquad (3.5)$$

or, in matrix notation.

$$A \underset{\sim}{X} = \underset{\sim}{\varepsilon} \quad . \qquad (3.6)$$

With appropriate scaling the errors ε_k all have variance equal to 1, and the covariance matrix of $\underset{\sim}{X}$ is

$$\operatorname{cov}(\underset{\sim}{X}, \underset{\sim}{X}^T) = \Sigma = A^{-1}(A^T)^{-1},$$

$$\Sigma^{-1} = A^T A \ . \qquad (3.7)$$

The models thus defined are formally equivalent to the simultaneous linear regression systems studied in econometrics (Christ, 1966). The equations 3.5 are linear structural equations, and Eq. 3.6 is the reduced form. However the reverse mapping of structural equations to path diagrams is not unique. If prior knowledge is available to restrict the possible directions of causation (usually in the form of a partial ordering of variables in time), then ambiguity is reduced, but not necessarily eliminated.

For model selection and testing it is the missing paths in the causal diagram that are of interest, each missing path corresponding to a zero in a fixed position of A. Since the elements of A are homogeneous functions (of degree $\frac{1}{2}$) in the elements of Σ, a vanishing path corresponds to a homogeneous constraint on Σ.

In summary, then, with regard to the joint distribution of the variables $X_1, \ldots, X_p$, there is an equivalence between linear causal path diagrams, systems of simultaneous linear regression equations, and constrained covariance models in which the constraints take the form of zeros in a certain factorization of Σ. These constraints are homogeneous, so with the additional assumption that the error distributions are normal, such models are examples of the constrained covariance model defined at the beginning of this chapter.

Comparing causal path models with the covariance selection model of Example 3.3, we note that the two classes of models are for the most part distinct: zeros in Σ^{-1} do not generally produce zeros in the A^TA factorization of Σ^{-1}, nor does the reverse hold. One interesting special case where the two classes of models do overlap is the following: Suppose that among the variable $X_1,\ldots,X_p$, direct causal links connect each variable to the q variables immediately following, but to no others. Then the matrix A has non-zero entries on the major diagonal and on q subdiagonals, with zeros elsewhere, and $\Sigma^{-1} = A^TA$ is non-zero on the diagonal, on q subdiagonals, and on q superdiagonals. The representations in terms of zero patterns in A and in Σ^{-1} are completely equivalent.

Example 3.5 - Non-cyclical causal path models

An important special case of Example 3.4 occurs when the causal path diagram contains no closed cycles. In this case the variables can always be ordered into a list in which all paths of causation point downward in the list (although the ordering is usually unique only up to permutions within certain subsets of the variables). If Σ is the covariance matrix of the variables so ordered, then the matrix factor A in Eq. 3.7 is lower triangular, that is, it has zeros above the major diagonal, and is the inverse of the Cholesky decomposition of Σ.

Each missing path in a non-cyclical path model has an interpretation in terms of conditional independence, as described in the following lemma.

Lemma 3.1

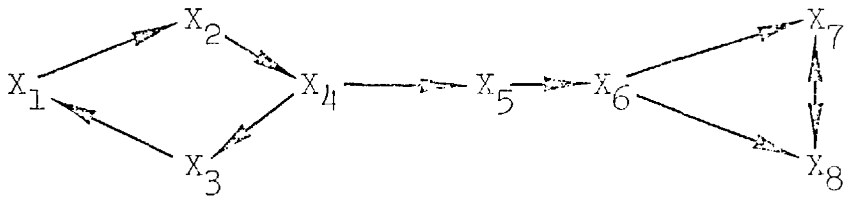
If X is a p×1 vector of random variables whose distribution is determined by a non-cyclical causal path diagram of the kind described above, and X_i are ordered as above, then a component variable X_j is conditionally uncorrelated with all predecessors which are not immediate causes, given those variables which are immediate causes.

A proof of this fairly obvious lemma can be based on the fact that an element a_{ij} of A for $i > j$ is $-a_{ii}^{-\frac{1}{2}}$ times $\beta_{ij\cdot 12\ldots i-1}$, the multiple regression coefficient of X_j when X_i is regressed on $X_1,\ldots,X_{i-1}$, (Dempster, 1969, p. 157) which in turn is equal to $\sigma_{ij\cdot 12\ldots i-1}/\sigma_{jj\cdot 12\ldots i-1}$.

It follows that a non-cyclical causal path model can be completely characterized as a constrained covariance model in which the constraints are all of the form $\sigma_{ij\cdot a} = 0$, where the conditioning set a depends on i and j.

Example 3.6 - Path models containing cycles

There has been some discussion in the literature on whether a set of random variables representing measurements of an actual physical system can ever logically be be said to have a closed loop of causation (Turner & Stevens, 1959; Wright, 1960). Nevertheless such models do exist formally as sets of simultaneous regression equations, even though their interpretation as causal systems requires some care. Consider the following example:



(3.8)

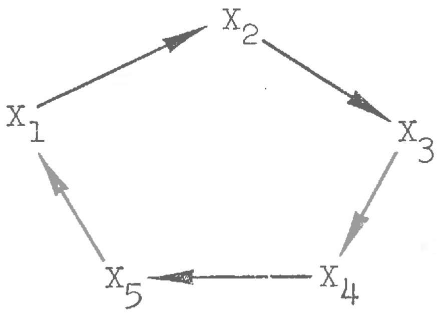
It is fairly easy to show that Lemma 3.1 extends to any variable such as X_5 which is not itself involved in a closed loop. For the purpose of the lemma, the variables need only be ordered in a way that puts those variables which are directly or indirectly affected by X_5 beyond X_5 in the list. For instance, either of the following orderings would suffice:

$$(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

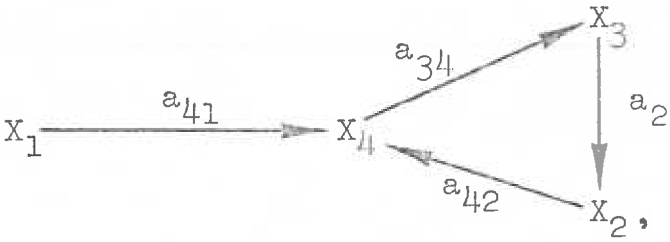
or

$$(X_2, X_3, X_4, X_1, X_5, X_8, X_7, X_6).$$

The principle of conditioning on variables that are immediate causes in order to produce conditional independence applies in some cases to variables involved in closed loops. . It works for instance when there is a single loop:



Here, $\sigma_{13.25} = 0$, $\sigma_{24.31} = 0$, ..., $\sigma_{52.14} = 0$. However it fails in



where the only conditional covariance that vanishes identically is $\sigma_{13.24}$ (which means, incidentally, that $\sigma^{13} = 0$). In this second example,

$$\sigma_{13.4} = a_{23}a_{42}a_{41}/(1+a_{41}^2+a_{42}^2+a_{42}^2a_{23}^2) ,$$

which is not zero unless either X_1 is independent of the rest ($a_{41} = 0$), or the diagram degenerates into one without a closed cycle ($a_{23} = 0$ or $a_{42} = 0$).

These examples serve to show that despite a close connection between conditional independence and linear causal path models, if closed cycles of causation are allowed then the effective covariance constraints cannot necessarily be written as vanishing conditional covariances.

Example 3.7 - Conditional covariance constraints

In one sense Examples 3.2 and 3.3 are at opposite extremes as special cases of a more general Wishart model which is constrained by having one or more conditional covariances vanish: in Example 3.2 the conditioning sets are empty and hence the smallest possible, whereas in Example 3.3 each contains all the other variables. The non-cyclical path models (and some with cycles) are intermediate between these two extremes. The general model, which is discussed in later sections, can be formulated as a Wishart model whose matrix parameter Σ satisfies a set of r homogeneous constraints,

$$h_k(\Sigma) = \sigma_{i_k j_k \cdot a_k} = 0, \qquad (k=1, \ldots, r),$$

where i_k is not equal to j_k, the pairs (i_k, j_k) are distinct for different k, and the conditioning sets may be empty or may overlap in any way.

It is worth noting that the specification of such a model in terms of conditional covariance constraints is not always unique. For instance if there are four variables with two constraints given by

$$\sigma_{12.4} = 0, \quad \text{and} \quad \sigma_{13.4} = 0,$$

then the first constraint can equivalently be written as

$$\sigma_{12.34} = 0,$$

which follows from the identity,

$$\sigma_{12.34} = \sigma_{12.4} - \sigma_{13.4}\sigma_{23.4}/\sigma_{33.4} \quad .$$

Then by symmetry, the second constraint can be replaced by

$$\sigma_{13.24} = 0.$$

3.3 Asymptotic covariances of functions of S

Subsequent chapters consider methods to test the adequacy of fit of various constrained Wishart models to observed data, the general approach being first to compute the constraint functions from the sample covariance matrix, then to decide whether the values obtained can be explained by statistical variation alone (Silvey, 1970, Section 7.3). Some knowledge of the joint distribution of the constraint functions under the model is required, and approximations to these distributions will be used, based on asymptotic moments for which formulas are derived here.

When calculating moments of functions of the elements of a sample covariance matrix S, the symmetry of S leads to certain complications. To simplify the formulas we introduce a special symmetric derivative notation: If f is a function of the elements of S and $\frac{df}{dS}$ is the matrix $(\frac{\partial f}{\partial s_{ij}})$ without taking the symmetry of S into account, then define a matrix derivative with a double bar as

$$\frac{df}{\overline{dS}} = \left(\frac{\partial f}{\overline{\partial s_{ij}}}\right) = \frac{1}{2}\left(\frac{df}{dS} + \frac{df}{dS^T}\right) . \qquad (3.9)$$

Three obvious points to note are, first, that $\frac{df}{\overline{dS}}$ is symmetric; second, that if f is symmetrically defined with respect to S and S^T (e.g., $f = \frac{1}{2}(s_{13}+s_{31})$), then $\frac{df}{dS} = \frac{df}{\overline{dS}}$; and third, that the total derivative of f with respect to s_{12}, say, with symmetry considered is $\frac{\partial f}{\overline{\partial s_{12}}} + \frac{\partial f}{\overline{\partial s_{21}}}$ or equivalently $2 \frac{\partial f}{\overline{\partial s_{12}}}$, and not simply $\frac{\partial f}{\overline{\partial s_{12}}}$.

With this notation the required asymptotic variances and covarinaces are given by Lemma 3.2 below.

Lemma 3.2

If f and g are continuous real-valued functions of the elements of S with first and second derivatives in a neighborhood of Σ, and S is distributed as $W(n,\Sigma)$, then to terms of order n^{-1} the covariance of f(S) and g(S) is

$$\text{cov}(f,g) = \frac{2}{n} \text{ tr } \left(\frac{df}{d\Sigma} \Sigma \frac{dg}{d\Sigma} \Sigma \right), \qquad (3.10)$$

where $\frac{df}{d\Sigma}$ is $\frac{df}{dS}$ evaluated at $S = \Sigma$. Alternatively,

$$\text{cov}(f,g) = \frac{2}{n} f(\Sigma)g(\Sigma) \text{ tr } \left(\frac{d \log f}{d\Sigma} \Sigma \frac{d \log g}{d\Sigma} \Sigma \right), \qquad (3.11)$$

and

$$\text{cov}(f,g) = - \frac{2}{n} \text{ tr } \left(\frac{df}{d\Sigma} \frac{df}{d\Sigma^{-1}} \right). \qquad (3.12)$$

All of these formulas are valid when $f \equiv g$, giving the asymptotic variance of f in that case.

Equation 3.11 is essentially like a formula given by Siotani (1968), who uses a different device to deal with the symmetry of S.

We indicate briefly a derivation of these formulas based on the notation of Browne (1974). If S is a pxp symmetric matrix, let $\text{vec}(S) = \underline{s}$ be a column vector formed from all the elements of S, and let $\underset{\sim}{s}$ be a column vector of the elements of S on and below the major diagonal. Let K_p^T be the matrix of 1's, 0's and ½'s that maps $\underline{s}$ into $\underset{\sim}{s}$ (by averaging elements in symmetrically opposite positions), so that

$$\underset{\sim}{s} = K_p^T \underline{s} \quad ,$$

and

$$\underline{s} = K_p^{-T} \underset{\sim}{s} \quad , \qquad (3.13)$$

where $K_p^- = (K_p^T K_p)^{-1} K_p^T$ is a left generalized inverse of K_p. Finally, let M_p be the symmetric idempotent matrix

$$M_p = K_p K_p^- = K_p (K_p^T K_p)^{-1} K_p^T \ .$$

It follows from Eq. 3.13 and Eq. A10 in the Appendix, that

$$\frac{df}{d\underset{\sim}{s}} = K_p^- \frac{df}{d\underline{s}} \quad .$$

Now the familiar covariance formula for pairs of elements from a Wishart matrix,

$$\mathrm{cov}(s_{ij}, s_{kl}) = \frac{1}{n} \ (\sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}),$$

can be written (as shown by Browne) as

$$\mathrm{cov}(\underset{\sim}{s},\underset{\sim}{s}^T) = \frac{2}{n} K_p^T \,(\Sigma\otimes\Sigma)\, K_p,$$

where $\otimes$ is the Kronecker outer product. It follows that

$$\mathrm{cov}(\underline{s},\underline{s}^T) = \frac{2}{n} M_p(\Sigma\otimes\Sigma)\, M_p.$$

Starting from the usual asymptotic covariance formula for two functions f and g of $\underset{\sim}{s}$, we have

$$\begin{aligned}
\mathrm{cov}(f,g) &= \frac{df}{d\underset{\sim}{\sigma}^T}\ \mathrm{cov}\ (\underset{\sim}{s},\underset{\sim}{s}^T)\ \frac{dg}{d\underset{\sim}{\sigma}} \\
&= \left\{ \frac{df}{d\underline{\sigma}^T} K_p^{-T} \right\} \left\{ \frac{2}{n} K_p^T\ (\Sigma\otimes\Sigma) K_p \right\} \left\{ K_p^- \frac{dg}{d\underline{\sigma}} \right\} \\
&= \frac{2}{n}\ \frac{df}{d\underline{\sigma}^T} M_p\ (\Sigma\otimes\Sigma)\ M_p \frac{dg}{d\underline{\sigma}}\ . \qquad (3.14)
\end{aligned}$$

The matrix M_p has the property $M_p\mathrm{vec}(X) = \mathrm{vec}\{\frac{1}{2}(X+X^T)\}$ for any square matrix X, whence

$$M_p \frac{dg}{d\underline{\sigma}} = \mathrm{vec} \left(\frac{dg}{d\Sigma}\right) . \qquad (3.15)$$

Also, using the following formula from Browne,

$$\mathrm{vec}(X)^T(A \otimes B)\mathrm{vec}(Y) = \mathrm{tr}\ (XAY^TB), \qquad (3.16)$$

which holds for any square matrices A, B, X and Y, one can derive Eq. 3.10 from Eqs. 3.14 and 3.14. Equation 3.11 follows as an immediate consequence, and Eq. 3.12 follows from Appendix formula A5.

3.4 Minimal sufficiency and ancillarity

It is interesting to consider the extent to which certain general principles of inference can be applied to the class of structured Wishart models defined in this chapter. Very little can be said concerning the most general formulation, so we confine our attention in this and the following section to the class of models formulated in Ex. 3.7 whose structure is defined by the vanishing of r conditional covariances.

It was shown in Ex. 2.3 that the Wishart distribution belongs to an exponential family whose natural parameters are the $q = \frac{1}{2}p(p+1)$ distinct elements σ^{ij} of Σ^{-1}. The distinct elements of S form a minimal set of sufficient statistics for a full (unconstrained) model. For a structured model where Σ is constrained to belong to a subset Ω_0 of Ω, let $\Omega_0^* = \{\Sigma^{-1}: \Sigma \in \Omega_0\}$ denote the restricted natural parameter space. So long as Ω_0^* does not lie in an affine space of dimension less than q, S remains the minimal sufficient statistic. It is clear from Eq. 3.3 that if one of the constraints is

$$\sigma_{ij.c} = 0, \qquad c = \{1,2,\ldots,p\} - \{i,j\}, \qquad (3.17)$$

then Ω_0^* is a subset of the linear subspace $\{\Sigma^{-1}: \sigma^{ij} = 0\}$, so that a constraint like 3.17 removes σ^{ij} from the set of natural parameters, and removes s_{ij} from the set of minimal sufficient statistics. There are other situations as shown in Ex. 3.7 where several constraints, none of which is explicitly of the form of Eq. 3.17, still force an element of Σ^{-1} to be zero, but these are the only circumstances under which Ω_0^* lies in an affine subspace.

Except for the special case where all r constraints are equivalent to the vanishing of r elements of Σ^{-1} (which is the "covariance selection" model described in Ex. 3.2), the dimension of the sufficient statistic will exceed the effective dimension of the parameter space, and it will follow that S is not a complete statistic. This can be seen directly by observing that if $\sigma_{ij.b} = 0$ is one of the constraints, then $s_{ij.b}$ being an unbiased estimate of $\sigma_{ij.b}$ is a function of S with zero expectation for every $\Sigma \in \Omega_0$.

In order to test proposed models against observed data, it is desirable to find ancillary statistics, i.e., statistics whose distributions under the model assumptions are completely known, and which are functions of the minimal sufficient statistic. Each single constraint $\sigma_{ij.b} = 0$ in the present model produces an ancillary statistic, namely the sample partial correlation coefficient,

$$r_{ij.b} = \frac{s_{ij.b}}{\sqrt{s_{ii.b}\, s_{jj.b}}} \quad .$$

That $r_{ij.b}$ has a marginal distribution not depending on $\Sigma \in \Omega_0$ follows from Fisher's (1924) result that if $F(r \mid n, \rho_{ij})$ is the distribution function of r_{ij} based on a sample of size n, then the distribution function of $r_{ij.b}$ is $F(r \mid n-q_b, \rho_{ij.b})$, where q_b is the number of indices in b.

Although there are r statistics that are individually ancillary when Ω_0 is defined by r constraints, they are not jointly ancillary in general. For example, if $p = 4$ and $\Omega_0 = \{\Sigma\colon \sigma_{12} = \sigma_{34} = 0\}$, then the asymptotic covariance of r_{12} and r_{34} is (Elston, 1975),

$$\operatorname{cov}(r_{12}, r_{34}) \doteq \frac{1}{n}(\rho_{13}\rho_{24} + \rho_{14}\rho_{23}),$$

which depends on other unknown parameters in the model. Since the $r_{ij.b}$ are not jointly ancillary, they cannot be used to form an exact test of the constrained model against the full model. Nevertheless, approximate tests can be based on them, since their joint moments can be estimated. This will be pursued further in Chapter 5. It does not seem possible to construct a set of r jointly ancillary statistics.

3.5 Invariance

Continuing with the Wishart model characterized by vanishing conditional covariances, we now apply invariance arguments to support the use of $r_{ij.b}$ as a test statistic, but we will show that, when there are several constraints, this line of argument also encounters some difficulties.

The group of linear scale transformations along individual coordinates in $\underset{\sim}{X}$-space leaves both the population and sample correlation matrices unchanged. Since $r_{ij.b}$ and $\rho_{ij.b}$ are functions only of the sample and population correlation matrices, respectively, and since $\sigma_{ij.b}$ is zero if and only if $\rho_{ij.b}$ is zero, it follows that Ω_0 and its complement $\Omega_A = \Omega - \Omega_0$ are invariant under the group of scale transformations, and that $\{r_{i_k j_k \cdot b_k}\}$ form a set of invariant statistics.

Ideally, we should like $\{r_{i_k j_k \cdot b_k}\}$ to be a maximal invariant set of statistics, but before examining this point we must identify the largest group of transformations under which the testing problem is invariant (Cox & Hinkley, 1974, Section 5.3). Suppose for the moment there is only one constraint, $\sigma_{12.b} = 0$. Partition the indices into three sets, a = {1,2}, b, and c = {complement of $a \cup b$}. Then define a group $\mathcal{G}$ of transformations on the sample space by

$$S \longrightarrow GSG^T,$$

where G is any non-singular matrix of the form

$$G = \left[\begin{array}{cc|c|c} g_{11} & 0 & & \\ & & G_{ab} & 0 \\ 0 & g_{22} & & \\ \hline \multicolumn{2}{c|}{0} & G_{bb} & 0 \\ \hline \multicolumn{2}{c|}{G_{ca}} & G_{cb} & G_{cc} \end{array}\right] . \qquad (3.18)$$

It is easily verified that $\mathcal{G}$ is a group, and the induced group of transformation on the parameter space is essentially the same,

$$\Sigma \longrightarrow G \, \Sigma \, G^T .$$

It can be shown directly that $r_{12.b}$ and $\rho_{12.b}$ are invariant under these transformations; furthermore, $\mathcal{G}$ is the largest group of linear transformations that leaves $r_{12.b}$ and $\rho_{12.b}$ unchanged.

Under this group $\mathcal{G}$ of transformations, $r_{12.b}$ is a maximal invariant statistic, which means that if S and S* are any two sample matrices such that $r_{12.b} = r^*_{12.b}$, then there is a transformation in $\mathcal{G}$ that maps S into S*. To establish this it is enough to show that there exists a $G \in \mathcal{G}$ that maps S into

$$\left[\begin{array}{cc|c|c} 1 & r_{12.b} & & \\ r_{12.b} & 1 & 0 & 0 \\ \hline \multicolumn{2}{c|}{0} & I & 0 \\ \hline \multicolumn{2}{c|}{0} & 0 & I \end{array}\right] . \qquad (3.19)$$

The required values in the upper-left four partitions of G are obtained by taking

$$G_{aa} = \begin{bmatrix} s_{11.\mathbf{b}}^{-\frac{1}{2}} & 0 \\ 0 & s_{22.\mathbf{b}}^{-\frac{1}{2}} \end{bmatrix} ,$$

$$G_{bb} = S_{bb}^{-\frac{1}{2}} ,$$

$$G_{ab} = - G_{aa}S_{ab}S_{bb}^{-1} ,$$

and since there are no restrictions on the space into which the coordinates c are mapped, suitable choices of G_{ca}, G_{cb} and G_{cc} can always be made to obtain the other partitions of G. It follows that $r_{12.b}$ is maximal invariant.

Nevertheless, as with ancillarity, the principle of maximal invariance runs into difficulties for the general model under consideration because of complications that arise when there are more constraints. The appropriate group of transformations should have the structure 3.18 with regard to all constraints simultaneously. For instance, if $\sigma_{12.ce} = \sigma_{34.de} = 0$, then we should define $\mathcal{G}$ by non-singular matrices of the form

$$G = \left[\begin{array}{cc|cc|c|c|c}
g_{11} & 0 & & & & & \\
 & & \multicolumn{2}{c|}{0} & G_{ac} & 0 & G_{ac} \\
0 & g_{22} & & & & & \\
\hline
 & & g_{33} & 0 & & & \\
\multicolumn{2}{c|}{0} & & & 0 & G_{bd} & G_{bc} \\
 & & 0 & g_{44} & & & \\
\hline
\multicolumn{2}{c|}{0} & \multicolumn{2}{c|}{0} & G_{cc} & 0 & G_{cc} \\
\hline
\multicolumn{2}{c|}{0} & \multicolumn{2}{c|}{0} & 0 & G_{dd} & G_{de} \\
\hline
\multicolumn{2}{c|}{0} & \multicolumn{2}{c|}{0} & 0 & 0 & G_{ee}
\end{array}\right] .$$

When we try to demonstrate maximal invariance by transforming S into a form analogous to Eq. 3.19, we are able to choose the various partitions of G as functions of S to obtain

$$GSG^T = \left[\begin{array}{cc|cc|c|c|c}
1 & r_{12.ce} & \multicolumn{2}{c|}{?} & 0 & ? & 0 \\
r_{12.ce} & 1 & & & & & \\
\hline
\multicolumn{2}{c|}{?} & 1 & r_{34.dc} & ? & 0 & 0 \\
 & & r_{34.de} & 1 & & & \\
\hline
\multicolumn{2}{c|}{0} & \multicolumn{2}{c|}{?} & I & ? & 0 \\
\hline
\multicolumn{2}{c|}{?} & \multicolumn{2}{c|}{0} & ? & I & 0 \\
\hline
\multicolumn{2}{c|}{0} & \multicolumn{2}{c|}{0} & 0 & 0 & ?
\end{array}\right] ,$$

but the elements in the partitions labeled "?" remain functions of S, and there is not in general any other transformation which completely removes this dependence. It follows that although two sample matrices S and S* might lead to the same values for $r_{12.ce}$ and $r_{34.de}$, there is in general no transformation in $\mathcal{G}$ that maps S into S*, so the set $\{r_{ij.b}\}_k$ is not maximal invariant.

We note a close connection here with the failure of joint ancillarity. For if $\underset{\sim}{t} = \{r_{i_k j_k . b_k}\}$ were a maximal invariant statistic, then $\underset{\sim}{\tau} = \{\rho_{i_k j_k . b_k}\}$ would be maximal invariant for the parameter Σ, and the distribution of $\underset{\sim}{t}$ would depend only on $\underset{\sim}{\tau}$. But under the model assumptions, $\underset{\sim}{\tau} = (0,0,\ldots,0)$, so the distribution of $\underset{\sim}{t}$ would not depend on any of the remaining parameters, and $\underset{\sim}{t}$ would be ancillary.

Example 3.8 Covariance selection (cont'd)

Returning to Dempster's model discussed in Ex.3.3 in which elements of Σ^{-1} are forced to zero, we recall that a constraint like $\sigma^{12} = 0$ is equivalent to $\sigma_{12.c} = 0$ where c is the complement of $\{1,2\}$ in $\{1,2,\ldots,p\}$. The invariant ancillary statistic corresponding to this constraint is therefore $r_{12.c}$. The partial correlation coefficients for several constraints of this kind can be obtained easily from S^{-1} using the following lemma.

Lemma 3.3

If R* is a correlation matrix formally computed from S^{-1}, then each off-diagonal element of R* is the negative of the conditional correlation coefficient of the corresponding pair of variables, given all the other variables, i.e.,

$$r^*_{ij} = \frac{s^{ij}}{\sqrt{s^{ii}\ s^{jj}}} = -\ r_{ij.c}, \qquad c = \{1,\ldots,p\} - \{i,j\}.$$

Proof

For a 2×2 covariance matrix, the result follows immediately. Now let $a = \{i,j\}$ and c = complement of a. Then $r_{ij.c}$ is the correlation coefficient computed from the 2×2 matrix $S_{aa.c}$. Letting S^{aa} represent the (a,a) partition of S^{-1}, we have $S^{-1}_{aa.c} = S^{aa}$, and the lemma follows by applying the 2x2 result to $S^{-1}_{aa.c}$.

We note that if R is the usual correlation matrix, then R* is not R^{-1}. However, R* can be calculated from R^{-1} by formally converting R^{-1} to a correlation matrix. i.e., by multiplying on the right and left by the diagonal matrix $\text{diag}(r^{ii})^{-\frac{1}{2}}$.

Chapter 4

MAXIMUM LIKELIHOOD ESTIMATION FOR CONSTRAINED WISHART MODELS

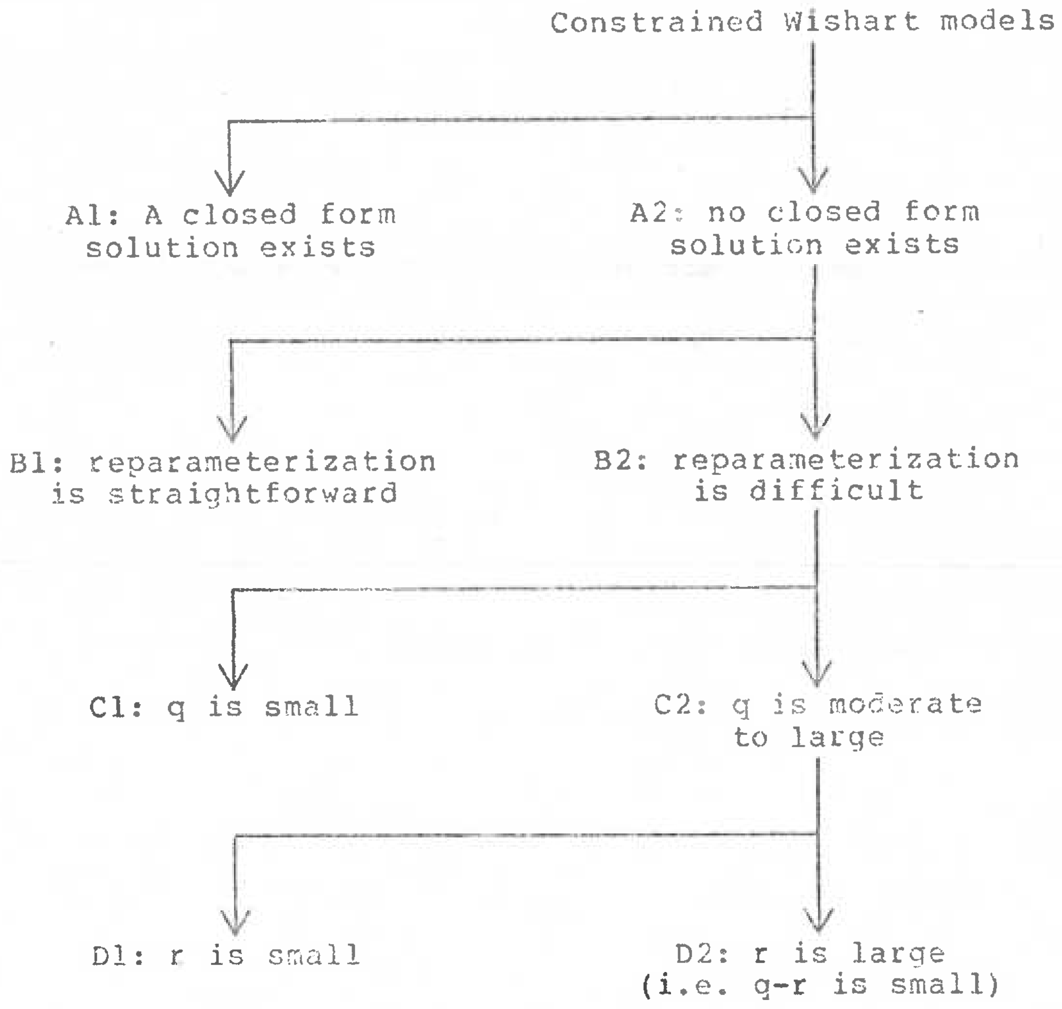
4.1 Introduction

Having defined a general class of patterned covariance models, one faces the problem of calculating constrained covariance estimates. If one's purpose is to explore various pattern hypotheses concerning Σ, it would be convenient to have a simple general method to do the fitting and avoid the need to develop specialized computer programs for each model, but the large number of parameters involved in many problems of practical interest makes it impossible to meet this objective.

When maximum likelihood is used for estimation, models, can be classified as in Fig. 4.1 according to the complexity of calculation. In category A1, direct calculation of maximum likelihood estimates is most efficient, although in some cases the closed-form solution may be difficult to obtain. Examples include some linear covariance structures (Anderson, 1969), and non-cyclical path models.

For many problems a closed-form solution cannot be obtained, but it may be possible to re-express the constrained covariances as continuous functions of a smaller set of unconstrained parameters. This defines category B1. Categories B1 and B2 involve the freedom-equation and constraint-equation representations discussed by Aitchison and Silvey (1960). Since

FIGURE 4.1 - Models classified by difficulty of fitting


Constrained Wishart models
A1: A closed form solution exists
A2: no closed form solution exists
B1: reparameterization is straightforward
B2: reparameterization is difficult
C1: q is small
C2: q is moderate to large
D1: r is small
D2: r is large (i.e. q-r is small)

each equality constraint formally reduces the dimension of the parameter space by 1, the number of parameters in principle is q-r (q is the number of variances and covariances and r is the number of constraints), and an explicit representation of the reduced parameter space can lead to considerable economy in numerical computation. Examples include Jöreskog's generalized factor analysis model (1970), general linear path models, block structure models, and the rest of Anderson's linear covariance models.

Computational methods based on the constraint formulation for models that are not easily reparameterized (category B2), are inherently more difficult. Aitchison and Silvey (1958; 1960) describe a general approach using Lagrange multipliers, which we adapt in Section 4.2 to the Wishart problem. This approach seems suitable for small problems (category C1), but the fact that it enlarges rather than reduces the working parameter set is a severe disadvantage for large problems (category C2).

Many problems in category D1, where the number of constraints is small, can be treated by an alternative approach developed in Section 4.3 which is based on a fix-point iteration and a series of r-dimensional optimizations.

The final category D2, remains difficult to treat in general, but models in it can often be approximated by certain reparameterized models in category B1.

4.2 Maximum Likelihood By Direct Iteration

This section adapts the constrained maximum Wishart likelihood problem to solution by a general least-squares optimization program.

There are in fact two kinds of constraints to consider: the explicit equality constraints defined by $\underset{\sim}{h}(\Sigma) = (0,\ldots,0)^T$, and an implicit set of inequality constraints that require Σ to be positive definite. To deal with the former we use a set of Lagrangian terms, and for the latter we shall reparametrize the problem in terms of the Cholesky lower triangular decomposition of Σ^{-1}, i.e., $\Sigma^{-1} = A^TA$. The $q = \frac{1}{2}p(p+1)$ non-zero elements of A form an unrestricted parametrization of the set of positive definite matrices Σ.

The unknown true A will be estimated by a stationary point of the function

$$\Phi(A,\underset{\sim}{\lambda}) = \tfrac{1}{2}\,\mathrm{tr}(A^TAS) - \tfrac{1}{2}\log|A^TA| - \underset{\sim}{\lambda}^T\underset{\sim}{h}(A), \qquad (4.1)$$

which, apart from terms not involving Σ, is $-\frac{1}{n}$ times the log-likelihood plus a Lagrangian term for each constraint. Here $\underset{\sim}{\lambda}$ is a column vector of Lagrangian multipliers, $\underset{\sim}{a}$ is a column vector composed of the elements on and below the diagonal of A, and $\underset{\sim}{\theta}$ is the combined vector $(\underset{\sim}{a}^T,\underset{\sim}{\lambda}^T)^T$.

A stationary point of Φ is a point where the vector $\underset{\sim}{\varphi}$ of partial derivatives vanishes, that is,

$$\underset{\sim}{\varphi}(\underset{\sim}{\theta}) = \frac{d\Phi(\underset{\sim}{\theta})}{d\underset{\sim}{\theta}} = \left[\begin{array}{c} \underset{\sim}{\psi} \\ \hline -\underset{\sim}{h} \end{array}\right] \begin{array}{l} 1 \\ \vdots \\ q \\ q+1 \\ \vdots \\ q+r \end{array} = 0 \; . \qquad (4.2)$$

Here $\underset{\sim}{\psi}$ is defined to be $\frac{d\Phi}{d\underset{\sim}{a}}$.

This set of simultaneous non-linear equations can be solved by treating the components of $\underset{\sim}{\varphi}$ as residuals and minimizing the sum of their squares, the minimum being attained when each component vanishes. A computer program for this based on the Gauss-Newton algorithm requires evaluation of both the residual vector $\underset{\sim}{\varphi}$, and the matrix G of its derivatives with respect to the parameters $\underset{\sim}{\theta}$, which can be partitioned as,

$$G = \frac{d\underset{\sim}{\varphi}}{d\underset{\sim}{\theta}^T} = \left[\begin{array}{c:c} \dfrac{d\underset{\sim}{\psi}}{d\underset{\sim}{a}^T} & \dfrac{d\underset{\sim}{\psi}}{d\underset{\sim}{\lambda}^T} \\ \hdashline -\dfrac{d\underset{\sim}{h}}{d\underset{\sim}{a}^T} & -\dfrac{d\underset{\sim}{h}}{d\underset{\sim}{\lambda}^T} \end{array}\right]$$

$$= \left[\begin{array}{c:c} \Gamma & -H \\ \hdashline -H^T & 0 \end{array}\right] \; . \qquad (4.3)$$

Here,

$$H = \frac{d\underset{\sim}{h}^T}{d\underset{\sim}{a}} \quad \text{and} \quad \Gamma = \frac{d^2\Phi}{d\underset{\sim}{a}\,d\underset{\sim}{a}^T} . \qquad (4.4)$$

To compute these quantities, we return to Eq. 4.1 and write Φ as $\Phi_W + \Phi_\lambda$, where $\Phi_W = \frac{1}{2}\,\mathrm{tr}(A^TAS) - \frac{1}{2}\log|A^TA|$ is the contribution from the Wishart likelihood, and $\Phi_\lambda = -\underset{\sim}{\lambda}^T\underset{\sim}{h}(A)$ represents the sum of the Lagrangian terms. Similarly, $\underset{\sim}{\psi} = \underset{\sim}{\psi}_W + \underset{\sim}{\psi}_\lambda$ and $\Gamma = \Gamma_W + \Gamma_\lambda$. For the likelihood contributions to ψ and Γ we have

$$\underset{\sim}{\psi}_W = \frac{d\Phi_W}{d\underset{\sim}{a}}$$

which is a column vector formed from the lower-triangle elements of

$$\frac{d\Phi_W}{dA} = AS - \mathrm{diag}\left(\frac{1}{a_{ii}}\right). \qquad (4.5)$$

This expression follows from Eq. A8 and the fact that $|A^TA| = \Pi_{i=1}^{p}\, a_{ii}^2$, since A is triangular. A general element of Γ_W is given by

$$\frac{\partial^2 \Phi_W}{\partial a_{ij} \partial a_{k\ell}} = \delta_{ik} s_{\ell j} + \delta_{ij} \delta_{jk} \delta_{k\ell} / a_{ii}^2 \ . \qquad (4.6)$$

where δ_{ij} is the Kronecker delta function.

Note the particularly simple form of these expressions, especially Eq. 4.6. The second derivative matrix of the Wishart likelihood, expressed in terms of Cholesky parameters, is nearly constant; only p diagonal elements depend on $\underset{\sim}{a}$, and out of the $q^2 = \{\frac{1}{2}p(p+1)\}^2$ elements, only 3q are non-zero.

For the Lagrangian contributions to ψ and Γ we have

$$\underset{\sim}{\psi}_\lambda = \frac{d}{d\underset{\sim}{a}} (-\underset{\sim}{h}^T \underset{\sim}{\lambda}) = -H\underset{\sim}{\lambda} \ , \qquad (4.7)$$

where H is defined by Eq. 4.4 , and

$$\Gamma_\lambda = \frac{d\underset{\sim}{\psi}_\lambda}{d\underset{\sim}{a}^T} = - \sum_{k=1}^{r} \lambda_k \frac{d^2 h_k}{d\underset{\sim}{a} d\underset{\sim}{a}^T} \ . \qquad (4.8)$$

There remains only the calculation of the first and second derivatives of the constraint functions h_k with respect to the elements of $\underset{\sim}{a}$, which are required for H and Γ_λ in Eqs. 4.3 , 4.7 and 4.8 . They must be calculated for each particular model. If the functions $h_k(\Sigma)$ can be expressed easily in terms of the a_{ij}, then this is straightforward, but in other cases the following formulas are helpful:

$$\frac{dh}{d\Sigma^{-1}} = - \Sigma \frac{dh}{d\Sigma} \Sigma , \tag{4.9}$$

$$\frac{dh}{dA} = 2 \cdot \text{lower triangle } \{ A \frac{dh}{d\Sigma^{-1}} \}, \tag{4.10}$$

and

$$\frac{\partial^2 h}{\partial a_{ij} \partial a_{\ell m}} = 2 \, \delta_{i\ell} \frac{\partial h}{\partial o^{mj}} + 4 \sum_{k=1}^{i} \sum_{n=1}^{\ell} a_{ik} a_{\ell n} \frac{\partial^2 h}{\partial \sigma^{kj} \partial^{nm}} . \tag{4.11}$$

The practical limitation of this computational method now becomes apparent, for at each iteration the (q+r) × (q+r) matrix G must be computed and the least-squares program must essentially solve the set of linear equations $G\underset{\sim}{X} = - \underset{\sim}{\varphi}$.

4.3 Constrained Estimation By Fix-Point Iteration

Another approach to the calculation of constrained estimates, which is suitable especially when the number of constraints is small, is based on a certain partially linear structure that exists in contrained Wishart models. This method can be used with several different estimation criteria and does not require the calculation of second derivatives of the constraint functions.

Consider the following four estimation criteria from among those discussed in Section 2.4, ML: maximum Wishart likelihood, LS: simple least squares applied to the elements of $\hat{\Sigma}$-S, LS*: least squares applied to the elements of $(\hat{\Sigma}^{-1}-S^{-1})$, and GLS: generalized least squares which weights the squares of the elements of $\hat{\Sigma}$-S by the inverse of their covariance matrix. In each case, one seeks to optimize the criterion within the class $\Omega_0 = \{\Sigma \in \Omega\colon \underset{\sim}{h}(\Sigma) = 0\}$. Writing Lagrangian terms for the constraints, one wants a stationary point of one of the following expressions:

$$\text{ML:}\quad \log|\Sigma^{-1}| - \operatorname{tr}(\Sigma^{-1}S) + \sum_k \lambda_k h_k(\Sigma)$$

$$\text{LS:}\quad \tfrac{1}{2}\operatorname{tr}(\Sigma-S)^2 + \sum_k \lambda_k h_k(\Sigma)$$

$$\text{LS*:}\quad \tfrac{1}{2}\operatorname{tr}(\Sigma^{-1}-S^{-1})^2 + \sum_k \lambda_k h_k(\Sigma)$$

$$\text{GLS:}\quad \tfrac{1}{2}\operatorname{tr}\{\Sigma^{-1}(\Sigma-S)\}^2 + \sum_k \lambda_k h_k(\Sigma)\ .$$

Equating the derivatives of these expressions to zero produces

$$\left.\begin{aligned} \text{ML:}\quad & \Sigma = S - \sum \lambda_k G_k^*(\Sigma), \\ \text{LS:}\quad & \Sigma = S - \sum \lambda_k G_k(\Sigma), \\ \text{LS*:}\quad & \Sigma^{-1} = S^{-1} - \sum \lambda_k G_k^*(\Sigma), \\ \text{GLS:}\quad & \Sigma^{-1} = S^{-1} - \sum \lambda_k (S^{-1} G_k^*(\Sigma) S^{-1}), \end{aligned}\right\} \qquad (4.12)$$

where

$$G_k(\Sigma) = \frac{dh_k(\Sigma)}{d\Sigma}, \quad (k=1,\ldots,r),$$

and

$$G_k^*(\Sigma) = \frac{dh_k(\Sigma)}{d\Sigma^{-1}} = -\,\Sigma \cdot G_k(\Sigma) \cdot \Sigma, \quad (k=1,\ldots,r).$$

We concentrate on the ML formulation, although it is clear from Eq. 4.12 that similar arguments could be applied to the other formulations. For fixed $\bar{\Sigma}$, let $\mathcal{L}(\bar{\Sigma})$ denote the linear subspace of $\mathbb{R}^{p^2}$ spanned by the matrices $G_k^*(\bar{\Sigma})$, $(k=1,\ldots,r)$, where $p \times p$ matrices represent points in $\mathbb{R}^{p^2}$.

Then Eq. 4.12(ML) says that if $\hat{\Sigma}$ is a maximum likelihood solution to the constrained problem, then S is in the linear manifold $\hat{\Sigma} + \mathcal{L}(\hat{\Sigma})$. Furthermore if $\mathbb{C}$ represents the set of all symmetric matrices that satisfy the constraints, then $\hat{\Sigma}$ is the linear projection of S into $\mathbb{C}$ along $\mathcal{L}(\hat{\Sigma})$. These facts suggest a fix-point iterative calculation:

1. First take $\Sigma^{(0)} = S$ to be an initial approximation to $\hat{\Sigma}$.
2. Denoting by $\Sigma^{(m)}$ the current approximation to $\hat{\Sigma}$, compute the matrices $G_k^*(\Sigma^{(m)})$ which define $\mathcal{L}(\Sigma^{(m)})$.
3. Project S into $\mathbb{C}$ along $\mathcal{L}(\Sigma^{(m)})$ to obtain a new estimate $\Sigma^{(m+1)}$, as in Fig. 4.2, and return to step 2.

If this process converges, then a solution to Eq. 4.12(ML) has been found.

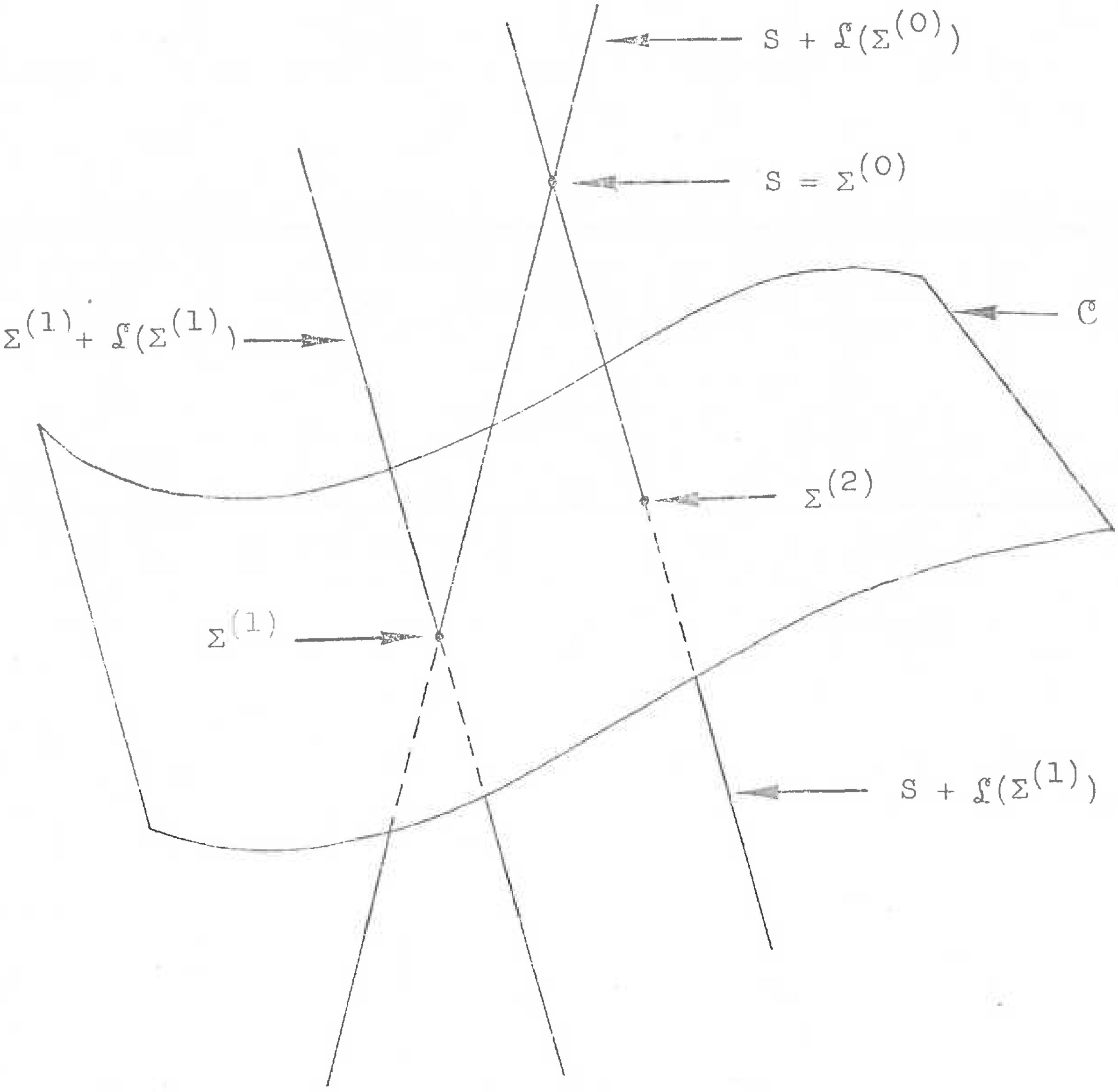
The crucial step is the projection of S into $\mathbb{C}$ along $\mathcal{L}(\Sigma^{(m)})$, which in itself requires solving a set of r non-linear equations,

$$h_k(\Sigma_\lambda) = 0, \quad (k=1,\ldots,r) \qquad (4.13)$$

in r variables, $\lambda_1,\ldots,\lambda_r$, where

$$\Sigma_\lambda = S-\lambda_1 G_1^* - \ldots - \lambda_r G_r^* .$$

FIGURE 4.2 - Fix-point algorithm


$S + \mathcal{L}(\Sigma^{(0)})$
$S = \Sigma^{(0)}$
$\mathcal{C}$
$\Sigma^{(1)} + \mathcal{L}(\Sigma^{(1)})$
$\Sigma^{(2)}$
$\Sigma^{(1)}$
$S + \mathcal{L}(\Sigma^{(1)})$

Here, the matrices

$$G_k^* = \frac{\partial h_k}{\partial \Sigma^{-1}}\bigg|_{\Sigma = \Sigma^{(m)}}, \quad (k=1,\ldots,r),$$

are fixed. As before, these equations can be solved with an iterative least-squares fitting program by minimizing a sum-of-squares objective function ξ,

$$\xi(\underset{\sim}{\lambda}) = \sum_{\ell} h_{\ell\lambda}^2 = \sum_{\ell} h_{\ell}^2(\Sigma_\lambda) \ .$$

The required derivatives of $h_{\ell\lambda}$ with respect to $\underset{\sim}{\lambda}$ are

$$\frac{\partial h_{\ell\lambda}}{\partial \lambda_k} = \sum_{i,j} \frac{\partial h_{\ell\lambda}}{\partial \sigma_{\lambda ij}} \frac{\partial \sigma_{\lambda ij}}{\partial \lambda_k}$$

$$= \mathrm{tr}(G_{\ell\lambda} G_k^*), \qquad (4.14)$$

where $G_{\ell\lambda} = G_\ell(\Sigma_\lambda)$.

The projection of S into $\mathbb{C}$ requires an initial value for $\underset{\sim}{\lambda}$ which one can take to be $\underset{\sim}{\lambda}_{(0)} = (0,\ldots,0)^T$ on the first outer step. On subsequent steps one can use the final $\underset{\sim}{\lambda}$ from the previous outer step as a new $\underset{\sim}{\lambda}_{(0)}$, although in some of the numerical examples it will be shown that better stability is achieved by taking $\lambda_{(0)} = (0,0,\ldots,0)^T$ at each step.

Solving the constrained likelihood equations in this way replaces the $\{\frac{1}{2}p(p+1)+r\}$-dimensional optimization problem of Section 4.2 by a fix-point iteration in $\frac{1}{2}p(p+1)$-dimensional space and a sequence of r-dimensional optimization problems. If r is small then the inner iterations will go quickly and the method will be effective even if p is fairly large. The main storage requirement is rp^2 locations for the matrices G_k^*, which can become a limitation.

A potential problem is that the fix-point procedure is not guaranteed to converge if a poor initial approximation to $\hat{\Sigma}$ is used, but in practice S is usually adequate. A more practical matter is that convergence when it occurs is linear and not quadratic as for the one-stage algorithm of Section 4.2, so high precision can be costly.

An interesting property of ML estimates for ξ can be seen by examining Eq. 4.12 . If all of the matrices $G_k^*(\hat{\Sigma})$ have a zero in the same position, then $\hat{\sigma}_{ij}$ is exactly s_{ij} in that position, and a similar property holds for LS and LS*. This is illustrated in the following example.

Example 4.1 - Covariance selection (cont'd)

In Dempster's covariance selection model (Ex. 3.3) the constraints are all of the form $\sigma^{ij} = 0$, so that each matrix G_k^* has a $\frac{1}{2}$ in position (i_k, j_k) and zeros elsewhere. Hence the solution $\hat{\Sigma}$ matches S in every unconstrained position, a property that Dempster establishes by a different argument. In particular, the variance estimates $\hat{\sigma}_{ii}$ are unaffected by imposing constraints of this type. We also note that since the matrices $G^*_k(\hat{\Sigma})$ do not depend on $\hat{\Sigma}$, the correct linear manifold $\mathcal{L}(\hat{\Sigma})$ for projecting S into $\hat{\Sigma}$ is known exactly, and final convergence occurs after one outer step.

4.4 Existence of Positive Definite Solutions

The foregoing sections tacitly assumed that the constrained optimization problem has a solution. We now show that a solution does exist, and briefly consider some related points.

As before, let Ω represent the space of p×p positive definite matrices, let $\Omega_0 = \{\Sigma \in \Omega\colon \ \underset{\sim}{h}(\Sigma) = 0\}$, and assume that Ω_0 is not empty. Then we have

Theorem 4.1

For fixed positive definite S the log Wishart likelihood $g(\Sigma) = g(\Sigma; S)$ assumes a maximum value for some $\hat{\Sigma}$ in Ω_0.

Proof. The set Ω_0 is not empty by assumption, so the constraints do not force all $\Sigma \in \Omega_0$ to be singular. Let Δ be some fixed element of Ω_0, and let $\overline{\Omega}$ be the closure of Ω. Since $g(\Sigma)$ is maximized over Ω when $\Sigma = S$, the constrained problem reduces to that of maximizing the continuous bounded function $g(\Sigma)$ over the compact set

$$\overline{\Omega} \cap h^{-1}(\underset{\sim}{0}) \cap \{\Sigma\colon \ g(\Delta) \leq g(\Sigma) \leq g(S)\} .$$

Hence the maximum is attained at some point $\hat{\Sigma}$ in $\bar{\Omega}$, and it remains to show only that $\hat{\Sigma}$ is not on the boundary of $\bar{\Omega}$. But $g(\Sigma)$ goes to $-\infty$ as Σ approaches any singular Σ^* from inside Ω, so $\hat{\Sigma}$ cannot be singular.

Some authors who have used general numerical methods to obtain maximum likelihood solutions have expressed concern over whether the "solutions" obtained might be matrices that are neither positive definite nor positive semidefinite (e.g., McDonald 1974). It is true that the Wishart likelihood is unbounded above in any neighborhood of the zero matrix for negative definite matrices of the form $-\alpha\hat{\Sigma}_{ML}$, $(\alpha > 0)$, but this is a minor practical issue, first because the problem can be reparametrized as in Section 4.3 to exclude consideration of matrices that are not positive definite and second because $g(\Sigma)$ goes to $-\infty$ as Σ approaches any boundary point from within Ω.

4.5 Derivatives for Conditional Covariance Constraints

Several of the examples discussed in Chapters 2 and 3 are defined by constraints that require conditional covariances to vanish, for example,

$$h(\Sigma) = \sigma_{12.b} = 0,$$

where b is some subset of the indices $\{3,4,\ldots,p\}$. This section obtains the derivatives that are required by the numerical algorithms of Sections 4.2 and 4.3.

First we consider the partial derivatives of $\sigma_{12.b}$ with respect to the elements of $\Psi = -\Sigma^{-1}$. Let a denote the complement of the set b of conditioning indices, so that in particular $(1,2) \in a \times a$. Since $\sigma_{12.b}$ is the (1,2) element of the matrix $SWP(b)\Sigma$, where SWP is an operator related to Beaton's SWP (see Appendix), we have using Eqs. A13 and A15,

$$\begin{aligned}\sigma_{12.b} &= [SWP(b)\Sigma]_{12} \\ &= [SWP(a)\Psi]_{12} \\ &= [-\Psi_{aa}^{-1}]_{12} \ .\end{aligned}$$

Thus, $\sigma_{12.b}$ is the (1,2) element in the inverse of the (a,a) partition of $-\Psi$. It follows that the derivatives of $\sigma_{12.b}$ with respect to elements of Ψ outside Ψ_{aa} are zero, and the derivatives inside Ψ_{aa} are obtained from Eq. A4,

$$\frac{\partial\sigma_{12.b}}{\partial\psi_{k\ell}} = \begin{cases} \psi_{aa}^{1k}\psi_{aa}^{\ell 2} & \text{for } (k,\ell) \in a \times a \\ 0 & \text{otherwise,} \end{cases} \qquad (4.15)$$

where ψ_{aa}^{ij} denotes the (i,j) element of Ψ_{aa}^{-1} . But the inverse of $-\Psi_{aa}$ is $\Sigma_{aa.b}$ which is the (a,a) partition of $SWP(b)\Sigma$, so Eq. 4.15 becomes

$$\frac{\partial\sigma_{12.b}}{\partial\sigma^{k\ell}} = \begin{cases} -\sigma_{1k.b}\sigma_{\ell 2.b} & \text{for } (k,\ell) \in a\text{x}a \\ 0 & \text{otherwise.} \end{cases} \tag{4.16}$$

In partitioned matrix notation, this is

$$\frac{\partial\sigma_{12.b}}{\partial\Sigma^{-1}} = \left[\begin{array}{c:c:c|c} \multicolumn{3}{c|}{} & 0 \\ \cdashline{4-4} \multicolumn{3}{c|}{} & 0 \\ \cdashline{4-4} \multicolumn{3}{c|}{-\Sigma_{a1.b}\Sigma_{2a.b}} & 0 \\ \hline 0 & 0 & 0 & 0 \end{array}\right]\begin{array}{l} \\ 1 \\ 2 \\ \\ \end{array} \quad \begin{array}{l} \left.\begin{array}{c} \\ \\ \\ \end{array}\right\}a \\ \left.\begin{array}{c} \\ \end{array}\right\}b \end{array} \quad . \tag{4.17}$$

(Rows and columns have been ordered in a way that facilitates comparisons with formulas below.) It is clear that if we use SWP(b)Σ to compute $h = \sigma_{12.b}$, then the derivatives follow with very little extra work.

When the symmetry of Σ is taken into account, as in Section 3.3, Eq. 4.16 becomes,

$$\frac{\partial\sigma_{12.b}}{\partial\sigma^{k\ell}} = \begin{cases} -\frac{1}{2}(\sigma_{1k.b}\sigma_{\ell 2.b} + \sigma_{1\ell.b}\sigma_{k2.b}) & (k,\ell) \in a\text{x}a \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore, the second derivatives of $\sigma_{12.b}$ which are required in Section 4.2 are obtained by applying Eq. 4.16 twice,

$$\frac{\partial^2\sigma_{12.b}}{\partial\sigma^{k\ell}\partial\sigma^{ij}} = \begin{cases} \sigma_{1k.b}\sigma_{\ell i.b}\sigma_{j2.b} + \sigma_{1i.b}\sigma_{jk.b}\sigma_{\ell 2.b} & i,j,k,\ell \in a \\ 0 & \text{otherwise.} \end{cases}$$

With symmetry explicitly allowed for, there are eight terms.

Next we consider the derivatives of $h = \sigma_{12.b}$ with respect to the elements of Σ. Although one could use Eq. 4.17 and the formula

$$\frac{\partial\sigma_{12.b}}{\partial\Sigma} = - \Sigma^{-1} \frac{\partial\sigma_{12.b}}{\partial\Sigma^{-1}} \Sigma^{-1} ,$$

a direct approach produces a more useful result. Writing

$$\sigma_{12.b} = \sigma_{12} - \Sigma_{1b}\Sigma_{bb}^{-1}\Sigma_{b2} ,$$

ignoring symmetry, and differentiating with respect to each partition of Σ in turn, gives

$$\frac{\partial\sigma_{12.b}}{\partial\Sigma} = \left[\begin{array}{c:c:c|c} 0 & 0 & 0 & 0 \\ \hdashline 0 & 0 & 1 & -\Sigma_{2b}\Sigma_{bb}^{-1} \\ \hdashline 0 & 0 & 0 & 0 \\ \hline 0 & 0 & -\Sigma_{bb}^{-1}\Sigma_{b1} & \Sigma_{bb}^{-1}\Sigma_{b1}\Sigma_{2b}\Sigma_{bb}^{-1} \end{array}\right] \begin{array}{l} \\ 1 \\ 2 \\ \\ \end{array} \begin{array}{l} \left.\begin{array}{l} \\ \\ \\ \end{array}\right\} a \\ \left.\begin{array}{l} \\ \end{array}\right\} b \end{array}$$

Observe that $\Sigma_{2b}\Sigma_{bb}^{-1}$, which we shall write as $\Sigma_{2b.b}$, is a row vector consisting of the multiple regression coefficients of variable X_2 regressed on variables X_j, $(j \in b)$, and it is also the (2,b) partition of the matrix SWP(b)Σ. Thus

$$\frac{\partial\sigma_{12.b}}{\partial\Sigma} = \left[\begin{array}{c:c:c|c} 0 & 0 & 0 & 0 \\ \hdashline 0 & 0 & 1 & -\Sigma_{2b.b} \\ \hdashline 0 & 0 & 0 & 0 \\ \hline 0 & 0 & -\Sigma_{b1.b} & \Sigma_{b1.b}\Sigma_{2b.b} \end{array}\right], \qquad (4.18)$$

which can also be calculated directly from SWP(b)Σ.

The relationship of a vanishing covariance constraint to Σ on the one hand and to Σ^{-1} on the other has an interesting symmetry, which we show by introducing a modified form of the constraint. Requiring $h(\Sigma)$ to vanish is equivalent to requiring

$$h^{\#}(\Sigma) = h(\Sigma)\ p(\Sigma) \qquad (4.19)$$

to vanish, where $p(\Sigma)$ is any continuously differentiable function without zeros or singularities in Ω. In particular, if $h = \sigma_{12.b}$, then define $h^{\#}$ to be

$$h^{\#} = \frac{-\sigma_{12.b}}{\sigma_{11.b}\sigma_{22.b}-\sigma_{12.b}\sigma_{21.b}} .$$

The functions h and $h^{\#}$ define the same constrained subspace of Ω. Whereas h is the (1,2) element of SWP(b)Σ, $h^{\#}$ is the (1,2) element of SWP(1,2,b)Σ. Now define new index sets α and β,

$$\alpha = b \cup \{1,2\}, \qquad \beta = a - \{1,2\} ,$$

which are depicted below,

$$\rlap{$\overbrace{}^{a}\overbrace{}^{b}$}\underbrace{x\;x\;\ldots\;x}_{\beta}\underbrace{\;1\;2\;x\;x\;\ldots\;x}_{\alpha} .$$

Let $\Psi = -\Sigma^{-1}$, as before. Since

$$\text{SWP}(1,2,b)\Sigma = \text{SWP}(\alpha)\Sigma = \text{SWP}(\beta)\Psi ,$$

it follows that the alternate constraint function $h^{\#}$ is just

$$h^{\#} = \psi_{12.\beta} ,$$

which is formally a conditional covariance derived from the "covariance" matrix Ψ. Noting that $\psi_{12.\beta}$ has the same relationship to $(\Psi,\Sigma,\alpha,\beta)$ as $\sigma_{12.b}$ has to (Σ,Ψ,a,b), we can write

$$\psi_{12.\beta} = \frac{-\sigma_{12.b}}{\sigma_{11.b}\sigma_{22.b}-\sigma^2_{12.b}} \quad \text{and} \quad \sigma_{12.b} = \frac{-\psi_{12.\beta}}{\psi_{11.\beta}\psi_{22.\beta}-\psi^2_{12.\beta}} \quad ,$$

and obtain the derivatives of $\psi_{12.\beta}$ directly from Eqs. 4.17 and 4.18 ,

$$\frac{\partial\psi_{12.\beta}}{\partial\Sigma} = -\frac{\partial\psi_{12.\beta}}{\partial\Psi^{-1}} = \left[\begin{array}{c|c:c:c} 0 & 0 & 0 & 0 \\ \hline 0 & & & \\ \hdashline 0 & \multicolumn{3}{c}{-\Psi_{\alpha 1.\beta}\Psi_{2\alpha.\beta}} \\ \hdashline 0 & & & \end{array}\right]\begin{array}{l} \\ 1 \\ 2 \\ \\ \end{array}\begin{array}{l} \}\,\beta \\ \\ \}\,\alpha \\ \\ \end{array} \qquad (4.20)$$

$$\frac{\partial\psi_{12.\beta}}{\partial\Sigma^{-1}} = -\frac{\partial\psi_{12.\beta}}{\partial\Psi} = \left[\begin{array}{c|c:c:c} -\Psi_{\beta 1.\beta}\Psi_{2\beta.\beta} & 0 & \Psi_{\beta 1.\beta} & 0 \\ \hline \Psi_{2\beta.\beta} & 0 & 1 & 0 \\ \hdashline 0 & 0 & 0 & 0 \\ \hdashline 0 & 0 & 0 & 0 \end{array}\right] \qquad (4.21)$$

Both of these matrices, as well as the value of $\psi_{12.\beta}$ itself can be computed easily from $SWP(\beta)\Psi = SWP(1,2,b)\Sigma$.

This duality between h and $h^{\#}$ has some theoretical and practical relevance. By differentiating Eq. 4.19 one sees that $\frac{\partial h}{\partial\Sigma^{-1}}$ and $\frac{\partial h^{\#}}{\partial\Sigma^{-1}}$ are proportional to each other when evaluated at any Σ that lies in the constrained subspace.

Therefore, the extra zeros that appear in Eq. 4.21 must also occur in Eq. 4.17 . Those in positions (1,1) and (2,2) are important because they remain zero when $G_k^* = \frac{\partial h_k}{\partial \Sigma^{-1}}$ is computed from $\frac{\partial h_k}{\partial \Sigma^{-1}}$. Having thus identified a larger set of zeros in each G_k^*, one can in some cases find additional elements of $\hat{\Sigma}$ that are equal to the corresponding elements of S.

The practical importance of this duality is that by using $h^{\#}$ instead of h to define the linear manifold $\mathcal{L}(\Sigma^{(m)})$ one can improve convergence of the fix-point calculation in some cases.

4.6 Exact Solution for Non-Cyclical Path Models

It was stated above that path models without cycles (Ex. 3.5) have closed-form maximum likelihood solutions. We develop this point here partly because of its intrinsic interest, and partly to provide numerical examples with known solutions for testing the iterative computational methods developed above.

Starting from Ex. 3.5, one can order the variables in a way that allows the inverse covariance matrix for a non-cyclical path model to be factored into

$$\Sigma^{-1} = C^T DC \;,$$

where D is a non-negative diagonal matrix and C is lower triangular with ones on the diagonal. (The matrix A in Ex. 3.5 is $D^{\frac{1}{2}}C$ in this notation.) The missing forward paths in the diagram define a fixed pattern of zeros in the lower triangle of C. Apart from constants, the log Wishart likelihood as a function of C and D is

$$L(C,D) = \mathrm{tr}(C^T DCS) - \sum_k \log d_{kk} \; . \tag{4.22}$$

Differentiating first with respect to an element of D we have,

$$\frac{\partial L}{\partial d_{kk}} = [CSC^T]_{kk} - \frac{1}{d_{kk}} \; ,$$

from which follows,

$$\hat{d}_{kk} = \sum_{i=1}^{k} \sum_{j=1}^{k} \hat{c}_{ki} \hat{c}_{kj} s_{ij} \; . \tag{4.23}$$

Now differentiate Eq. 4.22 with respect to C, using Eq. A8,

$$\frac{\partial L}{\partial C} = 2 \cdot \text{lower triangle } \{DCS\} \; .$$

For each c_{ij} $(i > j)$ not constrained to be zero we have,

$$\frac{\partial L}{\partial c_{ij}} = 2\, d_{ii}\Big(s_{ij} + \sum_{k<i} c_{ik}s_{kj}\Big) ,$$

so the likelinood equations become

$$\sum_{k<i} \hat{c}_{ik}s_{kj} = -s_{ij} \quad (j < i) .$$

The first few of these written out are

$$\left.\begin{array}{ccl}
i & j & \\
\hline
2 & 1 & c_{21}s_{11} = -s_{21} \\
\hline
3 & 1 & c_{31}s_{11} + c_{32}s_{21} = -s_{31} \\
3 & 2 & c_{31}s_{12} + c_{32}s_{22} = -s_{32} \\
\hline
4 & 1 & c_{41}s_{11} + c_{42}s_{21} + c_{43}s_{31} = -s_{41} \\
\begin{array}{c}4\\4\end{array} & \begin{array}{c}2\\3\end{array} & \begin{array}{l}c_{41}s_{12} +\\ c_{41}s_{13} +\end{array}\boxed{\begin{array}{l}c_{42}s_{22} + c_{43}s_{32} = -s_{42}\\ c_{42}s_{23} + c_{43}s_{33} = -s_{43}\end{array}} \\
\ldots & & \ldots
\end{array}\right\} \qquad (4.24)$$

Each group of equations exactly determines one row of C.

Now suppose the path from X_1 to X_4 is missing, so that c_{41} is zero by assumption. Then the (4,1) equation in 4.24 is removed since we no longer have $\frac{\partial \ell}{\partial c_{41}} = 0$, and the terms involving c_{41} vanish, so the equations in the third group reduce to those in the dashed box, and the number of equations matches the number of c's in the box. It follows that the non-zero elements in the k-th row of $\hat{C}$ are the negatives of the multiple regression coefficients of X_k regressed on just those variables that are its immediate causes in the path diagram. Essentially, maximum likelihood leads to the fitting of each regression equation separately.

The matrix $\hat{C}$ can be calculated easily using the SWP operator and the following scheme which is a modification of a procedure given by Dempster (1969, p. 63) for the unconstrained case.

i) At stage k, S has been swept on those indices in $\{1,\ldots,k-1\}$ that correspond to c_{kj}'s in the k-th row of C which are not constrained to be zero. Partition the resulting matrix as

$$S^* = \left[\begin{array}{c:c} G_{11} & H_{21}^T \\ \hdashline H_{21} & G_{22} \end{array}\right]\begin{array}{l} 1 \\ \vdots \\ k-1 \\ k \\ \vdots \\ k \end{array} .$$

ii) Take the non-zero elements of the k-th row of $\hat{C}$ from the corresponding positions in the first row of H_{21}. The rest of the k-th row of $\hat{C}$ is filled with zeros except for a 1 in position (k,k).

iii) SWP S* on the indices in the symmetric difference of the sets of indices not constrained in the k-th and (k+1)-st rows of C. Let k = k+1 and return to step (i).

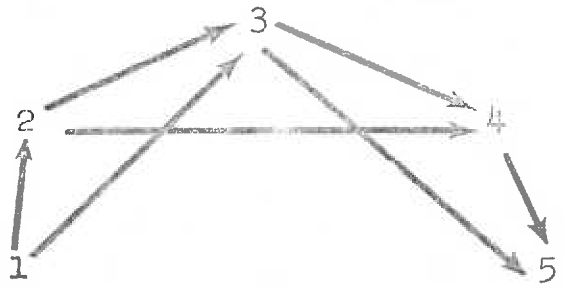
After $\hat{C}$ is computed in this way, $\hat{D}$ is obtained from Eq. 4.23 , and $\hat{\Sigma}^{-1}$ from $\hat{\Sigma}^{-1} = \hat{C}^T\hat{D}\hat{C}$.

4.7 Numerical Examples

Two examples are provided here to compare the performance of the computational algorithms developed above.

Example 4.2 - A five variable path model

Consider a path model with five variables, each affected only by its immediate two predecessors, i.e.,



This is a non-cyclical path model (cf. Ex. 3.4) with three missing forward paths, $1 \not\to 4$, $1 \not\to 5$ and $2 \not\to 5$, corresponding to three conditional covariance constraints,

$$\sigma_{14.23} = 0, \quad \sigma_{15.34} = 0, \quad \sigma_{25.34} = 0 \ . \qquad (4.25)$$

Suppose the following sample covariance matrix has been observed:

$$S = \begin{bmatrix} 1.0 & .5 & .5 & .5 & .5 \\ .5 & 1.0 & .5 & .5 & .5 \\ .5 & .5 & 1.0 & .5 & .5 \\ .5 & .5 & .5 & 1.0 & .5 \\ .5 & .5 & .5 & .5 & 1.0 \end{bmatrix} .$$

The direct calculation in Section 4.6 was used to obtain the exact solution which is shown in Fig. 4.3. The computation took .031 seconds on the Honeywell 6070, a computer which requires 3.1 microseconds for a floating point multiplication.

Note that the matrix $\hat{\Sigma}$ matches S in every unconstrained position. This is consistent with the fact mentioned in Ex. 3.4 that the constraints here are equivalent to three constraints on Σ^{-1}, namely, $\sigma^{14} = \sigma^{15} = \sigma^{25} = 0$. The equality between elements of $\hat{\Sigma}$ and S for such models was established in the discussion of Ex. 4.1.

FIGURE 4.3 - Maximum likelihood solution to Ex. 4.2

$$\hat{C} = \begin{bmatrix} 1.0 & 0 & 0 & 0 & 0 \\ -.50 & 1.0 & 0 & 0 & 0 \\ -.3\overline{3} & -.3\overline{3} & 1.0 & 0 & 0 \\ 0 & -.3\overline{3} & -.3\overline{3} & 1.0 & 0 \\ 0 & 0 & -.3\overline{3} & -.3\overline{3} & 1.0 \end{bmatrix}$$

$$\hat{D} = \begin{bmatrix} 1.0 & 0 & 0 & 0 & 0 \\ 0 & 1.3\overline{3} & 0 & 0 & 0 \\ 0 & 0 & 1.5 & 0 & 0 \\ 0 & 0 & 0 & 1.5 & 0 \\ 0 & 0 & 0 & 0 & 1.5 \end{bmatrix}$$

$$\hat{\Sigma}^{-1} = \hat{C}^T\hat{D}\hat{C} = \begin{bmatrix} 1.5 & -.5 & -.5 & 0 & 0 \\ -.5 & 1.6\overline{6} & -.33 & -.5 & 0 \\ -.5 & -.3\overline{3} & 1.83\overline{3} & -.3\overline{3} & -.5 \\ 0 & -.5 & -.3\overline{3} & 1.6\overline{6} & -.5 \\ 0 & 0 & -.5 & -.5 & 1.5 \end{bmatrix}$$

$$\hat{\Sigma} = \begin{bmatrix} 1.0 & .5 & .5 & .3\overline{3} & .27\overline{7} \\ .5 & 1.0 & .5 & .5 & .3\overline{3} \\ .5 & .5 & 1.0 & .5 & .5 \\ .3\overline{3} & .5 & .5 & 1.0 & .5 \\ .27\overline{7} & .3\overline{3} & .5 & .5 & 1.0 \end{bmatrix}$$

The single-stage iterative calculation of Section 4.2 was next applied to this model. The parameter space is effectively 18-dimensional, including the lower triangle of A and three Lagrange multipliers. The initial approximation was taken to be $A^{(0)} = I_{5\times5}$ and $\lambda^{(0)} = (0,0,0)^T$, and Table 4.1 shows the details of the run.

In Table 4.1, line (i) displays the Euclidean distances between the exact maximum likelihood estimate $\hat{A}$, and successive approximations $A^{(n)}$. Line (ii) shows the L_2 norm of the "residual" vector $\underset{\sim}{\varphi}$ at each iteration, i.e., the vector of derivatives of the Lagrangian log-likelihood function with respect to the parameters. Both lines clearly show that convergence of the algorithm is quadratic (up to the 8-digit precision of the computer), that is, the error is approximately squared at each step. The computation would normally be stopped when $\|\underset{\sim}{\varphi}\|$ became smaller than some preassigned constant, ε. In this example, an accuracy of $\varepsilon = 1E\text{-}5$ required three iterations and .487 seconds of computer time.

The fix-point iteration of Section 4.3 was also applied to this model, projecting S repeatedly into the constrained parameter space. Initially, $\underset{\sim}{\lambda}^{(0)}$ was taken to be $(0,0,0)^T$, and on subsequent outer steps to be the final $\underset{\sim}{\lambda}$ from the previous iteration.

TABLE 4.1 - Single stage iteration

		Iteration				
θ	Exact $\hat{\theta}$	0	1	2	3	4
a(1,1)	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
a(2,1)	-.577350	.000000	-.571429	-.577320	-.577350	-.577350
a(2,2)	1.154701	1.000000	1.142857	1.154639	1.154701	1.154701
a(3,1)	-.408248	.000000	-.400000	-.408163	-.408248	-.408248
a(3,2)	-.408248	.000000	-.400000	-.408163	-.408248	-.408248
a(3,3)	1.224745	1.000000	1.200000	1.224490	1.224745	1.224745
a(4,1)	.000000	.000000	.000000	.000000	.000000	.000000
a(4,2)	-.408248	.000000	-.400000	-.408163	-.408248	-.408248
a(4,3)	-.408248	.000000	-.400000	-.408163	-.408248	-.408248
a(4,4)	1.224745	1.000000	1.200000	1.224490	1.224745	1.224745
a(5,1)	.000000	.000000	.000000	.000000	.000000	.000000
a(5,2)	.000000	.000000	.000000	.000000	.000000	.000000
a(5,3)	-.408248	.000000	-.400000	-.408163	-.408248	-.408248
a(5,4)	-.408248	.000000	-.400000	-.408163	-.408248	-.408248
a(5,5)	1.224745	1.000000	1.200000	1.224490	1.224745	1.224745
$\lambda(1)$	.204124	.000000	.200000	.204082	.204124	.204124
$\lambda(2)$	.204124	.000000	.200000	.204082	.204124	.204124
$\lambda(3)$	.204124	.000000	.200000	.204082	.204124	.204124
(i) $\|A^{(n)} - \hat{A}\|$		1.228331	.049198	.4932E-3	.5662E-7	.6452E-8
(ii) $\|\underset{\sim}{\varphi}\|$		1.581139	.060434	.5963E-3	.7357E-7	.1893E-7
(iii) max $\|a_{ij}^{(n)} - a_{ij}^{(n-1)}\|$			-.571429	.2449E-1	.2550E-3	.2980E-7
(iv) time (sec.)			.1603	.1651	.1617	.1622
(v) cum. time			.1603	.3254	.4871	.6493

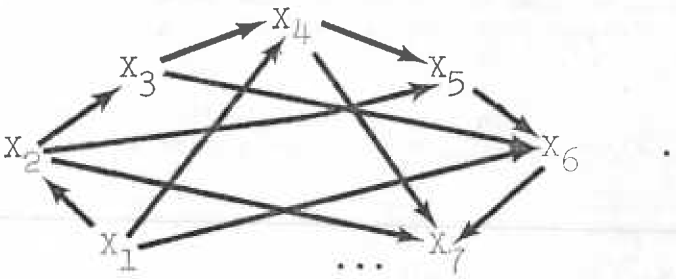
Table 4.2 shows the results after each outer iteration. The parameter space is represented here by Σ rather than A, but to facilitate comparison with Table 4.1 line (ii) shows the distance to the solution in terms of A. There are two stopping criteria to choose here. For the inner iteration the length of the constraint vector (line (iii)) was required to be less than $\varepsilon_1 = 1E-6$; the numbers of iterations needed to achieve this are shown in line (v). For the outer iteration the maximum change in σ_{ij} (line (iv)) was required to be less than ε_2. If $\varepsilon_2 = 1E-5$, then three steps and .212 seconds of computer time are required, which is less than half the computer time taken by the other iterative method. One must bear in mind, however, that exact comparisons are difficult since the efficiency of the coding is a critical factor. It is interesting to note that convergence is irregular here, with a significant jump in accuracy occuring in the second outer step. This illustrates the point made at the end of Section 4.5. Since the constraints in this example can be written in terms of Σ^{-1} alone, as soon as a trial value $\Sigma^{(1)}$ is obtained which lies in the constrained space $\mathcal{C}$ then the correct linear manifold $\mathfrak{L}(\Sigma^{(1)}) = \mathfrak{L}(\hat{\Sigma})$ is obtained exactly, and the next iterate coincides with the correct solution, $\Sigma^{(2)} = \hat{\Sigma}$, apart from round-off errors.

TABLE 4.2 - Fix-point iteration

		Outer iteration				
θ	Exact $\hat{\theta}$	0	1	2	3	4
$\sigma(1,1)$	1.000000	1.000000	.831978	1.000001	1.000000	1.000000
$\sigma(2,1)$	.500000	.500000	.419824	.500000	.500000	.500000
$\sigma(2,2)$	1.000000	1.000000	.919849	1.000000	1.000000	1.000000
$\sigma(3,1)$	.500000	.500000	.500000	.500000	.500000	.500000
$\sigma(3,2)$	.500000	.500000	.500000	.500000	.500000	.500000
$\sigma(3,3)$	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
$\sigma(4,1)$	.333333	.500000	.313381	.333333	.333333	.333333
$\sigma(4,2)$	.500000	.500000	.500000	.500000	.500000	.500000
$\sigma(4,3)$	.500000	.500000	.500000	.500000	.500000	.500000
$\sigma(4,4)$	1.000000	1.000000	.912179	1.000000	1.000000	1.000000
$\sigma(5,1)$	.277778	.500000	.268675	.277778	.277778	.277778
$\sigma(5,2)$	.333333	.500000	.323663	.333333	.333333	.333333
$\sigma(5,3)$	.500000	.500000	.500000	.500000	.500000	.500000
$\sigma(5,4)$	.500000	.500000	.445112	.500000	.500000	.500000
$\sigma(5,5)$	1.000000	1.000000	.849763	1.000000	1.000000	1.000000
$\lambda(1)$	-.750000	.000000	-.790387	-1.087261	-.749999	-.750000
$\lambda(2)$	-.500000	.000000	-.577568	-.895549	-.499999	-.500000
$\lambda(3)$	-.583333	.000000	-.576966	-.841130	-.583333	-.583333
(i) $\Vert\Sigma^{(n)} - \hat{\Sigma}\Vert$		.458123	.291496	1.0394E-6	4.6529E-8	3.4143E-8
(ii) $\Vert A^{(n)} - \hat{A}\Vert$		.548909	.271476	8.5847E-7	4.3579E-8	3.2127E-8
(iii) $\Vert \underset{\sim}{h} \Vert$		.288675	8.2127E-7	7.0224E-9	5.6563E-6	2.4970E-9
(iv) max $\vert\sigma_{ij}^{(n)} - \sigma_{ij}^{(n-1)}\vert$			.231325	.168023	7.1526E-7	2.2352E-8
(v) no. inner steps			2	1	1	1
(vi) time (sec.)			.0897	.0604	.0616	.0618
(vii) cum. time			.0897	.1501	.2117	.2735

Example 4.3 - A sequence of path models of increasing size

In order to explore the ways in which computer resource requirements for these computational methods increase with the number of variables and number of constraints, we consider a family of path models defined by an ordered sequence of random variables in which each variable is directly affected by its immediate predecessor and by predecessors at lags 3, 5, 7, etc. As before, these models have easily calculated exact solutions. The path diagram is,



The matrix factor A in $\Sigma^{-1} = A^TA$ has zeros in the subdiagonals numbered 2, 4, 6, etc. For this example we take a population covariance matrix $\Sigma_t = (A_t^TA_t)^{-1}$, where

$$A_t = \left[\begin{array}{ccccccc} 1 & & & & & & \\ -.6 & 1 & & & 0 & & \\ 0 & -.6 & 1 & & & & \\ (.6)^2 & 0 & -.6 & 1 & & & \\ 0 & (.6)^2 & 0 & -.6 & 1 & & \\ -(.6)^3 & 0 & (.6)^2 & 0 & -.6 & 1 & \\ 0 & -(.6)^3 & 0 & (.6)^2 & 0 & -.6 & 1 \ddots \\ & & \cdots & & & & \end{array}\right],$$

and we use an observation from a pseudo-random Wishart $W_{12\times 12}(30,\Sigma_t)$ generator to obtain a sample covariance matrix S.

In order to define an increasing sequence of models, we take initial sequences of p variables, for successive values of p. Each of these models corresponds to the upper-left partitions of the matrices Σ_t, A_t, and S. The number of covariances q and the number of constraints r both increase quadratically with p, and the amounts of computer storage for working data arrays required by the two iterative methods are shown in Table 4.3. Clearly the rapid increase in storage requirements sets a practical limitation of about p = 13 or 14 on the size of problem that can be accommodated by either method.

When applying the fix-point procedure in the form used in Ex. 4.3, a difficulty was encountered: the approximation $\underset{\sim}{\lambda}^{(1)}$ obtained after the first fix-point step was so far from the true $\underset{\sim}{\hat{\lambda}}$ that using it as an initial value for the second step caused the calculation to diverge. But when the procedure was modified to take $\underset{\sim}{\lambda}^{(0)} = \underset{\sim}{0}$ each time, satisfactory convergence was achieved.

An overall convergence criterion of ε_1 = 1E-4 was employed, along with a limit of 6 on the number of (outer) iterations. The inner stage of the fix-point was terminated by ε_2 = 1E-6. Table 4.4 compares the numbers of iterations

TABLE 4.3 - Storage requirements for Ex. 4.3

no. X´s	p	6	8	10	12	14
no. constraints	$r=\frac{1}{2}\{\frac{1}{2}(p^2+1)-p\}$	6	12	20	30	42
no. var´s & cov´s	$q=\frac{1}{2}p(p+1)$	21	36	55	78	105
no. param´s,1-stage	$q+r$	27	48	75	108	147
storage for 1-stage	$(q+r)^2+5(q+r)$	864	2544	6000	12204	22344
storage for 2-stage	$p^2(r+4)+r^2+5r$	426	1228	2900	5946	10990

TABLE 4.4 - Computer times for Ex. 4.3

	exact calc.	single-stage		fix-point (2-stage)		
p	time	number iter´s	time	number outer steps	number inner steps	time
5	.042	4	.648	5	6	.533
6	.075	4	1.451	5	7	1.420
8	.144	4	6.204	6*	10*	9.164
10	.305	4	21.148	6*	10*	30.628
12	.577	4	60.575	-	-	-

and computer times for the two iterative methods, as well as the time required by the exact calculation.

The starred entries in Table 4.4 are cases in which the iteration limit was reached before the convergence criterion was satisfied. In both such cases an accuracy of about 1E-3 had been attained, with convergence at the rate of about one order of magnitude for two outer steps.

As in the previous example a detailed comparison of computer times between methods is not really meaningful, since either might have been more or less efficiently coded. However the orders of magnitudes and trends are important. For problems of this sort where the number of constraints increases in fixed proportion to the number of covariances and where high accuracy is not required, the fix-point method does well for small problems but is soon surpassed by the single-stage method. If the number of constraints were to remain small, the competitive advantage of the fix-point method would increase, but if higher accuracy were required, it would decrease. It is interesting to note that the number of iterations to convergence for the one-stage procedure does not increase in the larger problems in this range, but computer times increase quadratically with the number of parameters, q+r. It is clear that a model with p=10 and r=20 is at or beyond the limit of routine calculation for exploratory purposes.

Chapter 5

THE USE OF CORRELATIONS IN THE STUDY OF STRUCTURED WISHART MODELS

5.1 Structured Correlation Models

A majority of the structured Wishart models discussed in Chapter 3 are invariant under coordinate scale transformations. They consequently can be characterized by patterns or constraints imposed on the population correlation matrix, with the sample correlation matrix forming a set of invariant statistics. It seems natural that methods for selecting appropriate models from families of this kind, for fitting them, and for testing their fit to observed data, should be based on the sample correlation matrix.

5.2 Normalizing Transformations

In certain respects sample correlations are difficult statistics to work with, because the distribution function of even a single sample correlation r under normal theory assumptions has a complicated form. Although the distribution is asymptotically normal with increasing sample size, the approach is too slow to be of much direct practical use for samples of moderate size. A standard method has been to apply a transformation to r to improve its distributional properties. Of several transformations proposed (Hotelling, 1953), the most useful has been that suggested by Fisher (1915),

$$z(r) = \tanh^{-1}(r) = \tfrac{1}{2} \log \left(\frac{1+r}{1-r}\right) . \qquad (5.1)$$

This transformation has the following desirable properties:

p1. Whereas r is confined to the interval (-1,1), z(r) has a distribution that ranges over the entire real line.

p2. The expectation of z(r) is $z(\rho) + O(n^{-1})$, where ρ is the population correlation and n is the sample size.

p3. The variance of z(r) is $1/(n-2) + O(n^{-2})$ which to this order of approximation does not depend on ρ. (The n here is the effective number of observations that remain after the mean and possibly other linear effects have been removed.)

p4. The skewness and kurtosis of z(r) are reduced, making the distribution of z(r) approach normality considerably faster than that of r.

As a result of these properties, even for small samples, the distribution of z(r) can be treated for most practical purposes as if it were $N(z(\rho), 1/(n-2))$.

When there are several variables to consider and thus a matrix of sample correlations, the situation is more complicated. By analogy with the bivariate case, one would like to transform the set of $q = \frac{1}{2}p(p-1)$ correlations to approximate joint normality so that statistical and data analytical techniques based on the multinormal distribution might be used. In place of p1, one now seeks to extend the range of the distribution from a subset of the unit hypercube in $\mathbb{R}^q$ to all of $\mathbb{R}^q$. The properties p2, p3 and p4 are still desirable for the marginal distributions of the transformed r_{ij}, but they are not sufficient for joint normality. And when a normalizing transformation is found, the transformed statistics will almost certainly have a correlation structure that depends on the unknown ρ_{ij}'s, and this will have an effect on inferences.

There are several ways to generalize $z(\cdot)$ for use with 3×3 and larger correlation matrices. One approach which will be pursued in the following section is to transform each r_{ij} individually. Another, which we mention only briefly here and which follows from a suggestion from J. W. Tukey (1973), is to ask whether there is a matrix analytic function (cf. Lancaster, 1969, p. 183) which generalizes $z(r)$. We find that the matrix logarithm, which appeared in a different context in Section 2.4 does this, since

$$\log \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{2}\log(1+r)(1-r) & \frac{1}{2}\log(1+r)/(1-r) \\ \frac{1}{2}\log(1+r)/(1-r) & \frac{1}{2}\log(1+r)(1-r) \end{bmatrix} .$$

Whether it is a useful normalizing transformation when applied to 3×3 and larger correlation matrices remains to be explored, but we note the following relevant points. First, log(·) maps the set of positive definite matrices onto the whole set of symmetric matrices (Nagao, 1973), so at least when applied to covariance matrices it removes the positive definiteness restriction. Second, it preserves constant-in-blocks structure, and thus might be a particularly useful transformation when block-structure models are contemplated. Third, since $\log R^{-1} = -\log R$, an analysis of R on the log scale is virtually the same as an analysis of R^{-1}.

5.3 The Element-By-Element Generalization of z

The normalizing transformation $z = \tanh^{-1}(\cdot)$ can be applied to a sample correlation matrix by transforming each correlation separately. Fisher (1924a) did this to a table of correlations derived from rainfall statistics, as did Hills (1969) in the paper mentioned in Ex. 3.2, but neither author examined the joint distribution of the transformed r_{ij}'s. That the r_{ij} are themselves correlated, and so are the statistics z_{ij}, was stressed by Elston (1974), but a thorough study of the joint distribution does not seem to have been published.

We note that z(·) can be used quite generally as a normalizing transformation for any conditional (i.e. partial) correlation, since such a statistic has the same marginal distribution as an ordinary sample correlation but with a modified effective sample size (cf. Section 3.5). Insofar as sample partial correlations are the invariant ancillary statistics for most of the models discussed in Chapter 3, $z(\cdot)$ is potentially useful in a fairly wide range of problems.

It is often implicitly assumed that the joint distribution of a set of z_{ij} is nearly normal. We will make this assumption as well, noting that it extends to sets of $z_{ij\cdot a} = z(r_{ij\cdot a})$, and citing the following points as support: (i) the statistics $z_{ij\cdot a}$ are all continuous twice-differentiable functions of the sample covariances which are asymptotically jointly normal as sample size increases (Cramér, 1946, p. 366; Anderson, 1958, p. 77); (ii) the marginal distributions approach normality quickly; and (iii) some preliminary computer simulations of pairs of $z_{ij\cdot a}$ have shown no systematic departure from joint normality.

Having assumed joint normality, we next consider the correlations among the $z_{ij\cdot a}$. One approach to estimating these correlations is to apply one of the asymptotic covariance formulas from Lemma 3.2, namely,

$$\operatorname{cov}(r,r') = \frac{2}{n}\, rr'\operatorname{tr}\left(\Sigma \frac{d \log r}{d\Sigma}\, \Sigma\, \frac{d \log r'}{d\Sigma}\right) \quad , \qquad (5.2)$$

and to substitute for the unknown true σ_{ij}'s values $\hat{\sigma}_{ij}$ estimated within the structured model. Covariances among the r's are converted to correlations, which are also the asymptotic correlations among the corresponding z's, and the asymptotic variance of $z_{ij \cdot a}$ is

$$\mathrm{var}(z_{ij.a}) = \frac{1}{n-2-q_a}\,, \tag{5.3}$$

where q_a is the number of indices in the set a. The matrix derivatives required in Eq. 5.2 are obtained from the formulas in Section 4.5, especially Eq. 4.18 . Although Eq. 5.2 does not simplify much for general pairs $r_{ij \cdot a}$ and $r_{k\ell .b}$, there are some interesting special cases of which six are discussed below.

<u>5.3.1</u>. The asymptotic covariances between pairs of ordinary correlations are obtained from Eq. 5.2 or as special cases of Elston's (1974) formulas for intraclass correlations. The two relevant cases are pairs of r_{ij} in the same row (or column) and pairs in distinct rows and columns:

$$\mathrm{cov}(r_{12},r_{23}) = \frac{1}{n}\left\{\rho_{13}(1-\rho_{12}^2-\rho_{23}^2)-\tfrac{1}{2}\rho_{12}\rho_{23}(1-\rho_{12}^2-\rho_{13}^2-\rho_{23}^2)\right\} \tag{5.4}$$

$$\begin{aligned}\mathrm{cov}(r_{12},r_{34}) = \frac{1}{n}\Big\{&\tfrac{1}{2}\rho_{12}\rho_{34}(\rho_{13}^2+\rho_{14}^2+\rho_{23}^2+\rho_{24}^2)\\ &-(\rho_{12}\rho_{13}\rho_{14}+\rho_{12}\rho_{32}\rho_{42}+\rho_{13}\rho_{23}\rho_{43}+\rho_{14}\rho_{24}\rho_{34})\\ &+\rho_{13}\rho_{24}+\rho_{14}\rho_{23}\Big\}\,.\end{aligned} \tag{5.5}$$

These are converted to correlations with the help of

$$\operatorname{var}(r_{12}) = \frac{1}{n}(1-\rho_{12}^2)^2 \quad . \tag{5.6}$$

(In fact, Eqs. 5.4 and 5.6 are really special cases of Eq. 5.5.)

5.3.2. The ancillary statistics for testing the presence of zeros in Σ^{-1} were shown in Ex. 3.8 to be conditional correlations $r_{ij}^* = r_{ij.c}$, where the set $c = c_{ij}$ contains all other indices besides i and j, and the r_{ij}^* are in turn the negatives of "correlations" formally computed from S^{-1},

$$r_{ij}^* = -s^{ij}/\sqrt{s^{ii}s^{jj}} \quad .$$

Asymptotically S^{-1} has a Wishart distribution with parameter Σ^{-1}, so it follows that the asymptotic variances and covariances of the z_{ij}^* are obtained from Eqs. 5.3 - 5.6 by replacing the parameters ρ_{ij} with

$$\rho_{ij}^* = \sigma^{ij}/\sqrt{\sigma^{ii}\sigma^{jj}} \ ,$$

where σ^{ij} is an element of Σ^{-1}.

5.3.3. The joint distribution of ordinary sample correlations r_{ij} is simplified when the population correlations

ρ_{ij} are all zero. This is the case considered in Hills' paper referred to in Ex. 3.2. He warned that the effect of the dependence between the correlation coefficients might be difficult to predict, but we are able to observe that according to Eqs. 5.4 and 5.5, all pairs of r_{ij} are asymptotically uncorrelated. In fact much stronger results obtain: all pairs of correlations are exactly uncorrelated and statistically independent. Furthermore, statistical independence holds among all correlations in any set of r_{ij} in a single column or row of R (Anderson, 1958, Ch. 7, Ex. 9), and indeed for any set of sample correlations with the property defined in the following lemma:

Lemma 5.1

Suppose that S is distributed as $W_p(n,\Sigma)$, that $\Sigma = I_{p\times p}$, and that $R = (r_{ij}) = (s_{ij}/\sqrt{s_{ii}s_{jj}})$. Let $J = \{(i_k,j_k)\}$ be a subset of the indices $\{(i,j),\ 1 \leq i < j \leq p\}$ which has been ordered so that for each k, at most one of the pair (i_k,j_k) appears as an i or a j earlier in the list. Then the statistics $r_{i_kj_k}$, $k = 1,2,\ldots$ are statistically independent.

A proof can be constructed by observing that the r_{ij}'s are cosines of angles among p independent spherically distributed random vectors in n-dimensional space. Then, by first conditioning on the variables $X_{i_1}, X_{j_1}, \ldots, X_{i_{k-1}}, X_{j_{k-1}}$, each $r_{i_kj_k}$ can be shown to be independent of $r_{i_1j_1}, \ldots, r_{i_{k-1}j_{k-1}}$.

The set J in this lemma can be equivalently defined in graph theoretic terms as follows: J has the property that the directed graph with nodes numbered 1,2,...,p, and an edge connecting node i_k to node j_k for each (i_k, j_k) in J, has no closed loops.

Naturally, if a transformation like $z(\cdot)$ normalizes the marginal distributions of two variables that are statistically independent, it normalizes their joint distribution. In the present case, although the z_{ij} are not independent as an entire group, all pairs, most triples, and many subsets of up to p-1 of the z_{ij} are. This is strong evidence to support the view that the multinormal distribution closely approximates the joint distribution of the z_{ij} when the ρ_{ij} are all zero.

5.3.4. Next take the case of a set of conditional correlations all with the same conditioning set, b. If a is the complement of the set b, and if S is distributed as $W_p(n,\Sigma)$, then $S_{aa.b}$ is distributed as $W_\alpha(n-\beta, \Sigma_{aa.b})$, where α and β are the numbers of indices in a and b. It follows that Eqs. 5.3 to 5.6 can be used to obtain asymptotic expressions for the variances and covariances of the $z_{ij.b}$'s provided that n, r_{ij} and ρ_{ij} are replaced by $(n-\beta)$, $r_{ij.b}$ and $\rho_{ij.b}$, respectively.

5.3.5. Consider a pair of partial correlations $r_{ij.a}$ and $r_{k\ell.b}$ with the property that the indices i,j and the conditioning set a are all contained in the set b. Then $r_{ij.a}$

and $r_{k\ell \cdot b}$ are statistically independent regardless of the values of the population correlations. This follows from a special property of the Wishart distribution (Dempster, 1969, p. 297), namely, that if S has a Wishart distribution, b and c are complementary sets of indices, and $S_{cc.b} = S_{cc} - S_{cb}S_{bb}^{-1}S_{bc}$, then S_{bb} and $S_{cc \cdot b}$ are statistically independent. In the case under consideration, $r_{ij.a}$ is a function only of the elements of S_{bb}, and $r_{k\ell .b}$ is a function only of the elements of $S_{cc.b}$. Hence, they are independent. The asymptotic covariance formula (Eq. 5.2), can be shown to give a zero covariance in this case, as it must do.

5.3.6. Finally, consider a case that is rather different from the rest but has some practical importance. Suppose that a certain fraction of the data is missing on some or all of the variables $X_1, \ldots, X_p$. If the mechanism which causes the data to be missing is statistically independent of the data values, then consistent estimates of the σ_{ij} and other parameters derived from them can still be obtained.

Suppose that $\overline{X}_i'$ denotes the mean of variable X_i calculated over the n_i observations available for that variable, and that s_{ij}' denotes the covariance calculated over the n_{ij} observations for which X_i and X_j are both present, (other definitions are possible). Thus the elements of the matrix $S' = (s_{ij}')$ are based on different, usually overlapping, subsets of the data. Conditionally, given the observed numbers and patterns of missing values, the statistics s'_{ij} are asymptotically jointly normal, as are the elements of the matrices R′ and Z′ formed from S′, but the variances and covariances of these statistics are no longer those obtained from the Wishart distribution. Instead, the covariance of s'_{ij} and $s'_{k\ell}$ is proportional to the number, $n_{ijk\ell}$, of observations present on all four variables. In particular, ignoring some small terms due to the particular definition of $\overline{X}'_i$, we have

$$\operatorname{cov}(s'_{ij}, s'_{k\ell}) = \frac{n_{ijk\ell}}{n_{ij}n_{k\ell}} (\sigma_{ik}\sigma_{j\ell}+\sigma_{i\ell}\sigma_{jk}) \qquad (5.7)$$

(5.7)

$$= \frac{nn_{ijk\ell}}{n_{ij}n_{k\ell}} \operatorname{cov}(s_{ij}, s_{k\ell}),$$

where any of the indices i,j,k,ℓ may be equal to each other, where s_{ij} and $s_{k\ell}$ represent covariances based on complete data (if it were available), and where, for example, $n_{112} = n_{12}$.

Equation 5.7 can be used in turn to obtain modified formulas for the asymptotic variances and covariances of the r'_{ij} and z'_{ij}. The new formulas resemble Eqs. 5.3 to 5.6, but each term has an additional factor of the form $(nn_{ijk\ell}/n_{ij}n_{k\ell})$. These formulas serve to show the general way in which the covariance structure of the z_{ij} is altered when data is missing, but since the effect of a small fraction of missing data on the correlations of the z_{ij} is small, and since we are dealing only with asymptotic approximations, it will not be worthwhile to do the extra calculations in most cases. We note that to use the modified formulas one must perform the additional task of tabulating the numbers n_i, n_{ij}, n_{ijk}, and $n_{ijk\ell}$ while the s'_{ij} are being calculated. For a problem with many variables this might require a prohibitive amount of additional computer storage, especially for $\{n_{ijk}\}$ and $\{n_{ijk\ell}\}$. An often useful compromise will be to tabulate only $\{n_i\}$ and $\{n_{ij}\}$, and to adjust only the standard errors of the z_{ij}. This will be illustrated in the example in Section 5.4.

5.4 Model Selection and Assessment

Assuming now that the application of z or some other transformation to a set of simple (or partial) correlation statistics renders their joint distribution approximately normal, we consider how these transformed statistics might be employed in the study of structured Wishart models. The paragraphs below outline three stages of such a study in which they are potentially useful.

5.4.1 Data Dependent Model Selection

Suppose that one is interested in Wishart models whose essential structure is preserved under the chosen transformation, as in for example the models of Section 3.2 that require some set of correlations or partial correlations to vanish. If it is assumed that such a class of models is appropriate, but it is not known in advance which member of the class to choose, the selection might be based on the observed magnitudes of the statistics $z_{ij.a}$. This is the course followed by Hills and by Fisher in seeking a small number of departures from $\rho_{ij} = 0$, but the approach can be applied to other problems such as the simultaneous-equation subset regression problem of Ex. 3.5, and Dempster's covariance selection model in Ex. 3.8 and Section 5.3.2. In the latter case, in which a pattern of zeros in Σ^{-1} is sought, the appropriate test statistics are $z^*_{ij} = z(r_{ij.c})$, where $c = \{1,2,\ldots,p\} - \{i,j\}$, and their empirical distribution should be compared to $N(0,1/(n-p))$, perhaps using a probability plot.

5.4.2 Model Fitting

If as in the last paragraph the model displays a simple structure after the normalizing transformation is applied, but if certain parameters must be estimated, the estimation might sensibly be done on the transformed scale, especially if the structure is linear on that scale. An illustration is provided by Example 3.1 in which correlations are assumed to equal each other within blocks. After transforming $\{r_{ij}\}$ to $\{z_{ij}\}$, one can apply usual normal-theory estimation procedures. If one chooses to ignore the correlation structure of the z_{ij}, this leads to calculating simple averages of the z_{ij}'s in blocks. Or, if the correlation structure is to be taken into account, then the set of z_{ij}'s should be weighted by the inverse of their estimated covariance matrix. Since the "correct" weights depend on the parameters being estimated, this could be done iteratively, perhaps stopping after two steps. In either case, the resulting point or interval estimates of $z(\rho_{ij})$ may be back-transformed to give estimates on the original correlation scale.

There are some situations in which the ρ_{ij} might be related to other quantitative variables also indexed by i and j. For instance, it might be appropriate to regress $z(r_{ij})$ linearly on certain other variables, suitably re-expressed.

Whether or not the fitting of structured covariance models on a z-transformed scale produces asymptotically efficient estimates is a matter for further study. We note that the procedure does provide an opportunity to use robust estimates, since averages of z_{ij}'s can easily be replaced by robust estimates of location (say, α-trimmed means). However, this should not be thought of as protection against individual outliers in the original data (which can be treated by robustly calculating the correlation matrix itself, as in Devlin, et al., 1975); rather it is protection against isolated discrepancies between the true and hypothesized models.

There is a connection here with the results of Chapter 2. Invariance arguments show that, in a structured-correlation model, maximum likelihood estimates of ρ_{ij} depend only on the sample correlation matrix R. Therefore estimation can proceed in two stages; first the correlations are estimated from R, then variances from s_{ii} and $(\hat{\rho}_{ij})$. If maximum likelihood in the first stage is replaced by a z-based procedure for expedience, but retained in the second stage where the number of remaining parameters is much smaller, then the property $\mathrm{tr}(\hat{\Sigma}^{-1}S) = p$ still holds.

5.4.3 Assessment of Residuals

Once a structured covariance model has been fitted, however the fitting was done, it may be useful to asses the quality of fit by examining residual correlations on a transformed scale. Let $\hat{\rho}_{ij}$ represent a correlation estimated by maximum likelihood (or by some other method). The statistics $e_{ij} = z(r_{ij}) - z(\hat{\rho}_{ij})$, or, generally, $e_{ij.a} = z(r_{ij.a}) - z(\hat{\rho}_{ij.a})$, can be regarded as residuals. If the original data is normally distributed and the chosen structured model is adequate, then the e_{ij} should be approximately jointly normally distributed with zero means and known variances, and a covariance structure that can be estimated from the matrix $(\hat{\rho}_{ij})$ using the asymptotic formulas in Eqs. 5.3 - 5.6 . One can construct a formal χ^2 statistic from such a set of residuals, but their real value will often lie in their use as a diagnostic aid to judge not only whether, but how the model fails to fit. By sorting and plotting these residuals against expected normal order statistics, one can explore to what extent they resemble a sample from an unstructured normal population, or, more precisely, from a normal population with a certain correlation structure. In particular, one might be able to isolate pairs or groups of the variables X_i whose correlations don't conform to the hypothesized pattern. This will be illustrated in Example 5.1.

There is a connection between this approach and the ideas developed by Wald (1943) and by Aitchison and Silvey (1960). Those authors propose to test a general model whose vector parameter θ must satisfy some restrictions, $h(\theta) = (h_1, \ldots, h_r) = 0$. They first compute $h(\hat{\theta})$, where $\hat{\theta}$ is the unrestricted maximum likelihood estimate, and then calculate a χ^2 statistic W from $h(\hat{\theta})$ and its estimated asymptotic covariance matrix. If their functions $h_k(\theta)$ are identified with $z(\rho_{ij})$, then their W statistic is the χ^2 statistic mentioned above. The present approach puts less emphasis on formal testing and more on the heuristic use of the residuals. Also by working in a more restricted framework, that of structured Wishart models, we are able to find a re-expression of the parameter space (namely, $\underset{\sim}{\zeta} = z(\underset{\sim}{\rho})$) which improves the distributional approximation for finite sample sizes.

We note that the points made in Section 5.4.1 and 5.4.3 above are quite closely connected, since it is usually by fitting "null" models and detecting systematic departures from them that more refined models are developed.

We also consider briefly the question of using normal probability plots, which assume independent observations, to study statistics that are correlated, as suggested in Sections 5.4.1 and 5.4.3. The issue has not received much attention in the statistical literature, although some authors have suggested that mild correlation should have little effect (e.g. Cox and Lauh, 1967). The main result of practical importance was mentioned briefly by Mallows (1969), namely, that if $X_1, X_2, \ldots, X_n$ are marginally and pair-wise normal, all with equal means and variances, and with correlations $\rho_{ij} = \text{corr}(X_i, X_j)$, then to a certain approximation the order statistics behave as though all correlations were equal to their average,

$$\bar{\rho} = \frac{2}{n(n-1)} \sum_{i<j} \rho_{ij} \,.$$

We observe further that the order statistics from an equicorrelated sample are like those from an independent sample, but with a modified variance and a randomly shifted sample mean. In particular, if $X_1, \ldots, X_n$ are jointly normal with mean zero, variance σ^2, and $\text{corr}(X_i, X_j) = \bar{\rho}$, then the sample configuration statistics

$$(X_1 - \bar{X}), \ldots, (X_n - \bar{X})$$

are like those from a set of independently normally distributed variables $Y_1, \ldots, Y_n$ with variance $\sigma^2(1-\bar{\rho})$. This means that a normal probability plot of the X's will on average be a straight line with slope $\sigma\sqrt{1-\bar{\rho}}$, but it will be subjected to a random vertical displacement with variance $\mathrm{var}(\bar{X}) = \frac{\sigma^2(1-\bar{\rho})}{n} + \sigma^2\bar{\rho}$. By contrast, the vertical displacement of a probability plot of the Y's has variance $\mathrm{var}(\bar{Y}) = \frac{\sigma^2(1-\bar{\rho})}{n}$. D. R. Cox (1975) has suggested that the quantity

$$\sigma_0 = \sigma\sqrt{1-\bar{\rho}} \tag{5.8}$$

be called the "effective standard deviation" because its square is on average the variance within samples from the correlated distribution.

From the arguments outlined here, and from studying a series of probability plots of correlated normal observations simulated by computer, we feel justified in using probability plots of z-transformed residual correlations in the study of structured Wishart models, bearing in mind that the slopes of these plots are affected by the correlations. This viewpoint is further strengthened by the fact that Mallows' approximation improves when the correlations are close to their average; when dealing with the statistics z_{ij} the correlations obtained from Eqs. 5.3 to 5.6 are typically more moderate than the ρ_{ij}'s themselves.

Example 5.1 - AT&T engineering and service indicators

To illustrate some of the points made above, we consider a set of data recently studied by Fowlkes (1975). The data consist of 66 variables which describe various characteristics of 61 geographical districts of several telephone operating companies in the Bell System. The variables apply to a single month and include such things as total number of customers, total expenditures of various kinds, numbers of new orders received and completed, quantity of equipment currently in use, number of customer trouble reports, etc. The study had the rather general objectives of summarizing the data, discovering underlying relationships, and exposing interesting and peculiar features.

We note that because of the large number of variables involved, this problem is beyond the capabilities of the iterative maximum likelihood methods developed in Ch. 4, so we employ analyses based on the approximate normality of the statistics $z(r_{ij})$.

After some preliminary analysis which suggested transformations of certain variables to improve marginal and pairwise distributions, Fowlkes computed a 66×66 correlation matrix R for variables across the 61 districts. Since a significant proportion of data was missing, each correlation was computed using all the data available for that pair of variables. The number of observation pairs entering each correlation ranged from 6 to 61, and averaged 38.

One of the analyses performed by Fowlkes was a hierarchical clustering of the variables using absolute value of correlation as a measure of similarity. Starting with Fowlkes' correlation matrix we consider a clustering with a different objective: in order to discover underlying symmetries among the variables, we seek to determine how well the 66×66 correlation matrix or some subset of it can be fit by a constant-in-blocks structure, of the kind described in Ex. 3.1. If such a structure can be found, then the variables entering each diagonal block form a symmetrical set, that is, in terms of their correlations with other variables, they are interchangeable.

An appropriate distance measure for this purpose is derived from the correlation matrix in the following manner. If two variables, say X_1 and X_2, are interchangeable, then pairs of correlations in the first two columns of R have identical expectations, as do their z-transforms. Hence, if

$$d_i = z_{i1} - z_{i2} ,$$

then

$$E(d_i) = 0, \quad i = 3,4,\ldots,66 . \qquad (5.9)$$

(In fact, Eq. 5.9 holds exactly when the numbers of observations n_{i1} and n_{i2} entering r_{i1} and r_{i2}, respectively, are equal, but only to order $O(\frac{1}{n_{i1}} - \frac{1}{n_{i2}})$ otherwise.) Since the d_i are approximately normally distributed with zero mean, they can be used to form a statistic

$$t_{12} = \sum_{i=3}^{66} d_i^2/w_i \ , \qquad (5.10)$$

which is an index of dissimilarity for the variables X_1 and X_2: a high value of t_{12} results when $E(d_i) \neq 0$ and X_1 and X_2 are not interchangeable. The quantities w_i are weights which should be proportional to the variances of the d_i. To obtain empirical weights, we ignored the covariance of z_{i1} and z_{i2}, but took account of the missing data. Hence,

$$w_i = \text{var}(z_{i1}) + \text{var}(z_{i2})$$

$$= \frac{1}{n_{i1}-3} + \frac{1}{n_{i2}-3} \quad .$$

(In fact, not only could the covariances be used and estimated as in Section 5.3.6, but the covariances of the d_i could be considered as well. Then t_{12} would be defined as $\underset{\sim}{d}^T W^{-1} \underset{\sim}{d}$, where W is a 64x64 matrix. But to do this for each of $\binom{66}{2}$ pairs of variables would be a prohibitively expensive calculation.)

The fact that the t_{ij} and r_{ij} provide different bases for clustering is shown by Fig. 5.1 in which t_{ij} is plotted against $z(r_{ij})$ for the 400 smallest values of t_{ij}. Bearing in mind that small values of t_{ij} correspond to greater interchangeability, one can clearly see that some pairs of variables are highly correlated but not highly interchangeable, and conversely.

A hierarchical clustering (Johnson, 1967; Warner, 1969) was performed using t_{ij} as a distance measure, and the results are displayed as a tree in Fig. 5.2. The identification numbers of the variables appear at the top, and the numbers on the left are average values of t_{ij} within the least compact cluster defined at each level of the tree.

The statistics t_{ij} are roughly distributed as χ^2_{64} for interchangeable pairs of variables. It seemed reasonable to cut the tree at the arbitrary level 8 which corresponds approximately to the .95 quantile of the χ^2_{64} distribution. This defined 18 clusters whose sizes range from 1 to 9 variables, of which seven clusters containing 5 or more variables account for 42 of the total, as shown in Table 5.1.

By rearranging rows and columns, we partition the corresponding 42x42 submatrix of z_{ij}'s into a 7X7 configuration of blocks, and fit a constant $\overline{z}_{ij}$ within each block. Again ignoring the correlations among the z_{ij}, we use weights that reflect the missing data, and trimmed averages to guard against

FIGURE 5.1 - Dissimilarity index vs. correlation for 400 pairs of variables

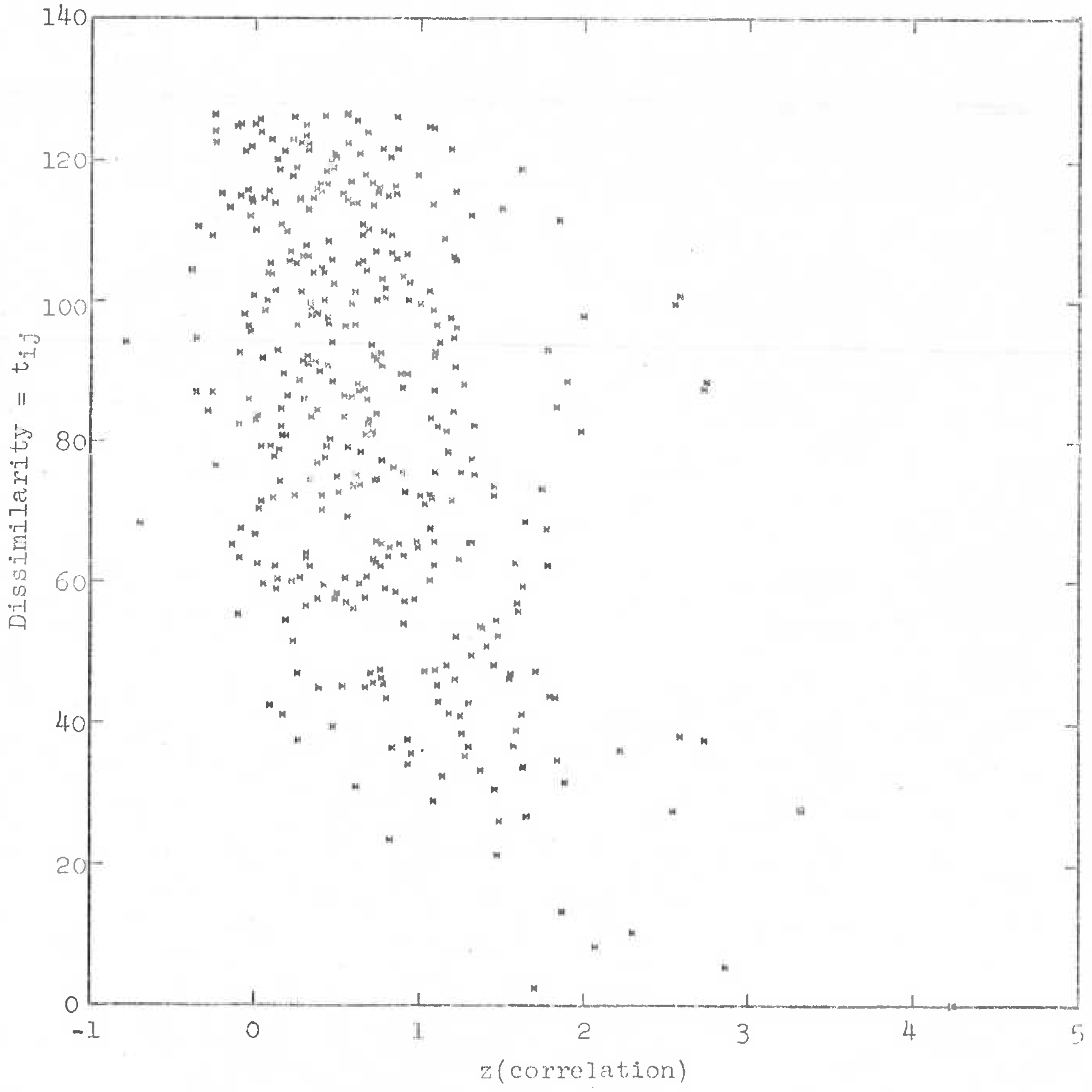


FIGURE 5.2 - Clustering of 66 variables based on dissimilarity index

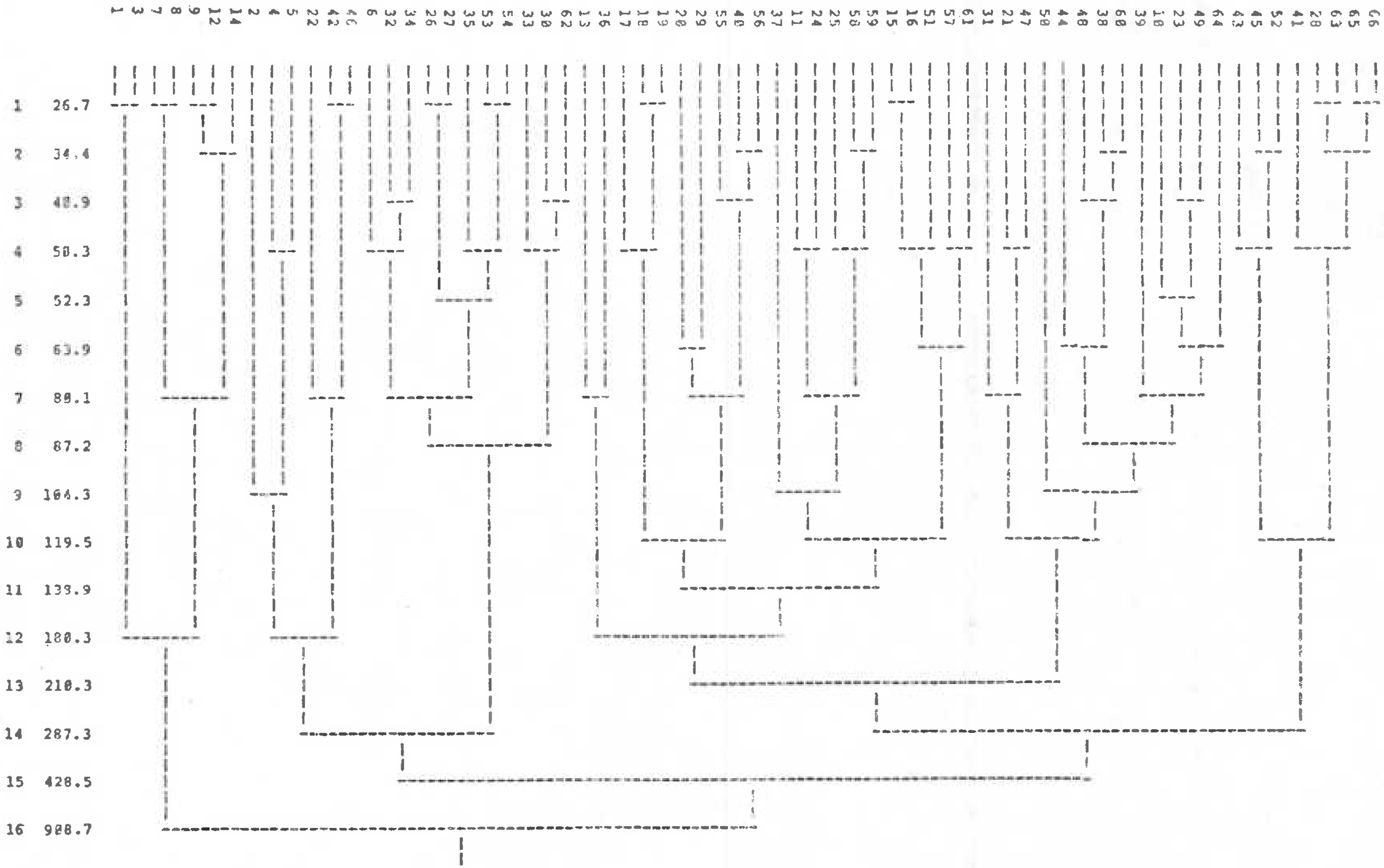


TABLE 5.1 - Clusters formed by cutting the tree at level 10

Cluster	Variables								
A	7	8	9	12	14				
B	6	32	34	26	27	35	53	54	
C	20	29	55	40	56				
D	11	24	25	58	59				
E	15	16	51	57	61				
F	44	48	38	60	39	10	23	49	64
G	41	28	63	65	66				

a few outliers. Table 5.2 shows the fitted $\bar{z}_{ij}$'s with their standard errors, and Table 5.3 the corresponding $r(\bar{z}_{ij})$'s.

Whether this fitted block correlation structure, which is completely described by 28 distinct correlation parameters, provides an adequate summary of the 42×42 correlation matrix with 861 distinct entries, is a question that can be answered in part by examining the residuals from the fit. The normal probability plot of the 861 residuals in Fig. 5.3 shows an appreciable curvature at the ends, suggesting a poor fit or longer-than-Gaussian tails. However, since missing data has caused the z_{ij} to have different variances, it seems more appropriate to standardize each z_{ij} by dividing by its standard error, $(n_{ij}-3)^{-\frac{1}{2}}$. Figure 5.4 is a probability plot of the standardized residuals with a superimposed line of unit slope and zero intercept which is the average configuration when the model assumptions are all satisfied. Except for about 15 extreme points out of 861, the empirical distribution of residuals agrees remarkably well with a normal configuration, suggesting that a block correlation structure is a reasonable model for a vast majority of this data. The improvement of linearity of Fig. 5.4 over Fig. 5.3 gives added support to the recommendation in Section 5.3.6 that a weighted analysis be done when a substantial amount of data is missing.

We note that this particular 7×7 block structure hypothesis cannot be formally tested with this set of data because the statistical significance level would be artifically

TABLE 5.2 - Average z_{ij} within blocks, and standard errors

	A	B	C	D	E	F	G
A	2.815 (0.046)	1.067 (0.026)	0.718 (0.031)	0.363 (0.034)	0.430 (0.041)	0.089 (0.030)	0.240 (0.031)
B	1.067 (0.026)	1.385 (0.035)	0.752 (0.028)	0.582 (0.030)	0.747 (0.035)	0.173 (0.025)	0.664 (0.028)
C	0.718 (0.031)	0.752 (0.028)	0.780 (0.054)	0.420 (0.036)	0.475 (0.043)	0.096 (0.031)	0.287 (0.032)
D	0.363 (0.034)	0.582 (0.030)	0.420 (0.036)	0.531 (0.062)	0.364 (0.048)	0.114 (0.034)	0.304 (0.035)
E	0.430 (0.041)	0.747 (0.035)	0.475 (0.043)	0.364 (0.048)	0.733 (0.089)	0.197 (0.041)	0.744 (0.042)
F	0.089 (0.030)	0.173 (0.025)	0.096 (0.031)	0.114 (0.034)	0.197 (0.041)	0.100 (0.045)	0.243 (0.031)
G	0.240 (0.031)	0.664 (0.028)	0.287 (0.032)	0.304 (0.035)	0.744 (0.042)	0.243 (0.031)	1.406 (0.054)

TABLE 5.3 - Fitted correlations derived from $\bar{z}_{ij}$

	A	B	C	D	E	F	G
A	0.993	0.788	0.616	0.348	0.405	0.089	0.236
B	0.788	0.882	0.637	0.524	0.634	0.171	0.581
C	0.616	0.637	0.653	0.397	0.442	0.096	0.279
D	0.348	0.524	0.397	0.486	0.349	0.114	0.295
E	0.405	0.634	0.442	0.349	0.625	0.195	0.632
F	0.089	0.171	0.096	0.114	0.195	0.099	0.238
G	0.236	0.581	0.279	0.295	0.632	0.238	0.887

FIGURE 5.3 - Normal probability plot of raw residuals

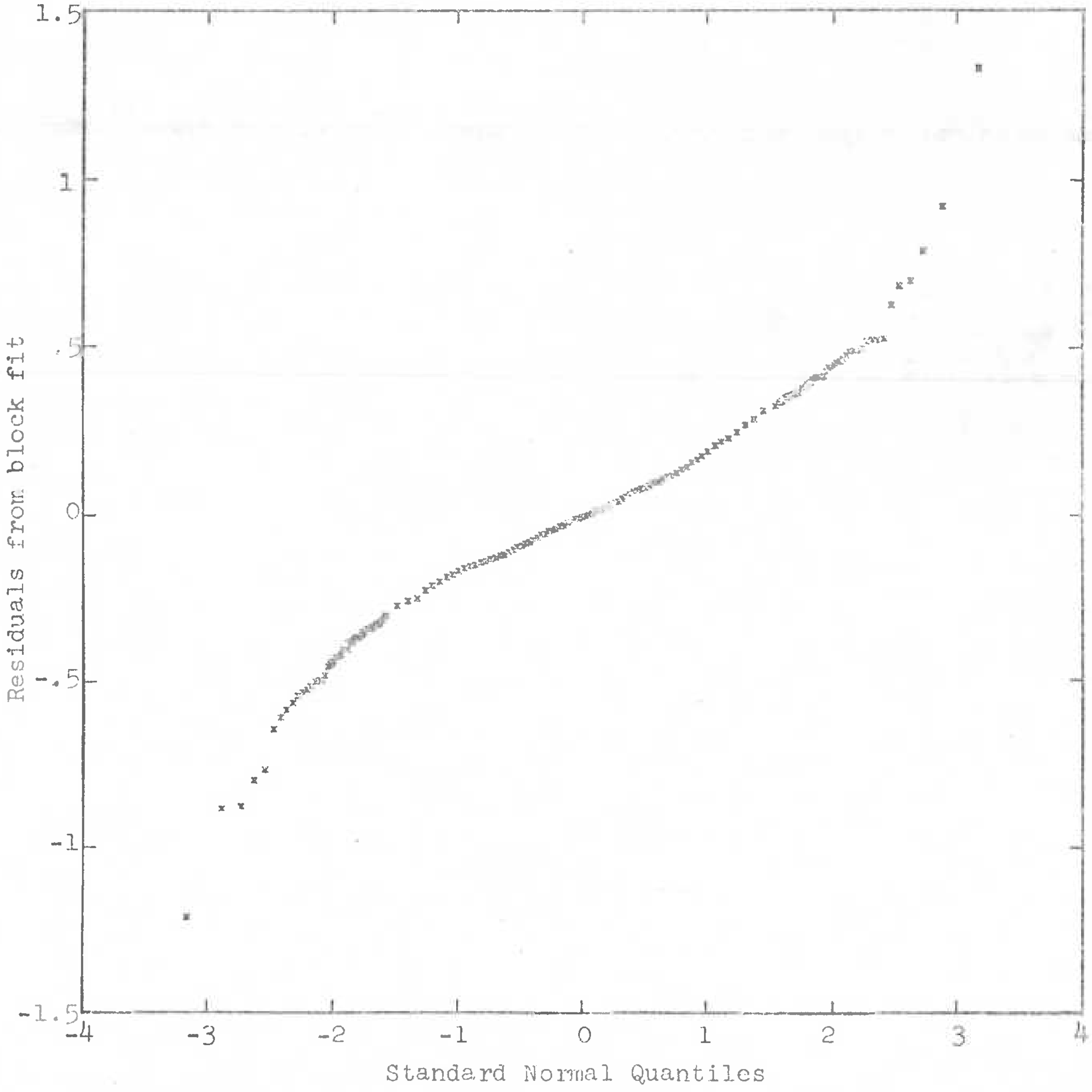
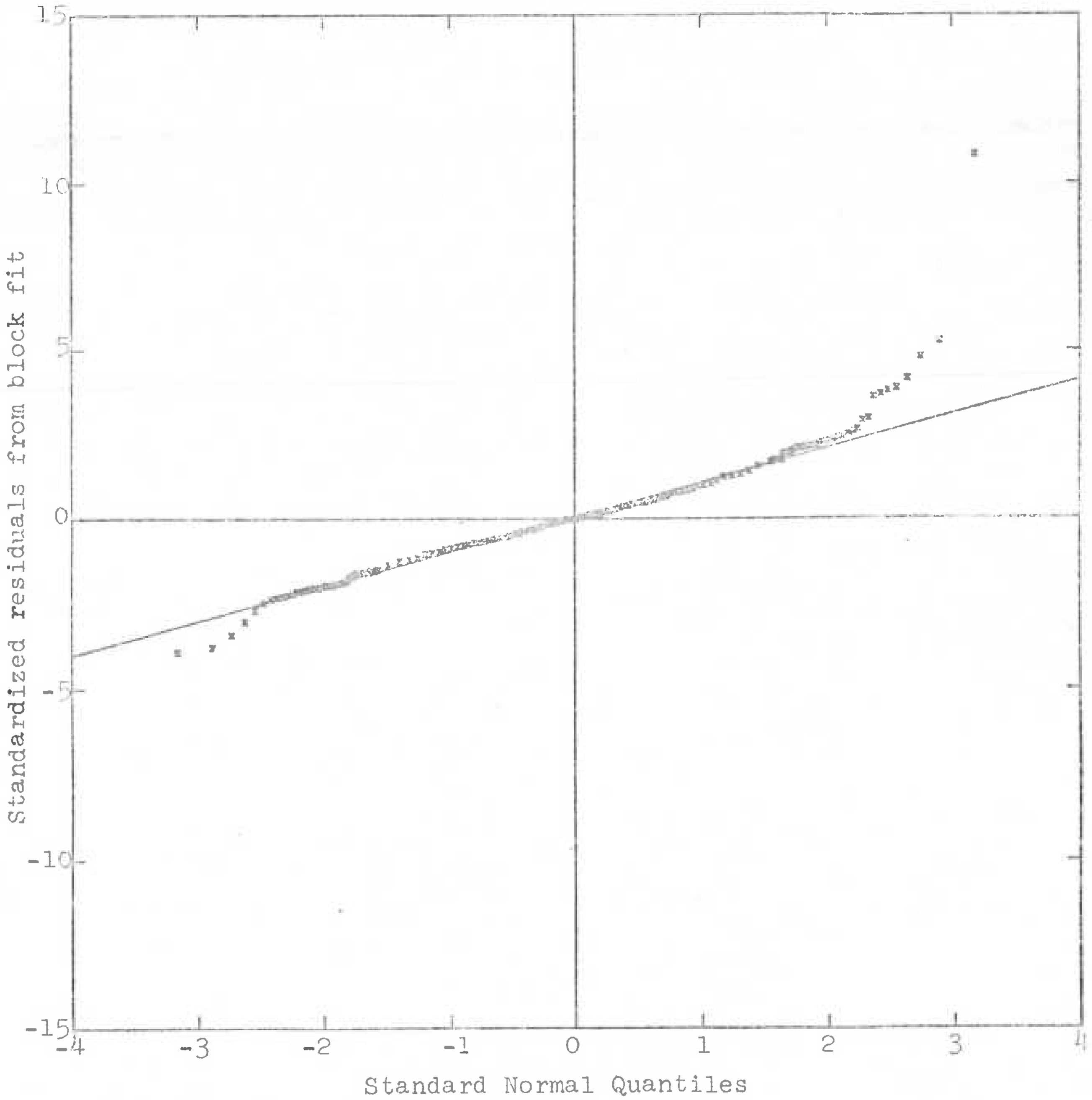


FIGURE 5.4 - Normal probability plot of standardized residuals



inflated from having used the same data in the clustering stage to select the model. An analysis of this kind is intended to be informal in any event, and to form the basis for further examination of the data. The few large residuals, for example, draw attention to pairs of variables that require closer scrutiny, and one would want to judge whether the clusters in Table 5.1 correspond to meaningful groupings in terms of the definitions of the variables.

These questions go beyond the scope of the present example, but we make one final point concerning the degree to which the block structure model adequately fits this set of data. The single probability plot of Fig. 5.4 is only a very rough guide. Ideally the residuals, which are shown in Table 5.4, should have no discernable structure beyond that imposed by the fitting process itself, but clearly there is some additional structure. Table 5.5 shows where the residuals of greatest magnitude fall, with residuals indicated by rank and negative values denoted by underscores. Since disproportionately many large residuals fall in diagonal blocks, we conclude that this particular block structure is more satisfactory for modelling correlations across rather than within groups. Also, we note that group F is rather different from the rest in that the within-group fitted correlation is small (.099), and it contains variables such as 39 and 49 which have excessively

TABLE 5.4 - Standardized residuals (x100) from block fit

	7	8	9	12	14	6	26	27	32	34	35	53	54	20	29	40	55	56	11	24	25	58	59	15	16	51	57	61	10	23	38	39	44	48	49	60	64	28	41	63	65	66
7		-391	-199	-101	-59	-7	-50	-10	138	-49	-93	-14	32	20	46	-53	69	-75	70	-75	15	-31	-53	-22	2	-10	-66	62	97	-46	48	19	-97	-55	-68	70	50	-100	83	-3	-10	15
8	-391		-51	-65	-205	65	-22	4	143	53	-82	18	92	8	38	-82	47	3	97	-30	71	-32	-16	-4	12	21	-41	61	102	-23	59	1	-46	-72	-55	25	64	-61	89	0	-8	26
9	-199	-51		1000	388	-5	-42	-10	108	-7	-123	-48	16	14	49	-73	72	-33	70	-61	46	-34	-39	-15	4	6	-56	56	90	-35	20	8	-85	-67	-57	33	49	-94	65	-5	-25	11
12	-101	-65	1000		385	3	-47	-13	111	-12	-123	-44	22	12	44	-82	67	-43	57	-70	44	-39	-36	-9	10	11	-54	59	80	-30	34	1	-85	-59	-58	35	47	-94	60	-5	-25	0
14	-59	-205	388	385		14	-26	10	135	-12	-98	-2	57	0	43	-79	58	-48	81	-69	47	-17	-29	0	24	3	-44	70	108	-32	36	6	-88	-60	-61	45	24	-69	97	18	-8	25
6	-7	65	-5	3	14		-126	-66	-127	259	-140	-270	-159	123	-82	12	-28	17	-143	-173	9	-104	-63	9	26	59	-174	36	-78	-87	125	-138	-7	-85	-86	23	-42	-160	-114	-76	-201	-197
26	-50	-22	-42	-47	-26	-126		155	-30	-121	478	222	117	-54	-84	90	140	227	-51	-6	142	171	127	-100	-63	-36	248	77	43	-18	-10	-12	173	61	-119	-83	52	101	214	43	-2	20
27	-10	4	-10	-13	10	-66	155		-45	-47	135	232	131	-98	-216	7	39	127	-90	-35	104	123	207	-57	-30	112	264	6	30	-56	16	-64	100	34	-55	-110	67	50	167	70	-1	-2
32	139	143	108	111	135	-127	-30	-45		-76	-133	52	99	13	-206	-105	47	-32	-112	-91	50	56	46	-97	-47	-69	52	11	44	-33	-14	-73	114	-96	-46	-23	30	-47	61	-86	-124	-50
34	-49	53	-7	-12	-12	259	-121	-47	-76		-170	30	40	42	-165	-4	-20	79	-52	-32	226	91	130	-50	-49	51	206	17	10	25	110	-116	99	-146	-23	-78	-60	28	20	-45	-55	-116
35	-93	-82	-123	-123	-90	-140	478	135	-133	-170		-154	20	-58	7	354	112	84	-151	-90	22	-28	8	-27	-68	-22	-30	147	14	-73	-32	0	55	10	-199	-1	79	-1	99	42	11	32
53	-14	18	-48	-44	-2	-270	222	232	52	30	-154		367	-74	-23	-149	-52	93	9	-77	-5	95	-39	-18	-36	-47	14	130	85	81	45	103	21	-82	-122	49	20	-48	196	58	-91	1
54	32	92	16	22	57	-158	117	131	99	40	20	367		-23	-36	-31	-8	54	-76	-72	35	81	20	-26	-45	-3	21	120	106	26	129	79	44	-48	-126	58	-3	-7	237	32	5	94
20	20	8	14	12	0	123	-54	-98	13	42	-38	-74	-23		-238	-60	-96	-137	14	-127	121	9	-15	-16	-37	-68	0	-16	3	-123	34	-124	-71	-8	-47	15	47	-68	32	-71	-38	-54
29	45	38	49	44	43	-82	-84	-216	-206	-165	7	-23	-36	-218		47	16	-32	159	17	-42	-189	-231	-124	-90	-163	-162	127	49	93	-57	152	-113	-121	-104	92	172	-61	20	-31	-59	-72
40	-53	-82	-73	-82	-79	12	90	7	-105	-4	154	-149	-31	-88	47		231	216	28	-104	-12	31	-57	84	45	40	-9	97	62	-37	-43	9	-21	-19	-101	-29	78	79	60	173	1	30
55	69	47	72	67	58	-28	140	39	47	-20	112	-52	-8	-96	16	231		91	-96	-8	-10	1	-134	36	49	35	39	-21	77	0	124	1	38	57	-137	-4	-118	-4	109	5	-21	53
56	-75	3	-33	-43	-48	17	227	127	-32	79	84	93	54	-137	-32	216	91		84	108	172	125	21	-41	-58	70	134	27	83	-10	-31	79	87	-10	-104	-38	123	49	37	44	-53	0
11	70	97	78	57	81	-143	-51	-90	-112	-52	-151	9	-76	14	159	28	-96	84		0	-25	-163	-189	-73	-56	-228	-110	-71	59	72	-130	128	-62	-199	-52	-115	52	-130	-17	-97	-132	-116
24	-75	-10	-61	-70	-69	-173	-6	-35	-91	-32	-80	-77	-72	-127	17	-104	-8	108	0		159	-89	-136	-95	-102	-207	2	-2	119	20	-73	165	209	37	-49	-163	19	-9	22	-158	-53	-70
25	15	71	46	44	47	9	142	104	50	226	22	-5	35	121	-42	-12	-10	172	-25	159		127	135	-15	-32	-156	394	35	99	-87	-188	-43	190	-18	-53	-136	58	78	48	-4	-9	-89
58	-31	-32	-34	-39	-17	-104	171	123	56	91	-28	95	81	9	-189	31	1	125	-163	-89	127		169	-26	18	47	413	97	42	-204	89	74	158	42	-162	44	41	207	251	149	77	115
59	-58	-16	-39	-36	-29	-63	127	207	46	138	8	-39	20	-15	-231	-57	-134	21	-169	-136	135	169		120	103	97	291	164	-8	-155	0	-63	144	-10	71	-39	-43	92	101	193	-68	-107
15	-22	-4	-15	-9	0	9	-100	-57	-97	-58	-27	-18	-26	-16	-124	84	36	-41	-73	-95	-15	-26	120		525	-6	-69	13	-63	-75	4	-300	14	78	2	-1	-64	-101	-72	-27	-78	-69
16	2	12	4	10	24	26	-63	-10	-47	-49	-68	-36	-46	-37	-90	45	49	-58	-56	-102	-12	18	103	525		-12	-86	10	-72	-51	20	-342	66	80	44	50	-8	-108	-99	9	-112	-39
51	-10	21	6	11	3	59	-36	112	-69	51	-22	-47	-3	-66	-163	40	35	70	-228	-207	-156	47	97	-6	-12		47	-123	5	20	50	-375	-87	163	-49	-71	-47	116	-42	6	-192	49
57	-66	-41	-56	-54	-44	-174	248	264	52	206	-30	14	21	0	-162	-9	39	134	-110	2	194	413	291	-69	-86	47		122	76	-78	190	-86	130	68	-91	83	12	216	204	164	-35	139
61	62	51	56	59	70	36	77	6	11	17	147	130	120	-16	127	97	-21	27	-71	-2	35	97	164	13	10	-123	122		-69	51	55	-116	84	78	12	214	-89	-7	160	107	-33	-13
10	97	102	90	80	108	-78	43	30	44	10	14	85	106	3	49	62	77	83	59	119	99	42	-8	-63	-72	5	76	-69		-4	-74	-92	50	94	7	-81	-115	-82	-111	-111	-249	-235
23	-46	-23	-35	-30	-32	-87	-18	-56	-33	25	-73	81	20	-123	93	-37	0	-10	72	20	-87	-204	-155	-75	-51	20	-72	51	-4		101	125	92	2	41	-31	-79	-14	-52	-55	-140	-85
18	48	59	20	34	36	125	-19	16	-14	110	-32	45	129	34	-57	-41	124	-31	-130	-73	-188	89	0	4	20	50	190	55	-74	101		-49	296	29	5	199	-156	193	208	44	241	225
39	19	1	0	1	6	-138	-12	-64	-73	-116	0	103	79	-124	152	9	1	79	128	165	-43	74	-53	-300	-342	-375	-86	-116	-92	125	-49		206	-217	-196	-160	-142	-140	25	-211	-1	-29
44	-97	-46	-85	-85	-88	-7	173	180	114	99	55	21	44	-71	-113	-21	38	87	-62	209	190	158	144	14	66	-67	130	84	50	92	296	206		64	-56	25	-3	131	44	-3	81	16
48	-55	-72	-67	-68	-60	-85	61	34	-96	-146	10	-82	-48	-8	-121	-19	57	-10	-199	37	-18	42	-15	78	80	163	68	70	94	2	29	-217	64		-16	122	9	150	82	125	36	124
49	-68	-55	-57	-58	-61	-86	-119	-65	-46	-73	-199	-122	-126	-47	-104	-101	-137	-104	-52	-49	-53	-162	71	2	44	-49	-91	12	7	41	5	-196	-56	-16		54	-15	-156	-222	-95	-194	-201
60	70	25	33	35	45	23	-83	-110	-23	-78	-1	49	58	15	92	-29	-4	-38	-115	-163	-136	44	-39	-1	50	-71	83	214	-81	-31	199	-160	25	122	54		125	112	101	239	172	210
64	50	64	49	47	24	-42	52	67	30	-60	79	20	-3	47	172	78	-118	123	52	19	56	41	-43	-64	-8	-47	12	-89	-115	-79	-156	-142	-3	9	-15	125		-21	-35	71	73	95
28	-100	-61	-94	-94	-69	-160	101	50	-47	28	-1	-48	-7	-68	-61	79	-4	48	-130	-9	70	207	92	-101	-109	116	216	-7	-82	-14	193	-140	131	150	-156	112	-21		29	47	-19	33
41	83	89	65	60	97	-114	214	167	81	28	99	196	237	32	20	60	109	37	-17	22	43	251	101	-72	-99	-42	204	160	-111	-52	208	25	44	82	-222	101	-35	29		-51	-31	362
63	-3	0	-5	-5	10	-76	43	70	-86	-45	42	58	32	-71	-31	173	5	44	-97	-158	-4	148	193	-27	9	6	164	107	-111	-55	44	-211	-3	125	-95	239	71	47	-51		-71	-62
65	-10	-9	-25	-25	-8	-201	-2	-1	-124	-66	11	-91	5	-39	-59	1	-21	-53	-132	-53	-9	77	-68	-78	-112	-192	-35	-33	-249	-140	241	-1	81	36	-194	172	73	-19	-31	-71		53
66	15	26	11	0	25	-197	20	-2	-50	-116	32	1	94	-54	-72	30	53	0	-116	-70	-89	119	-107	-69	-99	49	139	-13	-235	-85	225	-29	16	124	-201	210	95	33	162	-62	53	

TABLE 5.5 - Locations of largest residuals, by rank

	A	B	C	D	E	F	G
A	1,5 6,7	-	-	-	-	-	-
B		3,9 15,17	-	-	-	-	-
C				-	-	-	-
D					-	-	-
E		16		4,14	2	-	-
F					8,11 12	13	-
G						10	

many negative residuals. At a subsequent stage one might decide to exclude this group from the hypothesis of block structure, or to subdivide it in some appropriate way.

APPENDIX

A number of definitions and formulas are collected here to facilitate derivations in the text. Equations A1 through A10 refer to matrix derivatives, and the rest describe properties of a SWP operator similar but not identical to Beaton's (1964). All of these formulas can be verified easily from first principles.

A1. (Def.) If $f = f(X)$ is a scalar function of the elements of the matrix X, then $\frac{df}{dX}$ is defined to be the matrix $\left(\frac{\partial f}{\partial x_{ij}}\right)$.

A2. (Def.) If the elements of the matrix X are functions of the scalar y, then $\frac{dX}{dy}$ is defined to be the matrix $\left(\frac{\partial x_{ij}}{\partial y}\right)$.

A3. If the elements of the square matrix X are functions of the scalar y, then $\frac{dX^{-1}}{dy} = -X^{-1} \frac{dX}{dy} X^{-1}$.

A4. If the elements x^{ij} of the matrix X^{-1} are considered as functions of the elements of X, then $\frac{\partial x^{ij}}{\partial x_{k\ell}} = -x^{ik}x^{\ell j}$ and $\frac{\partial x_{ij}}{\partial x^{k\ell}} = -x_{ik}x_{\ell j}$.

A5. If f is a scalar function of the elements of X, then, $\frac{df}{dX^{-1}} = -X^{T} \frac{df}{dX} X^{T}$ and $\frac{df}{dX} = -(X^{T})^{-1} \frac{df}{dX^{-1}} (X^{T})^{-1}$.

A6. $\frac{d}{dX} \log \det(X) = (X^T)^{-1}$.

A7. $\frac{d}{dX} \mathrm{tr}(XA) = \frac{d}{dX} \mathrm{tr}(AX) = A^T$.

A8. $\frac{d}{dX} \mathrm{tr}(AXBX^T) = 2A^T XB^T$.

A9. (Chain rule) If the scalar y is a function of the elements of the matrix X which are in turn functions of a scalar z, then

$$\frac{dy}{dz} = \mathrm{tr}\left(\frac{dy}{dX}\,\frac{dX}{dz}^T\right) = \mathrm{tr}\left(\frac{dy}{dX^T}\,\frac{dX}{dz}\right) .$$

A10. If $f = f(x)$ is a scalar function of the elements of the vector x, and $x = Ay$, where A is a matrix of constants and the vector y may have fewer elements than x, then

$$\frac{df}{dy} = A^T \frac{df}{dx} .$$

A11. (Def.) If X is a square matrix then $Y = SWP(k)X$ is defined by

$$\begin{aligned}
y_{kk} &= -1/x_{kk} & & \\
y_{ik} &= x_{ik}/|x_{kk}| & &\text{for } i \neq k \\
y_{kj} &= x_{kj}/|x_{kk}| & &\text{for } j \neq k \\
y_{ij} &= x_{ij} - x_{ik}x_{kj}/x_{kk} & &\text{for } i \neq k,\ j \neq k.
\end{aligned}$$

This operator combines Beaton's SWP and RSW, and is its own inverse. If $x_{kk} > 0$, SWP coincides with Beaton's SWP; if $x_{kk} < 0$, it is equivalent to Beaton's RSW.

A12. If the covariance matrix Σ is partitioned as

$$\begin{array}{c} 1\ldots r\ldots p \\ \left[\begin{array}{c|c} \Sigma_{aa} & \Sigma_{ab} \\ \hline \Sigma_{ba} & \Sigma_{bb} \end{array}\right] \end{array},$$

then $\Sigma_{aa.b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$ is the conditional covariance of the variables $X_1,\ldots,X_r$ given $X_{r+1},\ldots,X_p$.

A13. $\mathrm{SWP}(a)\Sigma = \mathrm{SWP}(1,\ldots,r)\Sigma = \left[\begin{array}{c|c} -\Sigma_{aa}^{-1} & \Sigma_{aa}^{-1}\Sigma_{ab} \\ \hline \Sigma_{ba}\Sigma_{aa}^{-1} & \Sigma_{bb.a} \end{array}\right]$, and

$$\mathrm{SWP}(b)\Sigma = \left[\begin{array}{c|c} \Sigma_{aa.b} & \Sigma_{ab}\Sigma_{bb}^{-1} \\ \hline \Sigma_{bb}^{-1}\Sigma_{ba} & -\Sigma_{bb}^{-1} \end{array}\right].$$

A14. $\mathrm{SWP}(1,2,\ldots,p)\Sigma = -\Sigma^{-1}$.

A15. If a and b are complementary sets of indices then

$$\mathrm{SWP}(a)\Sigma = \mathrm{SWP}(b)(-\Sigma^{-1}).$$

REFERENCES

Aitchison, J. and Silvey, S.D. (1958). Maximum likelihood estimation of parameters subject to restraints. *Ann. Math. Statist. 29*, 813-828.

Aitchison, J. and Silvey, S.D. (1960). Maximum likelihood procedures and associated tests of significance. *J. Roy. Statist. Soc. B 22*, 154-171.

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

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

Arnold, S.F. (1973). Application of the theory of products of problems to certain patterned covariance matrices. *Ann. Statist. 1*, 1-18.

Beaton, A.E. (1964). The use of special matrix operators in statistical calculus. *ETS Res. Bul. 64-51*. Princeton: Educational Testing Service.

Blalock, H.M. (1968). Theory building and causal inference. Chapter 5 in *Methodology in Social Research*. Blalock, H.M. and Blalock, A. (eds.) New York: McGraw Hill.

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

Browne, M.W. (1974). Generalized least squares estimators in the analysis of covariance structures. *So. Afr. Statist. J. 8*, 1-24.

Christ, C.F. (1966). *Econometric Models and Methods*. New York: Wiley & Sons.

Cox, D.R. and Lauh, E. (1967). A note on the graphical analysis of multidimensional contingency tables. *Technometrics 9*, 481-488.

Cox, D.R. and Hinkley, D. (1974). *Theoretical Statistics*. London: Chapman and Hall.

Cox, D.R. (1975). Personal communication.

Cramèr, H. (1946). *Mathematical Methods of Statistics*. Princeton Univ. Press.

Dempster, A.P. (1969). *Elements of Continuous Multivariate Analysis.* Reading, Mass.: Addison-Wesley.

Dempster, A.P. (1971). An overview of multivariate data analysis. *J. Mult. Anal.* *1*, 316-346.

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

Devlin, S.J., Gnanadesikan, R. and Kettenring, J.R. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika* *62*.

Elston, R.C. (1975). On the correlation between correlations. *Biometrika* *62*, 133-140.

Fisher, R.A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* *10*, 507-521. Reprinted (1971) in *Collected Papers of R.A. Fisher.* Bennett, J.H. (ed.) Univ. of Adelaide.

Fisher, R.A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron* *1*, 3-32.

Fisher, R.A. (1924). The distribution of the partial correlation coefficient. *Metron* *3*, 329-332.

Fisher, R.A. (1924a). The influence of rainfall on the yield of wheat at Rothamsted. *Phil. Trans. B* *213*, 89-142.

Fowlkes, E.B., *et al.* (1975). Unpublished Bell Laboratories technical report.

Guttman, L. (1955). A generalized simplex for factor analysis. *Psychometrika* *20*, 173-191.

Hills, M. (1969). On looking at large correlation matrices. *Biometrika* *56*, 249-253.

Hotelling, H. (1953). New light on the correlation coefficient and its transforms. *J. Roy. Statist. Soc. B* *15*, 193-223.

Johnson, S.C. (1967). Hierarchical clustering schemes. *Psychometrika* *32*, 241-254.

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

Lancaster, P. (1969). *Theory of Matrices* New York: Academic Press.

Li, C.C. (1956). The concept of path coefficients and its impact on population genetics. *Biometrics 12*, 190-210.

Mallows, C.L. (1969). Techniques for non-standard order statistics. *ISI Proc. 37*, 164-166.

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

Mukherjee, B.N. (1970). Likelihood ratio tests of statistical hypotheses associated with patterned covariance matrices in psychology. *Br. J. Math. Soc. Psychol. 23*, 89-120. Nagao, H. (1973). On some test criteria for covariance matrix. *Ann. Statist. 1*, 700-709.

Olkin, I. and Press, S.J. (1969). Testing and estimation for a circular stationary model. *Ann. Math. Statist. 40*, 1358-1373.

Rao, C.R. (1972). Estimating variance and covariance components in linear models. *J. Amer. Statist. Assoc. 67*, 112-115.

Silvey, S.D. (1970). *Statistical Inference.* Harmondsworth: Penguin.

Siotani, M. (1968). Some methods for asymptotic distributions in the multivariate analysis. Univ. of N.C. Tech. Rep.

Tukey, J.W. (1954). Causation, regression and path analysis. In *Statistics and Mathematics in Biology.* Kempthorne, O., *et al.* (eds.) Iowa State College Press.

Tukey, J.W. and Wilk, M.B. (1966). Data analysis and statistics: an expository overview. *AFIPS Proc. 29*, 695-705.

Tukey, J.W. (1973). Personal communication.

Turner, M.E. and Stevens, C.D. (1959). The regression analysis of causal paths. *Biometrics 15*, 236-258.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc. 54*, 426-482.

Warner, J.L. (1969). Hierarchical clustering schemes. Unpublished Bell Laboratories technical report.

Wright, S. (1918). On the nature of size factors. *Genetics 3*, 367-374.

Wright, S. (1934). The method of path coefficients. *Ann. Math. Statist.* *5*, 161-215.

Wright, S. (1960). The treatment of reciprocal interaction, with or without lag, in path analysis. *Biometrics* *16*, 423-445.