This book is in the
ADDISON-WESLEY SERIES
IN BEHAVIORAL SCIENCES: QUANTITATIVE METHODS

Consulting Editor

FREDERICK MOSTELLER

Elements of CONTINUOUS MULTIVARIATE ANALYSIS

A. P. DEMPSTER
Department of Statistics, Harvard University

ADDISON-WESLEY PUBLISHING COMPANY
Reading, Massachusetts · Menlo Park, California · London · Don Mills, Ontario

CHAPTER 6

DUAL SPACES

6.1 BASIC DEFINITIONS AND THEORY

Given any vector space $\mathscr{E}$, it is possible to define as follows a new vector space $\mathscr{F}$ whose elements consist of the set of all linear functionals over $\mathscr{E}$. A *linear functional* over $\mathscr{E}$ is a real-valued function $v(V)$, defined for all V in $\mathscr{E}$, which satisfies the requirement that

$$v(\alpha_1 V_1 + \alpha_2 V_2) = \alpha_1 v(V_1) + \alpha_2 v(V_2) \tag{6.1.1}$$

for all vectors V_1 and V_2 in $\mathscr{E}$ and all real numbers α_1 and α_2. The addition of two linear functionals to give a linear functional and the multiplication of a linear functional by a real number to give a linear functional are both defined in the obvious manner, i.e., the relation $v_3 = \beta_1 v_1 + \beta_2 v_2$ for linear functionals v_1, v_2 and v_3 and real numbers β_1 and β_2 means that

$$v_3(V) = \beta_1 v_1(V) + \beta_2 v_2(V) \tag{6.1.2}$$

for all V in $\mathscr{E}$. It is easily checked that the set of linear functionals forms a vector space under the operations defined by (6.1.2), and this defines the vector space $\mathscr{F}$, as promised above.

The mathematical interest in considering $\mathscr{F}$ is that $\mathscr{E}$ and $\mathscr{F}$ have the mathematical properties of what will be called a pair of dual vector spaces. An abstract definition of this concept will be given shortly. The statistical interest in considering $\mathscr{F}$ is that if $\mathscr{E}$ is taken to be the variable-space of Example 2.1.1, then $\mathscr{F}$ is essentially the individual-space of Example 2.1.2. Note that any observation vector $[x_1, x_2, \ldots, x_p]$ on a basic set of variables $[V_1, V_2, \ldots, V_p]'$ defines an observation $\boldsymbol{\alpha}\mathbf{x}'$ on any variable $\boldsymbol{\alpha}\mathbf{V}$ in $\mathscr{E}$, and that the observation $\boldsymbol{\alpha}\mathbf{x}'$ may be regarded as the value of a functional v, where

$$v(\boldsymbol{\alpha}\mathbf{V}) = \boldsymbol{\alpha}\mathbf{x}'. \tag{6.1.3}$$

It may easily be checked that the functional v defined by (6.1.3) obeys (6.1.1), and so is a linear functional. Moreover, the vector operations for observation vectors defined in Example 2.1.2 and the vector operations for functionals defined in (6.1.2) determine the same operations on the functionals defined by (6.1.3), so that the concepts of individual-space and the space of linear functionals over variable-space are abstractly identical.

Returning to duality in general, any two vector spaces $\mathscr{E}$ and $\mathscr{F}$ will be called *dual* with respect to each other if they are related by a proper bilinear product function. A *bilinear product function* is a real-valued function $\{V, v\}$ defined for each V in $\mathscr{E}$ and v in $\mathscr{F}$, which satisfies the pair of linearity requirements that

$$\{\alpha_1 V_1 + \alpha_2 V_2, v\} = \alpha_1\{V_1, v\} + \alpha_2\{V_2, v\} \tag{6.1.4}$$

and

$$\{V, \beta_1 v_1 + \beta_2 v_2\} = \beta_1\{V, v_1\} + \beta_2\{V, v_2\}, \tag{6.1.5}$$

where α_1, α_2, β_1, and β_2 denote any real numbers, where V_1, V_2, and V denote any vectors in $\mathscr{E}$, and where v_1, v_2, and v denote any vectors in $\mathscr{F}$. From (6.1.4) and (6.1.5), it is clear that for each fixed v the relation

$$f_v(V) = \{V, v\} \tag{6.1.6}$$

defines a linear functional f_v over $\mathscr{E}$, and that for each fixed V the relation

$$F_V(v) = \{V, v\} \tag{6.1.7}$$

defines a linear functional F_V over $\mathscr{F}$. Furthermore, the mapping

$$v \rightarrow f_v \tag{6.1.8}$$

from the space $\mathscr{F}$ to the space of linear functionals over $\mathscr{E}$ is a linear transformation. Similarly the mapping

$$V \rightarrow F_V \tag{6.1.9}$$

from the space $\mathscr{E}$ to the space of linear functionals over $\mathscr{F}$ is a linear transformation. If the mappings (6.1.8) and (6.1.9) are both one-one in the sense that they define an isomorphism between one vector space and the space of linear functionals over another vector space, then the bilinear product $\{V, v\}$ is here called *proper*, and, in this case, the spaces $\mathscr{E}$ and $\mathscr{F}$ are dual spaces.

The definition of dual spaces treats $\mathscr{E}$ and $\mathscr{F}$ symmetrically, so that $\mathscr{F}$ dual to $\mathscr{E}$ implies $\mathscr{E}$ dual to $\mathscr{F}$. Unfortunately, however, it is not immediately clear that individual-space $\mathscr{F}$ and variable-space $\mathscr{E}$ are dual according to the definition. Specifically, it is clear that (6.1.8) is one-one but not that (6.1.9) is one-one. In other words, it needs to be shown that each linear functional over $\mathscr{F}$ corresponds to exactly one variable in $\mathscr{E}$ whose values are those given by the linear functional over $\mathscr{F}$. The missing link is provided by:

Lemma 6.1. *Suppose that two vector spaces $\mathscr{E}$ and $\mathscr{F}$ are related by a bilinear product $\{V, v\}$ such that one of the mappings* (6.1.8) *and* (6.1.9) *is one-one. Then the other mapping is also one-one, and $\mathscr{E}$ and $\mathscr{F}$ are dual spaces.*

It need only be proved that if (6.1.8) is a one-one mapping then (6.1.9) is also, for the symmetry of the hypotheses will then imply that the converse is also true. Assuming that (6.1.8) is one-one, it follows that $\mathscr{F}$ is abstractly identical to the space of linear functionals over $\mathscr{E}$, so it may be assumed that $\mathscr{F}$ is this space of linear functionals and that (6.1.6) holds. It is required to show that only one V corresponds to a given F_V, or that the relation $\{V_1, v\} = \{V_2, v\}$ for all v implies that $V_1 = V_2$, or, finally, that the relation $v(V_1) = v(V_2)$ for all linear functionals v implies that $V_1 = V_2$. But this last version is obvious, for, given $V_1 \neq V_2$, one can easily construct a linear functional v such that $v(V_1) \neq v(V_2)$, and so the lemma is proved by contradiction.

It is now clear that the space $\mathscr{F}$ of linear functionals over $\mathscr{E}$ should be regarded as a dual space for $\mathscr{E}$ according to the bilinear product

$$\{V, v\} = v(V) \tag{6.1.10}$$

for V in $\mathscr{E}$ and v in $\mathscr{F}$. The relations (6.1.1) and (6.1.2) imply (6.1.4) and (6.1.5), so that $\{V, v\}$ in (6.1.10) defines a bilinear product. Moreover, from (6.1.6) and (6.1.10) the functionals v and f_v are identical, so that (6.1.8) is trivially a one-one linear transformation, as required.

According to the given definition of dual spaces, it is possible for a given vector space $\mathscr{E}$ to be dual to a number of different vector spaces, or even dual to a given vector space $\mathscr{F}$ in a number of different ways corresponding to different bilinear product functions. Still, it makes sense to speak of *the* dual space $\mathscr{F}$ of a given vector space $\mathscr{E}$. The reason for this terminology is essentially given by the one-one relation (6.1.8) which asserts that all dual spaces are isomorphic to the dual space of linear functionals according to an isomorphism which preserves the values of the bilinear product function.

Virtually every concept and entity concerning a vector space $\mathscr{E}$ has a corresponding and generally different concept and entity in the dual space $\mathscr{F}$. Consequently, any statement concerning a vector space $\mathscr{E}$ may be rewritten as an equivalent but apparently different statement concerning the dual space $\mathscr{F}$. This translation is a useful device because certain statements may seem more familiar and therefore more comprehensible in terms of one space than in terms of the other. Some of these corresponding dual entities will be derived in the remainder of this section, beginning with dual bases and dual subspaces.

Suppose that $\mathbf{V} = [V_1, V_2, \ldots, V_p]'$ is any basis of a vector space $\mathscr{E}$ whose dual space is $\mathscr{F}$. Then a basis $\mathbf{v} = [v_1, v_2, \ldots, v_p]'$ of $\mathscr{F}$ dual to the basis $\mathbf{V}$ of $\mathscr{E}$ may be defined by the relations

$$\begin{aligned} \{V_i, v_j\} &= 1 \quad \text{if} \quad i = j \\ &= 0 \quad \text{if} \quad i \neq j, \end{aligned} \tag{6.1.11}$$

for i and $j = 1, 2, \ldots, p$. Note that the set of relations (6.1.11), as i ranges while j is fixed, determines the values of the functional v_j in $\mathscr{F}$ for the basis elements $V_1, V_2, \ldots, V_p$ in $\mathscr{E}$. Consequently, $v_j(V)$ is determined for any V in $\mathscr{E}$. The reader may easily check that the $v_1, v_2, \ldots, v_p$ defined in this way are linearly independent and that any v in $\mathscr{F}$ may be expressed as a linear combination of $v_1, v_2, \ldots, v_p$. In other words, $\mathbf{v}$ is in fact a basis of $\mathscr{F}$ and the concept of dual basis is well-defined. It follows incidentally that $\mathscr{E}$ and its dual space $\mathscr{F}$ have the same dimension, a fact already illustrated in the examples of variable-space and its dual individual-space.

Suppose that $\mathscr{U}$ is an r-dimensional subspace of a p-dimensional vector space $\mathscr{E}$ whose dual space is $\mathscr{F}$. A $(p - r)$-dimensional subspace $\mathscr{U}_d$ of $\mathscr{F}$ dual to $\mathscr{U}$ in $\mathscr{E}$ may be defined by the condition that v is in $\mathscr{U}_d$ if

$$\{V, v\} = 0 \qquad \text{for all } V \text{ in } \mathscr{U}. \tag{6.1.12}$$

Relation (6.1.5) is enough to ensure that $\mathscr{U}_d$ is a subspace of $\mathscr{F}$. To show that $\mathscr{U}_d$ has dimension $p - r$, suppose that $\mathbf{V} = [V_1, V_2, \ldots, V_p]'$ is a basis of $\mathscr{E}$ such that $V_1, V_2, \ldots, V_r$ span $\mathscr{U}$. It follows from (6.1.11) and (6.1.12) that the elements $v_{r+1}, v_{r+2}, \ldots, v_p$ of the dual basis $\mathbf{v} = [v_1, v_2, \ldots, v_p]'$ of $\mathscr{F}$ lie in $\mathscr{U}_d$. Moreover, for any $v = x_1v_1 + x_2v_2 + \cdots + x_rv_r$ with some $x_i \neq 0$, there exists a V in $\mathscr{U}$, for example V_i, such that $\{V, v\} \neq 0$. Consequently $\mathscr{U}_d$ is the subspace spanned by $v_{r+1}, v_{r+2}, \ldots, v_p$ and has dimension $p - r$.

A limiting case of the duality between $\mathscr{U}$ and $\mathscr{U}_d$ occurs when $\mathscr{U}$ is taken to be the subspace consisting only of $\emptyset$ in $\mathscr{E}$; then $\mathscr{U}_d$ is the subspace consisting of the whole of $\mathscr{F}$. Another important property of the duality is that, if $\mathscr{U}_d$ and $\mathscr{V}_d$ in $\mathscr{F}$ are the duals of $\mathscr{U}$ and $\mathscr{V}$ in $\mathscr{E}$, then $\mathscr{U}_d \cap \mathscr{V}_d$ in $\mathscr{F}$ is the dual of $\mathscr{U} \oplus \mathscr{V}$ in $\mathscr{E}$. Of course, if $\mathscr{U}_d$ is the dual of $\mathscr{U}$, then $\mathscr{U}$ is the dual of $\mathscr{U}_d$.

Suppose that A is a narrow sense linear transformation from $\mathscr{E}$ to $\mathscr{E}^*$ whose dual spaces are $\mathscr{F}$ and $\mathscr{F}^*$, corresponding to bilinear products $\{V, v\}$ and $\{V^*, v^*\}^*$, respectively. Then there is a unique narrow sense linear transformation A_d from $\mathscr{F}^*$ to $\mathscr{F}$ which satisfies

$$\{V, v\} = \{V^*, v^*\}^* \tag{6.1.13}$$

for all V in $\mathscr{E}$ and v^* in $\mathscr{F}^*$, where $V^* = \mathsf{A}V$ and $v = \mathsf{A}_dv^*$. The proof of this assertion requires only simple checking: (a) that $\{\mathsf{A}V, v^*\}^*$ defines a linear functional over $\mathscr{E}$ for each v^*, i.e., defines a member v of $\mathscr{F}$ for each v^*, and (b) that the mapping $v^* \to v = \mathsf{A}_dv^*$ so defined is a linear transformation. The linear transformation A_d will be called the *dual linear transformation of* A. Clearly, if A_d is the dual of A, then A is the dual of A_d. If $\mathscr{U}$ is the subspace in $\mathscr{E}$ which maps under A into the origin in $\mathscr{E}^*$ while $\mathscr{V}^*$ is the range space in $\mathscr{E}^*$ of the transformation A, then $\mathscr{V}^*_d$ is the subspace of $\mathscr{F}^*$ which maps under A_d into the origin in $\mathscr{F}$ and $\mathscr{U}_d$ is the range space in $\mathscr{F}$ of the transformation A_d.

Given an inner product function (V, U) defined for each pair V, U in $\mathscr{E}$, the next task is to define the dual inner product $(v, u)_d$ for each pair v, u in the dual space $\mathscr{F}$ of $\mathscr{E}$. For any v in $\mathscr{F}$, there exists a hyperplane of dimension $p - 1$ in $\mathscr{E}$ such that $\{V, v\} = 1$ for all V in the hyperplane, and V^* may be defined to be the orthogonal projection of $\emptyset$ into this hyperplane. Similarly, for u in $\mathscr{F}$ a corresponding U^* in $\mathscr{E}$ may be defined. With this structure in hand the dual inner product is defined by

$$(v, u)_d = \frac{(V^*, U^*)}{(V^*, V^*)(U^*, U^*)}\,. \tag{6.1.14}$$

To check that (6.1.14) defines an inner product, and incidentally to provide a simple alternative definition, consider any orthonormal basis $\mathbf{V}$ of $\mathscr{E}$ and the dual basis $\mathbf{v}$ of $\mathscr{F}$. Representing $v = \mathbf{xv}$, the hyperplane in $\mathscr{E}$ whose points V satisfy $\{V, v\} = 1$ may be represented analytically as the set of points $V = \boldsymbol{\alpha}\mathbf{V}$ such that $\boldsymbol{\alpha}\mathbf{x}' = 1$; the point V^* on this hyperplane which is closest to the origin is given by

$$V^* = [(\mathbf{xx}')^{-1}\mathbf{x}]\mathbf{V}. \tag{6.1.15}$$

Similarly, representing $u = \mathbf{yv}$, the corresponding

$$U^* = [(\mathbf{yy}')^{-1}\mathbf{y}]\mathbf{V}. \tag{6.1.16}$$

From (6.1.14), (6.1.15), and (6.1.16), together with the orthonormality of $\mathbf{V}$, one finds

$$\begin{aligned}(v, u)_d &= \frac{[(\mathbf{xx}')^{-1}\mathbf{x}][(\mathbf{yy}')^{-1}\mathbf{y}]'}{[(\mathbf{xx}')^{-1}\mathbf{x}][(\mathbf{xx}')^{-1}\mathbf{x}]' \cdot [(\mathbf{yy}')^{-1}\mathbf{y}][(\mathbf{yy}')^{-1}\mathbf{y}]'}\\ &= \mathbf{xy}'.\end{aligned} \tag{6.1.17}$$

It follows that $(v, u)_d$ as originally defined in (6.1.14) is identical to the inner product defined by asserting that $\mathbf{v}$ is orthonormal. Having thus shown that the coordinate-free definition (6.1.14) is legitimate, it follows that the dual basis $\mathbf{v}$ of *any* orthonormal basis $\mathbf{V}$ is orthonormal according to the dual inner product.

A Euclidean vector space $\mathscr{E}$ with a proper inner product is self-dual where the bilinear product relating $\mathscr{E}$ with itself is simply the given inner product over $\mathscr{E}$ which makes $\mathscr{E}$ Euclidean. The inner product is clearly a bilinear product and a proper one because $(V, W_1) = (V, W_2)$ for all V implies that $W_1 = W_2$. The one-one correspondence between $\mathscr{E}$ and the space of linear functionals over $\mathscr{E}$ implied by the representation of the dual space as $\mathscr{E}$ is

$$W \leftrightarrow v_W, \tag{6.1.18}$$

where

$$v_W(V) = (W, V) \tag{6.1.19}$$

for V and W in $\mathscr{E}$. If $\mathbf{W}$ is an orthonormal basis of $\mathscr{E}$, the isomorphism (6.1.18) carries

$$\mathbf{W} \leftrightarrow \mathbf{w}, \tag{6.1.20}$$

where $\mathbf{w}$ is the dual orthonormal basis of the space of linear functionals over $\mathscr{E}$. In this sense there is a natural isomorphism between any Euclidean vector space and its dual Euclidean vector space which carries every orthonormal basis into its dual orthonormal basis.

The foregoing theory provides an inner product defined on individual-space dual to a given covariance inner product on variable-space. Such a dual inner product will be called a *concentration inner product* in Chapter 7 and is an important concept in multivariate statistical theory.

6.2 DUAL GEOMETRIC SPACES

The idea of duality is often introduced in simple discussions of geometry. For example, in plane geometry the concept of a line is thought of as the dual of the concept of a point, the set of points on a line is the dual of a pencil of lines through a point, and the proposition that every two points define a line joining them is the dual of the proposition that every two lines intersect in a point. When this kind of duality is applied to p-dimensional space it asserts that points are dual to $(p - 1)$-dimensional hyperplanes, or, more generally, that r-dimensional hyperplanes are dual to $(p - r - 1)$-dimensional hyperplanes for $r = 0, 1, 2, \ldots, p - 1$.

To arrive at such duality concepts in vector space terms, the dual of a point V in a vector space $\mathscr{E}$ may be defined to be the $(p - 1)$-dimensional hyperplane consisting of points v in the dual space $\mathscr{F}$ of $\mathscr{E}$ which satisfy $\{V, v\} = 1$. The dual of a hyperplane in $\mathscr{E}$ may then be defined as the hyperplane of intersection of the family of $(p - 1)$-dimensional hyperplanes in $\mathscr{F}$ dual to the points of its original hyperplane in $\mathscr{E}$. The reader may wish to show that in this way a one-one correspondence is defined between r-dimensional hyperplanes in $\mathscr{E}$ and $(p - r - 1)$-dimensional hyperplanes in $\mathscr{F}$, and that, if two hyperplanes intersect in a third, then the duals of these two hyperplanes determine the dual of the third as the smallest hyperplane containing them both.

Note that the above concept of dual pairs of hyperplanes does not apply to hyperplanes through the origin, and is different from the concept of dual pairs of subspaces (i.e., hyperplanes through the origin) defined in Section 6.1. To see the relationship between these two types of duals, suppose that $u + \mathscr{V}$ in $\mathscr{F}$ denotes the dual hyperplane for the point V in $\mathscr{E}$. It is easily checked that, if $\mathscr{U}$ in $\mathscr{E}$ denotes the one-dimensional subspace consisting of the points αV for $-\infty < \alpha < \infty$, then $\mathscr{U}$ and $\mathscr{V}$ are a pair of dual subspaces in the sense of Section 6.1. Moreover, the dual hyperplane of αV is $\alpha^{-1}u + \mathscr{V}$. In other words, as V moves along a line towards the origin, the dual hyperplanes are a progression of parallel hyperplanes moving away from the origin, and vice versa. If V were to reach the origin then $u + \mathscr{V}$ would need to be a hyperplane at infinity. The concept of hyperplane at infinity is rigorously introduced in affine geometry, but need not be pursued here.

In this way the duality concept shows how the set of $(p - 1)$-dimensional hyperplanes of a given vector space $\mathscr{E}$ may themselves be regarded as forming a vector space, i.e., one need only consider the corresponding operations on the points of the dual space $\mathscr{F}$. Multiplying a hyperplane by α results in a hyperplane $1/\alpha$ times as distant from the origin, and the zero hyperplane is the hyperplane at infinity.

The dual of an origin-centered ellipsoid π in $\mathscr{E}$ is an origin-centered ellipsoid π_d in $\mathscr{F}$, where π and π_d are symbols for a pair of dual inner products or for their corresponding ellipsoids. However, the dual of a point on the surface of an ellipsoid π is a $(p - 1)$-dimensional hyperplane tangent to the dual ellipsoid π_d. To see this, consider any point V on the surface of π, i.e., a point such that $(V, V) = 1$ according to π. Then V corresponds to a hyperplane $v_V + \mathscr{F}_V$ in $\mathscr{F}$, where v_V may be chosen orthogonal to $\mathscr{F}_V$ according to π_d. Applying the dual form of (6.1.14) yields

$$1 = (V, V) = \frac{(v_V, v_V)_d}{(v_V, v_V)_d(v_V, v_V)_d} = (v_V, v_V)_d^{-1} = (v_V, v_V)_d, \tag{6.2.1}$$

so that v_V lies on the surface of π_d. Since $\mathscr{F}_V$ is orthogonal to v_V according to π_d, $v_V + \mathscr{F}_V$ is tangent to π_d at v_V, as required. It follows that the dual ellipsoid π_d may be regarded as the envelope of the family of $(p - 1)$-dimensional hyperplanes which are the duals of the points on π.

It is illuminating to describe how to determine the length of line segment $\emptyset W$ in $\mathscr{E}$ from the dual ellipsoid π_d. Recall that the length of $\emptyset W$ may be regarded as the affine ratio of the length of $\emptyset W$ to the length of the semi-axis of π in the same direction, i.e., as the ratio of the length of $\emptyset W$ to the length of $\emptyset V$ where $V = \alpha W$ and $\alpha = (W, W)^{-1/2}$. To dualize this characterization, consider the dual $v_W + \mathscr{F}_W$ of W where v_W is orthogonal to $\mathscr{F}_W$ according to π_d. The corresponding dual of $V = \alpha W$ is $\alpha^{-1}v_W + \mathscr{F}_W$. Thus the length of W is the ratio of $v_V = \alpha^{-1}v_W$ on π_d to the length of v_W. This is pictured in Fig. 6.2.1 in two dimensions. Note that, as pictured in Fig. 6.2.1, $v_W + \mathscr{F}_W$ lies outside π_d, which implies that W has length *less* than unity. In general, the farther from the origin that $v_W + \mathscr{F}_W$ moves, the smaller the length of W becomes.

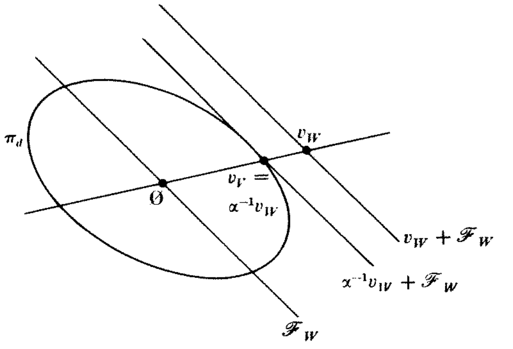

Fig. 6.2.1. The dual hyperplanes $v_W + \mathscr{F}_W$ and $\alpha^{-1}v_W + \mathscr{F}_W$ of the points W and V. The length of W is the ratio of the length of $\emptyset v_W$ to the length of $\emptyset v_V$.

From (6.1.14) it is clear that if the inner product π is scaled, i.e., multiplied by a factor λ, then the dual inner product is scaled by the factor λ^{-1}. At the same time, the ellipsoid π in $\mathscr{E}$ is scaled by the factor $\lambda^{-1/2}$, i.e., the length of each axis is multiplied by $\lambda^{-1/2}$, but the ellipsoid π_d in $\mathscr{F}$ is scaled by the direct factor $\lambda^{1/2}$.

6.3 SOME RELATED MATRIX THEORY

Suppose that $\mathbf{V} = [V_1, V_2, \ldots, V_p]'$ and $\mathbf{W} = [W_1, W_2, \ldots, W_p]'$ are two bases of $\mathscr{E}$ related by

$$\mathbf{V} = \mathbf{AW}, \qquad \boldsymbol{\beta} = \boldsymbol{\alpha}\mathbf{A}, \tag{6.3.1}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ denote coordinates of points of $\mathscr{E}$ relative to $\mathbf{V}$ and $\mathbf{W}$ respectively. Then the corresponding dual bases $\mathbf{v} = [v_1, v_2, \ldots, v_p]'$ and $\mathbf{w} = [w_1, w_2, \ldots, w_p]'$ of $\mathscr{F}$ are related by

$$\mathbf{w} = \mathbf{A}'\mathbf{v}, \qquad \mathbf{x} = \mathbf{y}\mathbf{A}', \tag{6.3.2}$$

where $\mathbf{x}$ and $\mathbf{y}$ are the coordinates of points of $\mathscr{F}$ relative to $\mathbf{v}$ and $\mathbf{w}$, respectively. To see this, note that either of the equations of (6.3.2) implies the other, and that the second equation of (6.3.2) is an immediate consequence of the first equation of (6.3.1); for, in the language of variable-space and individual-space, if variables obey a certain linear relation $\mathbf{V} = \mathbf{AW}$; then values of those variables obey the same linear relation $\mathbf{x}' = \mathbf{Ay}'$.

In matrix terms the concept of a dual linear transformation is very simple. Suppose that A is a linear transformation from a p-dimensional space $\mathscr{E}$ to a q-dimensional space $\mathscr{E}^*$, and that A is represented by

$$\mathbf{V} \rightarrow \mathbf{AW}, \tag{6.3.3}$$

where $\mathbf{V}$ is a basis of $\mathscr{E}$, $\mathbf{W}$ is a basis of $\mathscr{E}^*$, and $\mathbf{A}$ is the $p \times q$ matrix which determines A relative to these bases. It is easily checked from (6.1.13) that the dual linear transformation A_d may be represented by

$$\mathbf{w} \rightarrow \mathbf{A}'\mathbf{v}, \tag{6.3.4}$$

where $\mathbf{v}$ and $\mathbf{w}$ are the dual bases of $\mathbf{V}$ and $\mathbf{W}$ in the dual spaces $\mathscr{F}$ and $\mathscr{F}^*$ of $\mathscr{E}$ and $\mathscr{E}^*$, respectively.

Suppose that an inner product π on $\mathscr{E}$ has the inner product matrix $\mathbf{Q}$ relative to a basis $\mathbf{V}$. Then an important matrix result is that *the dual inner* *product π_d in $\mathscr{F}$ has the inner product matrix $\mathbf{Q}^{-1}$ relative to the dual basis* $\mathbf{v}$ *corresponding to* $\mathbf{V}$. To see this, suppose that $\mathbf{U} = \mathbf{CV}$ is orthonormal relative to π, so that

$$\mathbf{CQC}' = \mathbf{I} \qquad \text{or} \qquad \mathbf{Q} = [\mathbf{C}'\mathbf{C}]^{-1}. \tag{6.3.5}$$

Then the corresponding $\mathbf{u}$, where $\mathbf{v} = \mathbf{C}'\mathbf{u}$, is known to be orthonormal relative to π_d. Thus, if π_d has inner product matrix $\mathbf{P}$ relative to $\mathbf{v}$, then

$$\mathbf{P} = \mathbf{C}'\mathbf{IC} = \mathbf{C}'\mathbf{C}. \tag{6.3.6}$$

From (6.3.5) and (6.3.6),

$$\mathbf{P} = \mathbf{Q}^{-1}, \tag{6.3.7}$$

as required.

6.4 DUALITY ASPECTS OF THE PROCESS OF SUCCESSIVE ORTHOGONALIZATION

Another example of the inversion phenomena experienced in passing from one space to its dual is given by:

> ***Theorem 6.4.*** *Suppose that the basis* $\mathbf{U}$ *of a Euclidean space* $\mathscr{E}$ *is orthogonalized as in Section* 4.1 *to produce an orthogonal basis* $\mathbf{U}^*$ *and then an orthonormal basis* $\mathbf{U}^{**}$. *Suppose that* $\mathbf{U}$, $\mathbf{U}^*$, *and* $\mathbf{U}^{**}$ *have dual bases* $\mathbf{u}$, $\mathbf{u}^*$, *and* $\mathbf{u}^{**}$, *respectively, in the dual Euclidean space* $\mathscr{F}$ *of* $\mathscr{E}$. *Then, if the basis* $\mathbf{u}$ *is orthogonalized in the reverse of the given order, the resulting orthogonal basis is* $\mathbf{u}^*$ *and the corresponding orthonormal basis is* $\mathbf{u}^{**}$.

Since $\mathbf{U}^*$ is orthogonal in $\mathscr{E}$, its dual $\mathbf{u}^*$ is known to be orthogonal in $\mathscr{F}$, and similarly, $\mathbf{u}^{**}$ is known to be orthonormal in $\mathscr{F}$. Thus it remains only to show that $\mathbf{u}^*$ and $\mathbf{u}^{**}$ are the specific orthogonal bases indicated by the theorem. Using the notation of Section 4.3, it is known that

$$\mathbf{U}^* = \mathbf{AU} \tag{6.4.1}$$

may be characterized as that orthogonal basis of $\mathscr{E}$ such that $\mathbf{A}$ is a triangular matrix with unity along the main diagonal and zero above the main diagonal. Similarly

$$\mathbf{U}^{**} = \mathbf{CU} \tag{6.4.2}$$

may be characterized as that orthonormal basis of $\mathscr{E}$ such that $\mathbf{C}$ is triangular with zero elements above the main diagonal. From (6.3.2),

$$\mathbf{u}^* = \mathbf{B}'\mathbf{u} \qquad \text{and} \qquad \mathbf{u}^{**} = \mathbf{D}'\mathbf{u}, \tag{6.4.3}$$

where

$$\mathbf{B} = \mathbf{A}^{-1} \qquad \text{and} \qquad \mathbf{D} = \mathbf{C}^{-1}. \tag{6.4.4}$$

Thus $\mathbf{B}'$ is triangular with unity along the main diagonal and zero *below* the main diagonal and, similarly, $\mathbf{D}'$ is triangular with zero below the main diagonal.

These properties of $\mathbf{B}'$ and $\mathbf{D}'$ are sufficient to imply, respectively, that $\mathbf{u}^*$ and $\mathbf{u}^{**}$ are the particular reverse order bases specified in the theorem.

Theorem 6.4 supplies a concise explanation of the dual formulas which were profusely displayed in Section 4.2. Any formula involving $\mathbf{Q}$, $\mathbf{P}$, $\mathbf{T}$, $\mathbf{T}^{-1}$, $\mathbf{B}$, $\mathbf{A}$, $\mathbf{D}$, and $\mathbf{C}$ may be regarded as describing aspects of a basis $\mathbf{U}$ of a Euclidean space $\mathscr{E}$, where $\mathbf{U}$ has the inner product matrix $\mathbf{Q}$. The same formulas may equally well be applied to yield similar descriptions of the basis $\mathbf{u}$ of $\mathscr{F}$, treating the elements of $\mathbf{u}$ in the opposite order from the elements of $\mathbf{U}$. Thus, the dual of any formula may be immediately written down simply by replacing $\mathbf{Q}$, $\mathbf{P}$, $\mathbf{T}$, $\mathbf{T}^{-1}$, $\mathbf{B}$, $\mathbf{A}$, $\mathbf{D}$, and $\mathbf{C}$ by $\mathbf{P}^*$, $\mathbf{Q}^*$, $\mathbf{T}^{*-1}$, $\mathbf{T}^*$, $\mathbf{A}^*$, $\mathbf{B}^*$, $\mathbf{C}^*$, and $\mathbf{D}^*$, respectively, where the star notation means that the order of the rows and columns has been reversed, i.e., $\mathbf{P}^*$ is the same as $\mathbf{P}$ except for the reversing of the order of the rows and columns of $\mathbf{P}$, and so on.

Theorem 6.4 also suggests that the computational routines applicable to $\mathbf{Q}$, as described in Section 4.3, should also be interesting when applied to $\mathbf{P} = \mathbf{Q}^{-1}$. In particular, the dual of successively applying the sweep operator of (4.3.13) to the rows and columns of $\mathbf{Q}$ in order is the operation of successively applying the sweep operator to the rows and columns of $\mathbf{P}$ in reverse order. After s stages of the former, one has (4.3.14), which may be written

$$\begin{bmatrix} -\mathbf{Q}_{11}^{-1} & \mathbf{Q}_{11}^{-1}\mathbf{Q}_{12} \\ \mathbf{Q}_{21}\mathbf{Q}_{11}^{-1} & \mathbf{Q}_{22.1} \end{bmatrix}. \tag{6.4.5}$$

It follows dually that after $p - s$ stages of the latter, one must have

$$\begin{bmatrix} \mathbf{P}_{11.2} & \mathbf{P}_{12}\mathbf{P}_{22}^{-1} \\ \mathbf{P}_{22}^{-1}\mathbf{P}_{21} & -\mathbf{P}_{22}^{-1} \end{bmatrix}. \tag{6.4.6}$$

From the relations in Section 4.2, it follows that (6.4.5) is simply the negative of (6.4.6). Thus, the dual of the computations of the elimination method yields the same quantities as the original except for the changed signs and reversed order.

6.5 DUALITY RELATIONS CONCERNING A PAIR OF INNER PRODUCTS

Suppose that π_1 and π_2 are two inner products defined over $\mathscr{E}$ with dual inner products π_{1d} and π_{2d} defined over the dual space $\mathscr{F}$ of $\mathscr{E}$. If $\mathbf{W}$ is an orthogonal basis according to both π_1 and π_2, then the dual basis $\mathbf{W}$ of $\mathscr{F}$ is orthogonal according to both π_{1d} and π_{2d}, i.e., the dual of a basis of eigenvectors is also a basis of eigenvectors for the dual space. The eigenvalues of π_{1d} relative to π_{2d} are, however, the inverses of the eigenvalues of π_1 relative to π_2. To see this, suppose that $\mathbf{W}$ is scaled to be orthonormal relative to π_2. Then, relative to π_1,

$$\begin{aligned} (W_i, W_i)_1 &= \lambda_i, \qquad \text{for} \quad i = 1, 2, \ldots, p, \\ (W_i, W_j)_1 &= 0, \qquad \text{for} \quad i \neq j, \end{aligned} \tag{6.5.1}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the eigenvalues of π_1 relative to π_2. From (6.1.11), $\mathbf{w}$ is orthonormal relative to π_{2d}, but

$$(w_i, w_i)_{1d} = \frac{1}{\lambda_i} \qquad \text{for} \qquad i = 1, 2, \ldots, p, \tag{6.5.2}$$
$$(w_i, w_j)_{1d} = 0 \qquad \text{for} \qquad i \neq j,$$

whence the eigenvalues of π_{1d} relative to π_{2d} are $1/\lambda_1, 1/\lambda_2, \ldots, 1/\lambda_p$.

Geometrically, the basis $\mathbf{W}$ determines sets of conjugate axes for both of the ellipsoids π_1 and π_2 in $\mathscr{E}$, while $\mathbf{w}$ does the same for the dual ellipsoids π_{1d} and π_{2d} in $\mathscr{F}$. However, the ratio of the lengths of such principal axes is inverted by the passage from one space to its dual.

The inversion of the eigenvalues is also clear analytically. For suppose that $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are the inner product matrices of π_1 and π_2 relative to a basis $\mathbf{U}$ of $\mathscr{E}$. Then the eigenvalues are roots of the equation

$$\det (\mathbf{Q}_1 - \lambda \mathbf{Q}_2) = 0. \tag{6.5.3}$$

This equation is equivalent to

$$\lambda^p \det \mathbf{Q}_1\mathbf{Q}_1^{-1}\left(\frac{1}{\lambda}\,\mathbf{Q}_1 - \mathbf{Q}_2\right)\mathbf{Q}_2^{-1}\mathbf{Q}_2 = 0, \qquad \text{or}$$

$$(-\lambda)^p \det \mathbf{Q}_1 \det \mathbf{Q}_2 \det \left(\mathbf{Q}_1^{-1} - \frac{1}{\lambda}\,\mathbf{Q}_2^{-1}\right) = 0, \qquad \text{or}$$

$$\det \left(\mathbf{Q}_1^{-1} - \frac{1}{\lambda}\,\mathbf{Q}_2^{-1}\right) = 0, \tag{6.5.4}$$

which means that the inverses $1/\lambda_i$ are the roots of the determinantal equation relative to the dual basis $\mathbf{u}$ for the eigenvalues of π_{1d} relative to π_{2d}.

To pursue the analytic approach further, suppose that $\mathbf{V}$ is orthonormal relative to π_2, and has the inner product matrix $\mathbf{Q}$ according to π_1. Then the eigenvalues are the roots of the equation

$$\det (\mathbf{Q} - \lambda \mathbf{I}) = 0, \tag{6.5.5}$$

which is the standard form for solution. If the eigenvectors $\mathbf{W}$ are also chosen to be orthonormal according to π_2, then

$$\mathbf{W} = \mathbf{E}\mathbf{V}, \tag{6.5.6}$$

where $\mathbf{E}$ is an orthogonal matrix whose rows give the coordinates of $W_1, W_2, \ldots, W_p$ relative to $\mathbf{V}$. It is interesting to note that, since from (6.3.2)

$$\mathbf{v} = \mathbf{E}'\mathbf{w} \qquad \text{or} \qquad \mathbf{w} = \mathbf{E}\mathbf{v}, \tag{6.5.7}$$

the rows of $\mathbf{E}$ are also the coordinates of the dual eigenvectors $w_1, w_2, \ldots, w_p$ relative to $\mathbf{v}$.

A quite different theory relating two sets of eigenvalues will now be presented. Suppose that π denotes an inner product over a p-dimensional space $\mathscr{E}$, and that π_d denotes the dual inner product over $\mathscr{F}$. Next, consider a separate q-dimensional Euclidean space $\mathscr{E}^*$ whose inner product is denoted by π^* and whose dual and dual inner product are denoted by $\mathscr{F}^*$ and π_d^*. Suppose that A denotes a linear transformation from $\mathscr{E}^*$ to $\mathscr{E}$. With this structure a second inner product π_{A} may be defined over $\mathscr{E}$ from

$$(U, V)_{\mathsf{A}} = (\mathsf{A}U, \mathsf{A}V)^*. \tag{6.5.8}$$

Dually, a second inner product $\pi_{\mathsf{A}_d}^*$ may be defined over $\mathscr{F}^*$ from

$$(u^*, v^*)_{\mathsf{A}_d}^* = (\mathsf{A}_d u^*, \mathsf{A}_d v^*)_d. \tag{6.5.9}$$

Finally, suppose that A has rank $r \leq \min (p, q)$. *Then there are r nonzero eigenvalues of π_{A} relative to π, and $p - r$ zero eigenvalues. These same nonzero eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_r$ together with $q - r$ zeros make up the eigenvalues of $\pi_{\mathsf{A}_d}^*$ relative to π_d^*. If W_i denotes an eigenvector in $\mathscr{E}$ associated with λ_i, for $i = 1, 2, \ldots, r$, then an eigenvector w_i^* in $\mathscr{F}^*$ also corresponding to λ_i may be constructed by passing from W_i in $\mathscr{E}$ to w_i in $\mathscr{F}$ defined by the natural isomorphism* (6.1.18) *and then setting*

$$w_i^* = \mathsf{A}_d w_i. \tag{6.5.10}$$

The proof requires constructing a π^*-orthonormal basis $\mathbf{W}^*$ of $\mathscr{E}^*$ such that A may be described as

$$\mathbf{W}^* \rightarrow \mathbf{J}\mathbf{W}, \tag{6.5.11}$$

where $\mathbf{W}$ is a π-orthonormal basis of $\mathscr{E}$ whose first r elements $W_1, W_2, \ldots, W_r$ are eigenvectors as above and where $\mathbf{J}$ is a $q \times p$ matrix whose elements are all zero except that the (i, i) element is $\lambda_i^{1/2}$ for $i = 1, 2, \ldots, r$. The construction proceeds as follows: $W_{r+1}^*, W_{r+2}^*, \ldots, W_q^*$ is any π^*-orthonormal basis of the $(q - r)$-dimensional subspace of $\mathscr{E}^*$ which maps into $\emptyset$ in $\mathscr{E}$ under A. The r-dimensional subspace spanned by $W_1^*, W_2^*, \ldots, W_r^*$ is then determined as the π^*-orthogonal complement of the subspace spanned by $W_{r+1}^*, W_{r+2}^*, \ldots, W_q^*$. Under A, this r-dimensional subspace of $\mathscr{E}^*$ is in one-one correspondence with an r-dimensional range space in $\mathscr{E}$, and from (6.5.8) this range space must be the space spanned by $W_1, W_2, \ldots, W_r$. Finally $W_1^*, W_2^*, \ldots, W_r^*$ may be defined from (6.5.11), and the π^*-orthonormality of $W_1^*, W_2^*, \ldots, W_r^*$ follows from (6.5.8) together with the eigenvalue properties $(W_i, W_j)_{\mathsf{A}} = \lambda_i$ or 0 depending on whether $i = j$ or $i \neq j$.

In terms of the dual bases $\mathbf{w}^*$ and $\mathbf{w}$ of $\mathbf{W}^*$ and $\mathbf{W}$, the dual of (6.5.11) may be written

$$\mathbf{w} \rightarrow \mathbf{J}'\mathbf{w}^*. \tag{6.5.12}$$

It follows directly from (6.5.9) and (6.5.12) that the eigenvalues and eigenvectors of $\pi_{\mathsf{A}_d}^*$ relative to π_d^* are as stated.

The theorem just proved may be stated more simply in purely analytical terms as follows.

Theorem 6.5. *If* $\mathbf{A}$ *is a given* $q \times p$ *matrix of rank* r, *then the* r *nonzero eigenvalues* $\lambda_1, \lambda_2, \ldots, \lambda_r$ *of* $\mathbf{A}'\mathbf{A}$ *are the same as the* r *nonzero eigenvalues of* $\mathbf{AA}'$. *Moreover, if a set of eigenvectors of* $\mathbf{A}'\mathbf{A}$ *is given by the rows of an* $r \times p$ *matrix* $\mathbf{C}_1$, *then corresponding eigenvectors of* $\mathbf{AA}'$ *are given by the rows of the* $r \times q$ *matrix* $\mathbf{C}_1\mathbf{A}'$.

The proof here requires simply that $\mathbf{A}$ be interpreted as a linear transformation $\mathbf{V}^* \to \mathbf{AV}$ from a Euclidean space $\mathscr{E}^*$ with π^*-orthonormal basis $\mathbf{V}^*$ to a Euclidean space $\mathscr{E}$ with a π-orthonormal basis $\mathbf{V}$. Then $\mathbf{A}'\mathbf{A}$ is the inner product matrix of π_{A} relative to the bases $\mathbf{V}$ and, dually, $\mathbf{AA}'$ is the inner product matrix of $\pi^*_{\mathsf{A}_d}$ relative to the basis $\mathbf{v}^*$ of $\mathscr{F}^*$ dual to $\mathbf{V}^*$ in $\mathscr{E}^*$. With this identification, Theorem 6.5 is simply an analytic statement of the preceding vector result.

6.6 THE DUAL OF A SEMI-DEFINITE INNER PRODUCT

The definition (6.1.14) of the dual inner product assumed that $(V, V) > 0$ for all V. Suppose that an inner product π over $\mathscr{E}$ is of rank $f < p$ and that $(V, V) = 0$ for V in the $(p - f)$-dimensional subspace $\mathscr{U}$ of $\mathscr{E}$. The dual of $\mathscr{U}$ is an f-dimensional subspace $\mathscr{U}_d$ in $\mathscr{F}$. The natural dual of the semi-definite inner product π is a *partial inner product* π_d, defined only over the subspace $\mathscr{U}_d$ of $\mathscr{F}$.

Consider any basis $\mathbf{W}$ of $\mathscr{E}$ such that $W_1, W_2, \ldots, W_f$ span a subspace of $\mathscr{E}$ complementary to $\mathscr{U}$ and $W_{f+1}, W_{f+2}, \ldots, W_p$ span $\mathscr{U}$. Since the inner product has full rank over this f-dimensional subspace, the first set $W_1, W_2, \ldots, W_f$ may be chosen to be orthonormal according to π, while $W_{f+1}, W_{f+2}, \ldots, W_p$ must have zero norms and zero inner products with every V in $\mathscr{E}$. The dual basis $w_1, w_2, \ldots, w_p$ has the property that $w_1, w_2, \ldots, w_f$ span $\mathscr{U}_d$, so that an inner product may be defined over $\mathscr{U}_d$ taking $w_1, w_2, \ldots, w_f$ to be orthonormal, thus defining the dual partial inner product mentioned above. The reader may check that any basis $\mathbf{W}$ such that $W_1, W_2, \ldots, W_f$ are orthonormal according to π produces the same inner product π_d on $\mathscr{U}_d$, so that the definition is unique. He should also show that the uniqueness property does not hold if $w_1, w_2, \ldots, w_p$ are taken to be orthonormal to define an inner product over all of $\mathscr{F}$.

Conversely, if a partial inner product is defined over an f-dimensional subspace of $\mathscr{F}$, then one may recover the dual semi-definite inner product in $\mathscr{E}$ by specifying a basis $w_1, w_2, \ldots, w_p$ of $\mathscr{F}$ such that $w_1, w_2, \ldots, w_f$ are orthonormal according to the partial inner product. Note that $w_1, w_2, \ldots, w_f$ uniquely determine the $(p - f)$-dimensional subspace of $\mathscr{E}$ spanned by $W_{f+1}, W_{f+2}, \ldots, W_p$ so that the subspace $\mathscr{U}$ of $\mathscr{E}$ on which $(V, V) = 0$ is uniquely determined. The reader may check that any orthogonal transformation of $w_1, w_2, \ldots, w_f$ yields the same semi-definite inner product over $\mathscr{E}$ when $W_1, W_2, \ldots, W_f$ are taken to be orthonormal and $\mathscr{U}$ is uniquely determined as described.

Geometrically, the above theory states that the dual of an ellipsoidal cylinder extending to infinity along a family of $(p - f)$-dimensional hyperplanes is an ellipsoid lying in an f-dimensional hyperplane. This might have been expected, since to stretch an ellipsoid along conjugate axes by given factors is to shrink the dual ellipsoid along the dual conjugate axes by the same factors, so that infinite length axes in the original ellipsoid should result in zero length axes in the dual ellipsoid. Note also that, although the dual inner product is defined only over the hyperplane $\mathscr{U}_d$ through the origin, it can be used to define the concepts of length, volume, and angle in any hyperplane $v + \mathscr{U}_d$ parallel to $\mathscr{U}_d$. One need only translate the geometric figures in $v + \mathscr{U}_d$ back to $\mathscr{U}_d$ and use the definitions applicable in $\mathscr{U}_d$.

Some related analytic theory follows. Suppose that $\mathbf{Q}$ is a positive semi-definite $p \times p$ symmetric matrix of rank f which is regarded as an inner product matrix relative to the basis $\mathbf{U}$ of $\mathscr{E}$. It may be of interest to locate the subspace $\mathscr{U}_d$ of $\mathscr{F}$ over which the partial inner product is defined. A way to do this is to find a $p \times f$ matrix $\mathbf{D}_1$ such that

$$[w_1, w_2, \ldots, w_f]' = \mathbf{D}_1'\mathbf{u}, \tag{6.6.1}$$

where $\mathbf{u}$ is the dual basis of $\mathbf{U}$ and $w_1, w_2, \ldots, w_f$ is an orthonormal set spanning $\mathscr{U}_d$ as above. It will now be proved that a $p \times f$ *matrix* $\mathbf{D}_1$ *obeys the relation* (6.6.1) *for some choice of* $w_1, w_2, \ldots, w_f$ *if and only if*

$$\mathbf{Q} = \mathbf{D}_1\mathbf{D}_1'. \tag{6.6.2}$$

Suppose first that (6.6.1) holds. Then $\mathbf{D}_1$ is the first f columns of a matrix $\mathbf{D}$ such that $\mathbf{w} = \mathbf{D}'\mathbf{u}$ or equivalently that $\mathbf{U} = \mathbf{DW}$. Now $\mathbf{W}$ has the inner product matrix $\mathbf{I}_f$ whose first f diagonal elements are unity and whose remaining elements are zero. It follows that $\mathbf{Q} = \mathbf{DI}_f\mathbf{D}'$. But $\mathbf{DI}_f\mathbf{D}' = \mathbf{D}_1\mathbf{D}_1'$ and thus (6.6.2) follows. To prove the converse result, noting that (6.6.2) implies that $\mathbf{D}_1$ has rank f, add any $p - f$ columns to $\mathbf{D}_1$ to make it a $p \times p$ nonsingular matrix $\mathbf{D}$. Then the argument simply operates in reverse.

The computations required to produce an instance of $\mathbf{D}_1$ from a given $\mathbf{Q}$ were essentially given in Section 4.4. By carrying out successive orthogonalization on $\mathbf{Q}$, one gets f columns of $\mathbf{B}$ corresponding to nonzero (U_s^*, U_s^*). Dividing these f columns by the corresponding $(U_s^*, U_s^*)^{-1/2}$ gives a particular choice of $\mathbf{D}_1$.

In Section 3.6 it was shown that a linear transformation may be used to carry an inner product in the reverse direction in a natural way. Dualizing that theory shows how to carry a partial inner product in the forward direction into another partial inner product. Specifically, suppose that A_d denotes any linear transformation from $\mathscr{F}$ to $\mathscr{F}^*$ and that π_d is a partial inner product over $\mathscr{F}$. Denoting the duals of A_d, $\mathscr{F}$, $\mathscr{F}^*$, and π_d by A, $\mathscr{E}$, $\mathscr{E}^*$, and π, the theory of

Section 3.6 shows how the transformation A from $\mathscr{E}^*$ to $\mathscr{E}$ induces from π a wide sense inner product π^* over $\mathscr{E}^*$, whose dual in turn defines the desired partial inner product π_d^* over $\mathscr{F}^*$ induced by A_d from π_d.

The roundabout definition of π_d^* from π_d via dual theory may be replaced by the simple and direct geometric characterization given in Theorem 6.6 which follows. Let π_d and π_d^* denote ambiguously either the partial inner products π_d and π_d^* or their corresponding ellipsoids lying in the subspaces over which the partial inner products are defined. Define the *shadow of* π_d *in* $\mathscr{F}$ *under the transformation* A_d to be the set of points in $\mathscr{F}^*$ which are the transforms of some point in π_d.

Theorem 6.6. *π_d^* is the shadow of π_d under the transformation A_d.*

In other words, the set of points v in $\mathscr{F}$ for which $(v, v)_d$ is defined and less than or equal to unity maps into the set of points v^* in $\mathscr{F}^*$ for which $(v^*, v^*)_d^*$ is defined and less than or equal to unity. The proof requires consideration of several subspaces: the subspace $\mathscr{W}_d$ in $\mathscr{F}$ in which π_d lies, the subspace $\mathscr{V}_d$ in $\mathscr{F}$ which maps into the origin in $\mathscr{F}^*$ under A_d, the subspace $\mathscr{N}_d$ of $\mathscr{W}_d$ orthogonal to $\mathscr{W}_d \cap \mathscr{U}_d$ according to π_d^*, and the subspace $\mathscr{U}_d^*$ of $\mathscr{F}^*$ consisting of the maps of points in $\mathscr{U}_d$, this being the subspace in which the shadow lies. Perhaps the most straightforward approach is to set up an orthonormal coordinate system in $\mathscr{W}_d$ a part of which spans $\mathscr{U}_d$ and a subpart of which transforms into the range space $\mathscr{U}_d^*$. In these terms it is obvious that the shadow π_d^* of π_d is an ellipsoid in $\mathscr{U}_d^*$. To check that π_d^* is the right ellipsoid requires carefully setting up dual concepts and checking out the original definition of π_d^* by the roundabout route. Further details are left to the reader.

It is worth stating formally that:

Corollary 6.6. *The shadow of any ellipsoid in a hyperplane under a linear transformation is again an ellipsoid in a hyperplane. The center of the shadow ellipsoid is the transform of the center of the original ellipsoid.*

For hyperplanes through the origin and narrow sense linear transformations the corollary is an immediate consequence of Theorem 6.6. But the corollary is obvious for translations of a space into itself, and so holds for wide sense linear transformations and for ellipsoids in arbitrary hyperplanes with arbitrary centers.

Corollary 6.6 is illustrated in Figure 7.3.2.

6.7 EXERCISES

6.1.1 Show that a linear functional v defined over a vector space $\mathscr{E}$ is uniquely determined by its values over any basis of $\mathscr{E}$.

6.1.2 What is meant by the assertion that two linear functionals v_1 and v_2 over $\mathscr{E}$ are different? Show that v_1 and v_2 may agree on a subspace of dimension $p - 1$ and still be different.

6.1.3 Suppose that $\{V, v\}$ is defined to be zero for all V in $\mathscr{E}$ and v in $\mathscr{F}$. Does this define a bilinear product over $\mathscr{E}$ and $\mathscr{F}$? If so, can this bilinear product be used in showing that $\mathscr{E}$ and $\mathscr{F}$ are dual?

6.1.4 Suppose that $V_1, V_2, \ldots, V_p$ is a basis of $\mathscr{E}$ and $v_1, v_2, \ldots, v_p$ is the dual basis in $\mathscr{F}$. Express in terms of $v_1, v_2, \ldots, v_p$ the dual basis in $\mathscr{F}$ of the basis $V_1, V_2 + V_1, V_3 + V_1, \ldots, V_p + V_1$ in $\mathscr{E}$.

6.1.5 Suppose that $\mathscr{U}$ and $\mathscr{V}$ are complementary subspaces of $\mathscr{E}$. Show that the dual subspaces $\mathscr{U}_d$ and $\mathscr{V}_d$ are complementary in $\mathscr{F}$.

6.1.6 Show that the isomorphism between $\mathscr{E}$ and $\mathscr{F}$ defined by

$$\boldsymbol{\alpha}'\mathbf{V} \to \boldsymbol{\alpha}'\mathbf{v},$$

where $\mathbf{V}$ and $\mathbf{v}$ are dual bases, is not coordinate-free. Show further that the isomorphism is the same for a basis $\mathbf{W} = \mathbf{C}\mathbf{V}$ and its dual $\mathbf{w}$ if and only if $\mathbf{C}$ is an orthogonal matrix.

6.1.7 Since variable-space is the dual of individual-space, it must follow that variable-space has a natural isomorphism to the space of linear functionals over individual-space. What is this isomorphism, and what is the bilinear product which it preserves? In other words, how does a linear functional over individual-space determine a variable?

6.1.8 Suppose that $\mathscr{E}$ and $\mathscr{F}$ are dual spaces with bilinear product $\{V, v\}$. Suppose that $V_1, V_2, \ldots, V_p$ and $v_1, v_2, \ldots, v_p$ are dual bases of $\mathscr{E}$ and $\mathscr{F}$, and that a is any element of $\mathscr{F}$. Show that

$$a = \sum_{i=1}^{p} \{V_i, a\}v_i.$$

What is the statistical interpretation of this formula? What is the dual formula and what is the statistical interpretation of the dual formula?

6.1.9 Suppose that $\mathscr{E}$ and $\mathscr{F}$ are dual spaces with bilinear product function $\{V, v\}$. Show that $\mathscr{E}$ and $\mathscr{F}$ are also dual spaces with the rescaled bilinear product function $\lambda\{V, v\}$ for any $\lambda \neq 0$. Show that the duality of a pair of subspaces $\mathscr{U}$ and $\mathscr{U}_d$ is not affected by rescaling the bilinear product function, but that the concepts of dual basis and dual inner product are also subject to rescaling.

6.1.10 Show that the natural isomorphism between a pair $\mathscr{E}$ and $\mathscr{F}$ of dual Euclidean spaces defined in Section 6.1 is simply the identity relationship when $\mathscr{E}$ is regarded as self-dual.

6.1.11 Show that a definition of dual inner product alternative to (6.1.14) is given by

$$(v, v)_d = \sup_V \frac{\{V, v\}^2}{(V, V)}.$$

6.2.1 What is the dual of the parallelotope in $\mathscr{E}$ with vertices $W + \sum_1^p c_iV_i$ where each c_i is 0 or 1?

6.2.2 The dual of a pair of r-dimensional hyperplanes lying in an $(r + 1)$-dimensional hyperplane through the origin is a pair of $(p - r - 1)$-dimensional hyperplanes with a common intersection in the hyperplane at infinity. Show that these two $(p - r - 1)$-dimensional hyperplanes are parallel.

6.2.3 Suppose that V in $\mathscr{E}$ and v in $\mathscr{F}$ are said to be *biorthogonal* if $\{V, v\} = 0$. Show that V and v are biorthogonal if and only if the line V is parallel to the dual hyperplane in $\mathscr{E}$ of v in $\mathscr{F}$. A subspace $\mathscr{U}$ of $\mathscr{E}$ is said to be *biorthogonal* to a subspace $\mathscr{V}$ of $\mathscr{F}$ if $\{V, v\} = 0$ for all V in $\mathscr{U}$ and v in $\mathscr{V}$. What is the geometric interpretation in $\mathscr{E}$ of the relationship between $\mathscr{U}$ and the dual of $\mathscr{V}$?

6.3.1 Suppose that $\mathscr{E}$ and $\mathscr{F}$ are dual p-dimensional spaces with bilinear product function $\{V, v\}$ for V in $\mathscr{E}$ and v in $\mathscr{F}$. Suppose that $\mathbf{V}$ is any basis in $\mathscr{E}$ and $\mathbf{w}$ any basis in $\mathscr{F}$. Define the bilinear product matrix $\mathbf{R}$ of $\mathbf{V}$ and $\mathbf{w}$ to be a $p \times p$ matrix whose (i, j) element is $\{V_i, w_j\}$. Show how to express $\{\boldsymbol{\alpha}\mathbf{V}, \boldsymbol{\beta}\mathbf{w}\}$ in terms of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\mathbf{R}$. How is $\mathbf{R}$ related to the bilinear product matrix of bases $\mathbf{V}^* = \mathbf{AV}$ and $\mathbf{w}^* = \mathbf{Bw}$? What is the bilinear product matrix of a pair of dual bases?

6.3.2 Suppose that $\mathbf{V}$ and $\mathbf{W}$ are any two orthonormal bases of a Euclidean space $\mathscr{E}$, and that $\mathbf{v}$ and $\mathbf{w}$ are the corresponding dual bases of $\mathscr{F}$. Show directly that the inner products defined by regarding $\mathbf{v}$ or $\mathbf{w}$ to be orthonormal are identical. Show likewise that the isomorphisms between $\mathscr{E}$ and $\mathscr{F}$ defined by $\mathbf{V} \to \mathbf{v}$ or $\mathbf{W} \to \mathbf{w}$ are identical.

6.3.3 A linear transformation of $\mathscr{E}$ into itself has rank p when it carries a basis $\mathbf{V}$ of $\mathscr{E}$ into another basis $\mathbf{V}^*$ of $\mathscr{E}$. The dual linear transformation carries the dual basis $\mathbf{v}$ of $\mathbf{V}$ into the dual basis $\mathbf{v}^*$ of $\mathbf{V}^*$. Show that the correspondence between points of $\mathscr{E}$ and $(p - 1)$-dimensional hyperplanes of $\mathscr{F}$ is preserved after dual linear transformations are applied to both $\mathscr{E}$ and $\mathscr{F}$.

6.4.1 Show that $\mathbf{Q}$ is a positive definite symmetric matrix if and only if $\mathbf{Q}^{-1}$ is also.

6.4.2 Show that the dual of formula (4.2.17) is

$$\mathbf{u}_{1.2} = \mathbf{u}_1 + \mathbf{H}'_{21}\mathbf{u}_2.$$

6.5.1 Draw the duals of Figs. 5.2.1 and 5.2.2, showing λ_1 and λ_2 as ratios of lengths in $\mathscr{F}$.

6.5.2 Give a purely analytic proof of Theorem 6.5.

6.6.1 Check the statements made in the last two sentences of paragraph two of Section 6.6, and in the last sentence of paragraph three.

6.6.2 Show how definition (6.1.14) may be modified to provide alternative definitions of the dual of a semi-definite inner product and the dual of a partial inner product.

6.6.3 Suppose that $\mathbf{D}_1$ and $\mathbf{D}_1^*$ are both $p \times f$ matrices satisfying (6.6.2). Show that $\mathbf{D}_1 = \mathbf{D}_1^*\mathbf{G}'$ for some orthogonal matrix $\mathbf{G}$.

6.6.4 Suppose that $\dot{\mathbf{Q}}$ is a pseudoinverse of a positive semi-definite symmetric $p \times p$ matrix $\mathbf{Q}$. The inner products π and $\dot{\pi}$ defined by $\mathbf{Q}$ and $\dot{\mathbf{Q}}$ relative to a pair $\mathbf{V}$ and $\mathbf{v}$ of dual bases are both semi-definite. How does $\dot{\pi}$ relate to the partial inner product π_d?

6.6.5 Suppose that $\mathscr{E}$ and $\mathscr{F}$ are a pair of dual spaces with full rank inner products π and π_d. Suppose that π^* is an inner product of rank $r \leq p$ over $\mathscr{E}$ and that π^* is its dual (partial) inner product over $\mathscr{F}$. Show that the r nonzero eigenvalues of π^* relative to π are the inverses of the r eigenvalues of π^* relative to π_d in the subspace over which π^* is defined. Show also that the isomorphism (6.1.18) carries corresponding sets of eigenvectors into one another, where if one set is π-orthonormal its image is π_d-orthonormal.

6.6.6 Demonstrate Corollary 6.6 directly.

6.6.7 Complete the proof of Theorem 6.6.

PART 3

DATA ANALYSIS

CHAPTER 7

ONE SAMPLE OF INDIVIDUALS: BASIC THEORY

7.1 INTRODUCTION

This is a purely theoretical chapter, while the remaining chapters of Part 2 mix theory with examples presenting analyses of observed data. The concern of this chapter is to introduce various concepts related to the sample mean and sample covariance of a sample of n individuals each observed on a set of p variables. The sample individuals will be denoted by $a_1, a_2, \ldots, a_n$ and the observable variables by $V_1, V_2, \ldots, V_p$. The term *p-variate sample of size n* will be used to describe the resulting data.

Such a sample may be identified mathematically with a set of points $a_1, a_2, \ldots, a_n$ in the individual-space $\mathscr{F}$ dual to the variable-space $\mathscr{E}$ spanned by $V_1, V_2, \ldots, V_p$. The theory of Chapters 7 through 11 is completely derived from this simple mathematical formulation.

The variables $V_1, V_2, \ldots, V_p$ and individuals $a_1, a_2, \ldots, a_n$ define $p \times n$ *quantities* or *statistics*. The quantity associated with V_i and a_j may be denoted by $V_j^{(i)}$, and the value of $V_j^{(i)}$ may be denoted by $X_j^{(i)}$ for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$. In principle any numerical operation with the $X_j^{(i)}$ defines the value of another quantity which is a function of the $V_j^{(i)}$.

In practice it is unwieldy to carry along double notation and terminology for such quantities and their values. Consequently, accepting the lesser evil of some ambiguity, symbols appropriate to the values of quantities will be used throughout. At the same time the names of quantities may be used in reference to the values of those quantities. For example, $\bar{\mathbf{X}}$ will be used to denote a $1 \times p$ vector of sample mean values, but for brevity $\bar{\mathbf{X}}$ will usually be referred to simply as the sample mean vector.

Note that the methods of Chapters 7 through 11 may be applied to a finite population in place of a finite sample. The methods can also be extended to infinite populations through a limiting argument, but this is usually done within the framework of probability theory, i.e., the concept of an infinite population may be related to the concept of a probability distribution. In this sense, the discussion of infinite populations is postponed until Chapter 12.

7.2 DEFINITIONS

The *sample mean (individual)* m of a given sample $a_1, a_2, \ldots, a_n$ is

$$m = \frac{1}{n}\sum_{i=1}^{n} a_i. \tag{7.2.1}$$

The value of m on a variable V in variable-space $\mathscr{E}$ will be denoted by $m(V)$ and should be called the *sample mean (value) of the variable V*. If the values of the individuals $a_1, a_2, \ldots, a_n$ on a variable V are denoted by $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$, then from (6.1.2) and (7.2.1)

$$m(V) = \frac{1}{n}\sum_{i=1}^{n} X^{(i)}, \tag{7.2.2}$$

where the right side of (7.2.2) is often abbreviated to $\bar{X}$. The idea here is simple and familiar, but note the distinction between a sample mean m, which is a point in the individual-space $\mathscr{F}$ dual to $\mathscr{E}$, and a value $\bar{X}$ of such a sample mean.

For purposes of computation rather than interpretation, it is convenient to have terminology for an alternative but equivalent linear quantity and its associated value. These are the *sample sum individual*

$$t = \sum_{i=1}^{n} a_i = nm \tag{7.2.3}$$

and its associated *sample sum value*

$$t(V) = \sum_{i=1}^{n} X^{(i)} = n\bar{X}, \tag{7.2.4}$$

for the variable V.

The *sample covariance* is an inner product function defined over variable-space $\mathscr{E}$. If the sample values of the variables V and W on individuals $a_1, a_2, \ldots, a_n$ are denoted by $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ and $Y^{(1)}, Y^{(2)}, \ldots, Y^{(n)}$, respectively then the *sample covariance (value) of V and W is defined to be*

$$\operatorname{cov}(V, W) = \frac{1}{n-1}\sum_{i=1}^{n}(X^{(i)} - \bar{X})(Y^{(i)} - \bar{Y}) = \frac{1}{n-1}\left[\sum_{i=1}^{n} X^{(i)}Y^{(i)} - n\bar{X}\bar{Y}\right]. \tag{7.2.5}$$

In the special case where V and W are the same, covariance is called *variance* and is written $\operatorname{cov}(V, V) = \operatorname{var}(V)$. The square root of $\operatorname{var}(V)$, which is a quantity in the same units as V, is called the *standard deviation* of V. The reader should recheck that the sample covariance defined by (7.2.5) satisfies the requirements for an inner product over $\mathscr{E}$, where this inner product may be definite or semi-definite.

Terminology for several related sample-based inner products will also be useful. The *sample raw sum inner product* may be defined, in the notation of (7.2.5), by

$$(V, W)_Q = \sum_{i=1}^{n} X^{(i)}Y^{(i)}, \tag{7.2.6}$$

and similarly the *sample corrected sum inner product* may be defined by

$$(V, W)_T = \sum_{i=1}^{n}(X^{(i)} - \bar{X})(Y^{(i)} - \bar{Y}). \tag{7.2.7}$$

These different inner products are related by

$$(n-1)\operatorname{cov}(V, W) = (V, W)_T = (V, W)_Q - nm(V)m(W). \tag{7.2.8}$$

The definitions of sample mean, sample covariance, and related concepts have been given in coordinate-free terms. Most often, however, a sample is specified by its $n \times p$ *data matrix* $\mathbf{X}$ whose (i, j) element $X_j^{(i)}$ gives the observed value of the variable V_j on the sample individual a_i, for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$.

The sample mean individual m is determined by its $1 \times p$ coordinate vector $\bar{\mathbf{X}}$ relative to the basis $\mathbf{v}$ in $\mathscr{F}$ dual to $\mathbf{V}$ in $\mathscr{E}$, i.e., by

$$m = \bar{\mathbf{X}}\mathbf{v}. \tag{7.2.9}$$

If the rows of $\mathbf{X}$ are denoted by $\mathbf{X}^{(i)}$ for $i = 1, 2, \ldots, n$, then the sample individual a_i may be expressed as

$$a_i = \mathbf{X}^{(i)}\mathbf{v} \tag{7.2.10}$$

for $i = 1, 2, \ldots, n$ (cf. Exercise 6.1.8), whence $\bar{\mathbf{X}}$ in (7.2.9) is given by the ordinary mean of the n rows of $\mathbf{X}$, i.e., by

$$\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}^{(i)}. \tag{7.2.11}$$

The elements of $\bar{\mathbf{X}}$ are the values of m for the basis $\mathbf{V}$ of $\mathscr{E}$, so that the value $m(V)$ for any $V = \boldsymbol{\alpha}\mathbf{V}$ in $\mathscr{E}$ is given by

$$m(V) = \boldsymbol{\alpha}\bar{\mathbf{X}}'. \tag{7.2.12}$$

The formulas analogous to (7.2.9), (7.2.11), and (7.2.12) when m is replaced by the sample sum individual t are left for the reader to express.

Similarly, the sample covariance inner product is determined by its inner product matrix $\mathbf{S}$ relative to a basis $\mathbf{V}$, i.e., by the matrix $\mathbf{S}$ whose (i, j) element S_{ij} is given by

$$S_{ij} = \operatorname{cov}(V_i, V_j). \tag{7.2.13}$$

$\mathbf{S}$ will be called the *sample covariance matrix* of the set of variables $\mathbf{V}$. For any $V = \boldsymbol{\alpha}\mathbf{V}$ and $W = \boldsymbol{\beta}\mathbf{V}$ in $\mathscr{E}$,

$$\operatorname{cov}(V, W) = \boldsymbol{\alpha}\mathbf{S}\boldsymbol{\beta}' \tag{7.2.14}$$

as in (3.1.7), i.e., as with any inner product only the covariance matrix for any basis is required to determine the covariance between any pair of variables.

The additional inner products $(V, W)_Q$ and $(V, W)_T$ defined in (7.2.6) and (7.2.7) have inner product matrices relative to **V** which will be denoted by **Q** and **T**, respectively. In matrix terms, **Q** and **T** may be expressed using the data matrix **X** via

$$\mathbf{Q} = \sum_{i=1}^{n} \mathbf{X}^{(i)\prime}\mathbf{X}^{(i)} = \mathbf{X}'\mathbf{X} \tag{7.2.15}$$

and

$$\mathbf{T} = \sum_{i=1}^{n} (\mathbf{X}^{(i)} - \bar{\mathbf{X}})'(\mathbf{X}^{(i)} - \bar{\mathbf{X}}) = \mathbf{X}'\mathbf{X} - n\bar{\mathbf{X}}'\bar{\mathbf{X}}, \tag{7.2.16}$$

while the relations (7.2.8) may be written

$$(n - 1)\mathbf{S} = \mathbf{T} = \mathbf{Q} - n\bar{\mathbf{X}}'\bar{\mathbf{X}}. \tag{7.2.17}$$

The dual of the sample covariance inner product over $\mathscr{E}$ will be called the *sample concentration inner product* over $\mathscr{F}$. If the sample covariance matrix is **S** relative to a basis **V** of $\mathscr{E}$, and if the sample covariance has full rank p, then the inner product matrix of the sample concentration relative to the dual basis **v** in $\mathscr{F}$ of **V** in $\mathscr{E}$ is given by $\mathbf{S}^{-1}$. The case of semi-definite sample covariance is pursued further in Section 7.7. Dual inner products to the sample raw sum inner product and the sample corrected sum inner products will occasionally have roles to play, but no specific terminology will be assigned to them.

7.3 REPRESENTATION OF A SAMPLE WITH INDIVIDUALS AS POINTS

A given p-variate sample of size n may be visualized geometrically as consisting of the n points $a_1, a_2, \ldots, a_n$ in the p-dimensional affine individual-space $\mathscr{F}$. The sample mean and sample covariance may be described geometrically by the sample mean point m and the inner product ellipsoid of the sample concentration. This ellipsoid consists of points at unit distance or less from the origin $\varnothing$ in $\mathscr{F}$ according to the concentration inner product, and will be called the (*origin-centered*) *ellipsoid of concentration.* The result of translating this ellipsoid into an ellipsoid with center m will be called the *mean-centered ellipsoid of concentration* (cf. Cramér, 1946). The mean-centered ellipsoid of concentration consists of points at most unit distance from the mean, and provides a method for representing the location and scatter of a sample in a single geometric figure, at least to the extent to which the sample mean and sample covariance alone are able to provide such a representation.

Before illustrating these concepts with drawings it may help to be specific about the mechanics of plotting the n sample points in $\mathscr{F}$ from an $n \times p$ data matrix **X** representing a p-variate sample of size n. Each individual a_i is represented by the corresponding row i of **X**, namely

$$\mathbf{X}^{(i)} = [X_1^{(i)}, X_2^{(i)}, \ldots, X_p^{(i)}], \tag{7.3.1}$$

which gives the coordinates of a_i relative to the dual basis **v** of **V** as in (7.2.10). The first step in plotting is to lay out the coordinate axes $v_1, v_2, \ldots, v_p$ forming **v**. This may be done physically on a plane piece of paper if $p = 2$, or in space if $p = 3$, but may only be done conceptually for $p > 3$. The case $p = 2$ is pictured in Fig. 7.3.1 where the axes v_1 and v_2 were first laid down and the points a_i were plotted using the formula $a_i = X_1^{(i)}v_1 + X_2^{(i)}v_2$ together with the standard physical representations of vector multiplication and addition (cf. Fig. 2.4.1). Note that while v_1 and v_2 are pictured in Fig. 7.3.1 as orthogonal, this is merely a convention inessential in such representations of the sample. The sample mean point is plotted in the same way from $m = \bar{\mathbf{X}}\mathbf{v}$. The points $v = \mathbf{x}\mathbf{v}$ belonging to the origin-centered ellipsoid of concentration satisfy the relation

$$\mathbf{x}\mathbf{S}^{-1}\mathbf{x}' \leq 1 \tag{7.3.2}$$

and those belonging to the mean-centered ellipsoid of concentration satisfy the relation

$$(\mathbf{x} - \bar{\mathbf{X}})\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{X}})' \leq 1, \tag{7.3.3}$$

assuming of course that $\mathbf{S}^{-1}$ exists.

The remainder of this section is concerned with the behavior of samples and their associated mean points and concentration ellipsoids under linear transformations operating on $\mathscr{F}$. The important fact here, namely Theorem 7.3, relates the shadow of a sample with the shadow of its mean-centered concentration ellipsoid. Recall the definition and theory of shadows given in Section 6.6.

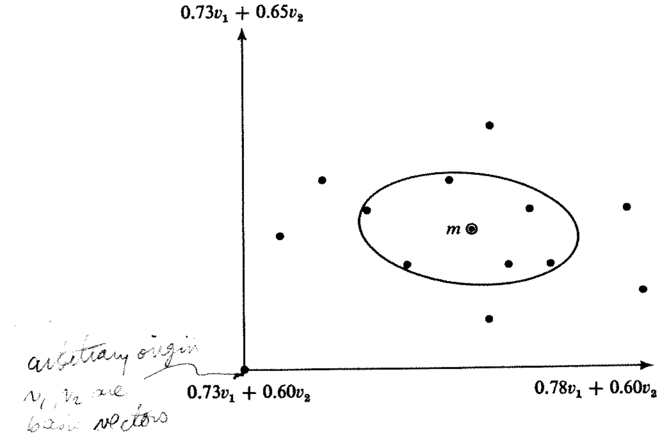

Fig. 7.3.1. The sample data of V_1 and V_2 from Example 8.1 plotted as 12 points in the individual-space dual to the variable-space spanned by V_1, V_2. The mean point and mean-centered concentration ellipse of the sample are also shown.

Theorem 7.3. *Suppose that π_d is the mean-centered ellipsoid of concentration of a sample $a_1, a_2, \ldots, a_n$ in $\mathscr{F}$. Suppose that a linear transformation A_d from $\mathscr{F}$ to $\mathscr{F}^*$ carries the sample $a_1, a_2, \ldots, a_n$ in $\mathscr{F}$ into a sample a_1^*, $a_2^*, \ldots, a_n^*$ in $\mathscr{F}^*$. Then the mean-centered concentration ellipsoid π_d^* of $a_1^*, a_2^*, \ldots, a_n^*$ in $\mathscr{F}^*$ is the shadow of π_d under the transformation A_d.*

This theorem holds when A_d is a wide sense linear transformation, but it is trivial under translations, and so need only be proved when the sample mean in $\mathscr{F}$ is the origin and A_d carries the origin in $\mathscr{F}$ into the origin in $\mathscr{F}^*$. In this case the sample mean of $a_1^*, a_2^*, \ldots, a_n^*$ is the origin in $\mathscr{F}^*$ and both concentration ellipsoids are origin-centered. The individual-spaces $\mathscr{F}$ and $\mathscr{F}^*$ have dual variable-spaces $\mathscr{E}$ and $\mathscr{E}^*$. The transformation A_d determines its dual transformation A from $\mathscr{E}^*$ to $\mathscr{E}$ whose basic property (6.1.13) implies in the present context that the sample values are preserved under the transformation A, i.e., that $a_i(V) = a_i^*(V^*)$ for $i = 1, 2, \ldots, n$ where $V = \mathsf{A}V^*$ and V^* is any variable in $\mathscr{E}^*$. This preservation of sample values means that the inner product over $\mathscr{E}$ induced by A from the covariance inner product over $\mathscr{E}^*$ of the sample a_1^*, $a_2^*, \ldots, a_n^*$ is the sample covariance inner product of the sample $a_1, a_2, \ldots, a_n$. Thus π_d^* may be produced by the roundabout route described in Section 6.6, and Theorem 7.3 follows directly from Theorem 6.6.

The shadow theory is easily visualized when A_d is a linear projection. For example, consider projection along the family of hyperplanes parallel to the subspace spanned by $v_{r+1}, v_{r+2}, \ldots, v_p$ into the subspace spanned by $v_1, v_2, \ldots, v_r$ where $\mathbf{v} = [v_1, v_2, \ldots, v_p]'$ is a basis of $\mathscr{F}$. Of course, any narrow sense linear projection may be defined in this way for properly chosen bases. Such a projection carries the sample defined by the data matrix $\mathbf{X}$ into the sample defined by the same data matrix with the last $p - r$ columns replaced by zeros. Thus the projected sample provides a representation in $\mathscr{F}$ for the reduced sample in which the data are available for $V_1, V_2, \ldots, V_r$ only. Fig. 7.3.2 is a plane drawing for the case $p = 3$ and $r = 2$ of the type of linear projection just described. It also illustrates the shadow theory of Theorem 7.3.

The sample data for $V_1, V_2, \ldots, V_r$ may be equally well represented in $\mathscr{F}$ by projection into any subspace complementary to the subspace spanned by $v_{r+1}, v_{r+2}, \ldots, v_p$ or, more generally, by projection into any hyperplane parallel to such a subspace. For example, one could have projected in Fig. 7.3.2 into any plane making a positive angle with v_3, and the projected sample would have retained essentially the same information. The reason for this is that the subspace spanned by $V_1, V_2, \ldots, V_r$ determines only the dual subspace spanned by $v_{r+1}, v_{r+2}, \ldots, v_p$ on which all of $V_1, V_2, \ldots, V_r$ assume zero values, and consequently the choice of a complementary subspace is left open. In other words, to say that the sample data are available on $V_1, V_2, \ldots, V_r$ is to say not only that the sample data on $V_{r+1}, V_{r+2}, \ldots, V_p$ are unknown but also that the sample data on any complementary subspace to $V_1, V_2, \ldots, V_r$ are unknown.

Sometimes it is helpful to consider projecting a sample into a hyperplane which contains its mean. This provides a representation of the sample data restricted to a subspace of variable-space, where the projected sample has the same mean as the original sample.

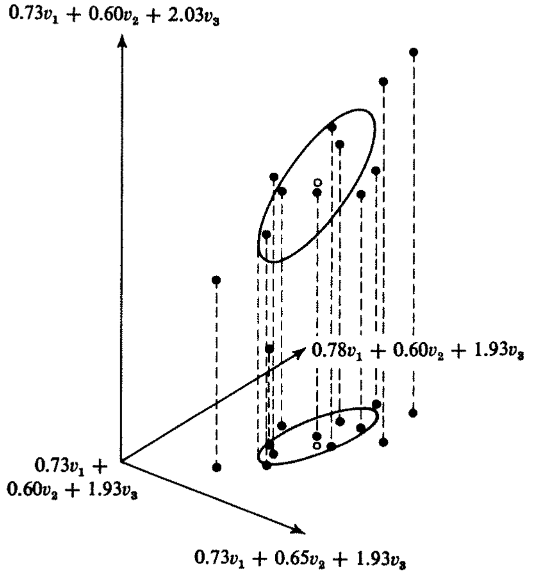
$0.73v_1 + 0.60v_2 + 2.03v_3$
$0.78v_1 + 0.60v_2 + 1.93v_3$
$0.73v_1 + 0.60v_2 + 1.93v_3$
$0.73v_1 + 0.65v_2 + 1.93v_3$

Fig. 7.3.2. The data of Example 8.1 plotted in the 3-dimensional individual-space spanned by v_1, v_2, v_3. The projection along lines of constant x_1 and x_2 into the hyperplane $\mathrm{x}_3 = 1.93$ is also shown, along with the original mean-centered ellipsoid of concentration and its shadow.

7.4 REPRESENTATION OF A SAMPLE WITH VARIABLES AS POINTS

In contrast to the geometric representation of Section 7.3, a given p-variate sample of size n may also be regarded as p points plotted in an n-dimensional space, where the points correspond to the basic variables $V_1, V_2, \ldots, V_p$. To achieve this, create an n-dimensional vector space $\mathscr{N}$ by imagining n basis vectors $N_1, N_2, \ldots, N_n$ which correspond to the n sample individuals $a_1, a_2, \ldots, a_n$, and then plot the points

$$P_j = X_j^{(1)}N_1 + X_j^{(2)}N_2 + \cdots + X_j^{(n)}N_n \tag{7.4.1}$$

for $j = 1, 2, \ldots, p$. Here P_j is a geometric representation in $\mathscr{N}$ of V_j whose coordinates are given by the jth column of the data matrix $\mathbf{X}$. More generally, any variable V in $\mathscr{E}$ with sample values $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ may be represented by

$$P = \sum_{i=1}^{n} X^{(i)} N_i. \tag{7.4.2}$$

From a mathematical point of view, (7.4.2) determines a linear transformation

$$V \to P \tag{7.4.3}$$

from $\mathscr{E}$ to $\mathscr{N}$. It is obvious but worth noting that *knowing the linear transformation* (7.4.3) *is equivalent to knowing the sample data* $\mathbf{X}$. Note also that the range space in $\mathscr{N}$ of the mapping (7.4.3) has at most dimension p which may be much less than n.

The sample mean and sample covariance have special relationships with $\mathscr{N}$, which suggests that $\mathscr{N}$ be regarded as a Euclidean space with inner product defined by regarding $N_1, N_2, \ldots, N_n$ to be orthonormal. Suppose that $\mathscr{N}_I$ denotes the one-dimensional subspace of $\mathscr{N}$ spanned by $\sum_1^n N_j$ and that $\mathscr{N}_{II}$ denotes the $(p - 1)$-dimensional subspace of $\mathscr{N}$ orthogonal to $\mathscr{N}_I$ according to the suggested inner product. Then the components of P in (7.4.2) along $\mathscr{N}_I$ and $\mathscr{N}_{II}$ are, respectively,

$$P_I = \sum_{j=1}^{n} \bar{X} N_j \tag{7.4.4}$$

and

$$P_{II} = \sum_{j=1}^{n} (X^{(j)} - \bar{X}) N_j. \tag{7.4.5}$$

From these representations the following theorem is immediately apparent.

Theorem 7.4.1. *Suppose that a given sample is represented by the linear transformation* (7.4.3) *from* $\mathscr{E}$ *to the Euclidean space* $\mathscr{N}$ *with orthonormal basis* $N_1, N_2, \ldots, N_n$. *Suppose that the variables* V *and* W *have sample values* $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ *and* $Y^{(1)}, Y^{(2)}, \ldots, Y^{(n)}$. *Suppose that the transforms* P *and* Q *of* V *and* W *decompose into* $P = P_I + P_{II}$ *and* $Q = Q_I + Q_{II}$ *along the orthogonal subspaces* $\mathscr{N}_I$ *and* $\mathscr{N}_{II}$ *defined above. Then*

$$|m(V)| = \left[\frac{1}{n}(P_I, P_I)\right]^{1/2} \tag{7.4.6}$$

and similarly for W, *while*

$$\operatorname{cov}(V, W) = \frac{1}{n-1}(P_{II}, Q_{II}), \tag{7.4.7}$$

where (P_I, P_I) *and* (P_{II}, Q_{II}) *refer to the given inner product over* $\mathscr{N}$.

Formula (7.4.6) simply says that the absolute value of $\sqrt{n}m(V)$ is the length of P_I, and this follows directly from (7.4.4). Similarly, from (7.4.5) and its analogue for Q,

$$(P_{II}, Q_{II}) = \sum_{i=1}^{n} (X^{(i)} - \bar{X})(Y^{(i)} - \bar{Y}), \tag{7.4.8}$$

which yields (7.4.7) directly.

Formula (7.4.7) suggests that for certain purposes the linear transformation

$$V \to P_{II} \tag{7.5.9}$$

from $\mathscr{E}$ to $\mathscr{N}_{II}$ may be more useful than (7.4.3). The reader may easily check that the transformation (7.4.9) together with the sample mean uniquely determines the sample, while the transformation (7.4.9) by itself determines the sample covariance. Indeed, the transformation (7.4.9) may be said to induce the inner product $(V, W)_T$ on $\mathscr{E}$ from the given inner product on $\mathscr{N}_{II}$ by setting $(V, W)_T = (P_{II}, Q_{II})$. Recall from (7.2.8) that the inner product $(V, W)_T$ is simply $\operatorname{cov}(V, W)$ rescaled by the factor $n - 1$.

The *sample correlation coefficient* between the variables V and W is defined to be

$$r = \frac{\operatorname{cov}(V, W)}{\operatorname{var}(V)^{1/2} \operatorname{var}(W)^{1/2}}. \tag{7.4.10}$$

Alternative expressions for r are

$$\begin{aligned} r &= \frac{(V, W)_T}{(V, V)_T^{1/2}(W, W)_T^{1/2}} \\ &= \frac{(P_{II}, Q_{II})}{(P_{II}, P_{II})^{1/2}(Q_{II}, Q_{II})^{1/2}}. \end{aligned} \tag{7.4.11}$$

Any of the expressions (7.4.10) or (7.4.11) show that r should be thought of as

$$r = \cos\theta, \tag{7.4.12}$$

where θ denotes the angle between V and W in variable-space $\mathscr{E}$ with covariance as inner product, or equivalently θ is the angle between P_{II} and Q_{II} in $\mathscr{N}_{II}$. A sample correlation coefficient between V and W will sometimes be denoted by $\operatorname{cor}(V, W)$.

From (7.4.12) it is clear that

$$-1 \leq r \leq 1. \tag{7.4.13}$$

Moreover $r = \pm 1$ if and only if

$$P_{II} = \pm\delta Q_{II} \tag{7.4.14}$$

for some $\delta > 0$. Since $P_{II} = \sum_1^n (X^{(i)} - \bar{X})N_i$ and $Q_{II} = \sum_1^n (Y^{(i)} - \bar{Y})N_i$, the condition (7.4.14) may be written

$$X^{(i)} - \bar{X} = \pm\delta(Y^{(i)} - \bar{Y}) \tag{7.4.15}$$

for $i = 1, 2, \ldots, n$.

Actual graphic plotting of a sample as p points in $\mathcal{N}$ or $\mathcal{N}_{II}$ is impractical since $n = 2$ or $n = 3$ rarely occurs. Such plotting is, however, a useful conceptual device.

7.5 COMPUTATION-ORIENTED THEORY

This section considers some standard computations based on single sample data matrices. The first concern will be to describe the computation of a sample mean vector and a sample covariance matrix in terms of the computing language introduced in Section 4.3. Later, the discussion will turn to deeper matters involving partially swept inner product matrices together with the addition and deletion of either variables or individuals.

Given an $n \times p$ data matrix $\mathbf{X}$, the inner product matrix of the raw sum inner product may be computed as indicated in (7.2.15) by the single matrix multiplication

$$\mathbf{Q} = \mathbf{X}'\mathbf{X}. \tag{7.5.1}$$

Similarly, the mean vector $\bar{\mathbf{X}}$ may be computed as indicated by (7.2.11), and thence $\mathbf{T}$ and $\mathbf{S}$ from $\mathbf{Q}$ and $\bar{\mathbf{X}}$ as indicated by (7.2.17). An alternative computing scheme uses the device of adding a column to $\mathbf{X}$ whose n elements are all unity, thus forming an augmented $n \times (p + 1)$ data matrix $\mathbf{X}_{(+)}$. $\mathbf{X}_{(+)}$ should be regarded as the data matrix for a $(p + 1)$-variate sample whose variables are $V_1, V_2, \ldots, V_p$, and V_0 whose value is unity for all individuals.

Extending (7.5.1) gives an augmented raw sum inner product matrix

$$\mathbf{Q}_{(+)} = \mathbf{X}'_{(+)}\mathbf{X}_{(+)} \tag{7.5.2}$$

of dimension $(p + 1) \times (p + 1)$. It is easily checked that the partition of $\mathbf{Q}_{(+)}$ into p and 1 rows and columns yields

$$\mathbf{Q}_{(+)} = \left[\begin{array}{ccc:c} & & & \sum_1^n X_1^{(i)} \\ & & & \sum_1^n X_2^{(i)} \\ & & & \cdot \\ & \mathbf{Q} & & \cdot \\ & & & \cdot \\ & & & \sum_1^n X_p^{(i)} \\ \hdashline \sum_1^n X_1^{(i)} & \sum_1^n X_2^{(i)} & \ldots\ \sum_1^n X_p^{(i)} & n \end{array}\right] \tag{7.5.3}$$

so that the last row or column of $\mathbf{Q}_{(+)}$ yields the sample sums of the p variables $V_1, V_2, \ldots, V_p$ while the last diagonal element records the sample size. Next, applying the sweep operator SWP$[p + 1]$ defined in Section 4.3.2 yields

$$\text{SWP}[p + 1]\mathbf{Q}_{(+)} = \left[\begin{array}{ccc:c} & & & \bar{X}_1 \\ & & & \bar{X}_2 \\ & & & \cdot \\ & \mathbf{T} & & \cdot \\ & & & \cdot \\ & & & \bar{X}_p \\ \hdashline \bar{X}_1 & \bar{X}_2 & \ldots\ \bar{X}_p & -1/n \end{array}\right], \tag{7.5.4}$$

where $\mathbf{T}$ denotes the sample corrected sum inner product matrix and $\bar{X}_j$ denotes the sample mean value of V_j for $j = 1, 2, \ldots, p$. Clearly, the computing sequence of finding $\mathbf{Q}_{(+)}$ and then SWP$[p + 1]\mathbf{Q}_{(+)}$ is an easily programmable description of a way to find sample sums, sample means, sample raw sum inner products, and sample corrected sum inner products. The final step of computing $\mathbf{S} = \mathbf{T}/(n - 1)$ is trivial. A word of caution is appropriate here. $\mathbf{T}$ in (7.5.4) is found by subtraction of $n\bar{\mathbf{X}}'\bar{\mathbf{X}}$ from $\mathbf{Q}$ in (7.5.3), and may be a small difference between large values, thus acquiring a large component of rounding error. In practice, therefore, it is usually preferable to compute $\bar{\mathbf{X}}$ directly and thence compute $\mathbf{T}$ directly from the first line of (7.2.16).

A device similar to that of creating $\mathbf{X}_{(+)}$ from $\mathbf{X}$ is to add columns to $\mathbf{X}$ corresponding to individuals in the sample. This device is computationally useful for adding and deleting individuals from a sample. Suppose that the data matrix $\mathbf{X}$ augmented by columns corresponding to individuals $a_{i_1}, a_{i_2}, \ldots, a_{i_r}$ is denoted by

$$\mathbf{X}^* = [\mathbf{X}, \mathbf{1}_{i_1}, \mathbf{1}_{i_2}, \ldots, \mathbf{1}_{i_r}], \tag{7.5.5}$$

where $\mathbf{1}_{i_s}$ denotes an $n \times 1$ vector with zero elements except for unity in position i_s for $s = 1, 2, \ldots, r$. Again, $\mathbf{X}^*$ should be regarded as a data matrix for a sample with additional artificial variables, where the artificial variable corresponding to an individual is the *indicator* variable which takes the value unity for that individual and zero for all other individuals. Next consider

$$\mathbf{Q}^* = \mathbf{X}^{*\prime}\mathbf{X}^* \tag{7.5.6}$$

whose rows and columns may be partitioned into p and r to produce

$$\mathbf{Q}^* = \begin{bmatrix} \mathbf{Q} & \mathbf{F}' \\ \mathbf{F} & \mathbf{I} \end{bmatrix}, \tag{7.5.7}$$

where $\mathbf{F}$ denotes the $r \times p$ data matrix of the individuals $a_{i_1}, a_{i_2}, \ldots, a_{i_r}$ on variables $V_1, V_2, \ldots, V_p$ and $\mathbf{I}$ denotes the $r \times r$ identity matrix.

To see how $\mathbf{Q}^*$ relates to adding and deleting individuals define

$$\dot{\mathbf{Q}}^* = \text{SWP}[p + 1, p + 2, \ldots, p + r]\mathbf{Q}^*, \tag{7.5.8}$$

where, of course,

$$\mathbf{Q}^* = \text{RSW}[p + 1, p + 2, \ldots, p + r]\dot{\mathbf{Q}}^*. \tag{7.5.9}$$

By directly carrying out the sweep operations indicated in (7.5.8), it is easily seen that (7.5.7) is modified into

$$\dot{\mathbf{Q}}^* = \begin{bmatrix} \dot{\mathbf{Q}} & \mathbf{F}' \\ \mathbf{F} & -\mathbf{I} \end{bmatrix}, \tag{7.5.10}$$

where $\dot{\mathbf{Q}}$ denotes the raw sum inner product matrix of the sample of size $n - r$ formed by deleting the individuals $a_{i_1}, a_{i_2}, \ldots, a_{i_r}$ from the given sample. Thus it follows that the operations (7.5.8) and (7.5.9) may be used to delete and add individuals, respectively, from a raw sum inner product matrix.

A somewhat deeper look at the computational problems of adding variables or of adding or deleting individuals is facilitated by the use of the assimilation operator defined in Section 4.3.3. With a given sample of size n, the addition of variables $V_{p+1}, V_{p+2}, \ldots, V_{p+r}$ to a basic set $V_1, V_2, \ldots, V_p$ means that the basic $p \times p$ raw sum inner product matrix $\mathbf{Q}$ acquires r additional rows and columns and becomes a $(p + r) \times (p + r)$ raw sum inner product matrix $\mathbf{Q}^*$. Suppose that earlier statistical analysis of the original p variables has produced SWP$[1, 2, \ldots, s]\mathbf{Q}$. Then the ASM$[p + 1, p + 2, \ldots, p + r;\ 1, 2, \ldots, s]$ operator is designed to assimilate the last r rows and columns of $\mathbf{Q}^*$ together with SWP$[1, 2, \ldots, s]\mathbf{Q}$ to produce SWP$[1, 2, \ldots, s]\mathbf{Q}^*$.

The ASM operator is also useful in generalizing the operations of passing between (7.5.7) and (7.5.10). The aim of the generalization is to pass back and forth between SWP$[1, 2, \ldots, s]\mathbf{Q}$ and SWP$[1, 2, \ldots, s]\dot{\mathbf{Q}}$. The partition of SWP$[1, 2, \ldots, s]\mathbf{Q}$ will be denoted as usual by

$$\text{SWP}[1, 2, \ldots, s]\mathbf{Q} = \begin{bmatrix} -\mathbf{Q}_{11}^{-1} & \mathbf{H}_{12} \\ \mathbf{H}_{21} & \mathbf{Q}_{22.1} \end{bmatrix}, \tag{7.5.11}$$

and corresponding notation

$$\text{SWP}[1, 2, \ldots, s]\dot{\mathbf{Q}} = \begin{bmatrix} -\dot{\mathbf{Q}}_{11}^{-1} & \dot{\mathbf{H}}_{12} \\ \dot{\mathbf{H}}_{21} & \dot{\mathbf{Q}}_{22.1} \end{bmatrix} \tag{7.5.12}$$

and

$$\mathbf{F} = [\mathbf{F}_1, \mathbf{F}_2] \tag{7.5.13}$$

will be used.

Theorem 7.5. *The operator* ASM$[p + 1,\ p + 2, \ldots, p + r;\ 1,\ 2, \ldots, s]$ *followed by* SWP$[p + 1, p + 2, \ldots, p + r]$ *carries*

$$\begin{bmatrix} -\mathbf{Q}_{11}^{-1} & \mathbf{H}_{12} & \mathbf{F}_1' \\ \mathbf{H}_{21} & \mathbf{Q}_{22.1} & \mathbf{F}_2' \\ \mathbf{F}_1 & \mathbf{F}_2 & \mathbf{I} \end{bmatrix} \tag{7.5.14}$$

into

$$\begin{bmatrix} -\dot{\mathbf{Q}}_{11}^{-1} & \dot{\mathbf{H}}_{12} & \dot{\mathbf{Q}}_{11}^{-1}\mathbf{F}_1' \\ \dot{\mathbf{H}}_{21} & \dot{\mathbf{Q}}_{22.1} & \mathbf{F}_2' - \dot{\mathbf{H}}_{21}\mathbf{F}_1' \\ \mathbf{F}_1\dot{\mathbf{Q}}_{11}^{-1} & \mathbf{F}_2 - \mathbf{F}_1\dot{\mathbf{H}}_{12} & -\mathbf{I} - \mathbf{F}_1\dot{\mathbf{Q}}_{11}^{-1}\mathbf{F}_1' \end{bmatrix}. \tag{7.5.15}$$

Similarly, the operator ASM$[p + 1, p + 2, \ldots, p + r;\ 1, 2, \ldots, s]$ *followed by* SWP$[p + 1, p + 2, \ldots, p + r]$ *carries*

$$\begin{bmatrix} -\dot{\mathbf{Q}}_{11}^{-1} & \dot{\mathbf{H}}_{12} & \mathbf{F}_1' \\ \dot{\mathbf{H}}_{21} & \dot{\mathbf{Q}}_{22.1} & \mathbf{F}_2' \\ \mathbf{F}_1 & \mathbf{F}_2 & -\mathbf{I} \end{bmatrix} \tag{7.5.16}$$

into

$$\begin{bmatrix} -\mathbf{Q}_{11}^{-1} & \mathbf{H}_{12} & -\mathbf{Q}_{11}^{-1}\mathbf{F}_1' \\ \mathbf{H}_{21} & \mathbf{Q}_{22.1} & -\mathbf{F}_2' + \mathbf{H}_{21}\mathbf{F}_1' \\ -\mathbf{F}_1\mathbf{Q}_{11}^{-1} & -\mathbf{F}_2 + \mathbf{F}_1\mathbf{H}_{12} & \mathbf{I} - \mathbf{F}_1\mathbf{Q}_{11}^{-1}\mathbf{F}_1' \end{bmatrix}. \tag{7.5.17}$$

The first part of Theorem 7.5 is proved by noting that the three steps of passing from (7.5.14) to (7.5.7) to (7.5.10) to (7.5.15) require the successive operations:

$$\begin{aligned} &\text{RSW}[1, 2, \ldots, s] \quad \text{on the upper left } p \times p \text{ part only}, \\ &\text{SWP}[p + 1, p + 2, \ldots, p + r], \\ &\text{SWP}[1, 2, \ldots, s]. \end{aligned} \tag{7.5.18}$$

The last two operations in (7.5.18) may be carried out in reverse order, since sweep operators commute, and after reversing this order the first two operations combine to form ASM$[p + 1, p + 2, \ldots, p + r; 1, 2, \ldots, s]$, by the final characterization of the ASM operator given in Section 4.3.3. The second part of the theorem is proved in a similar way, so the details are omitted.

The computing rules of Theorem 7.5 may be made to yield a set of mathematically elegant but computationally inefficient formulas concerning the addition and deletion of individuals. Carrying out the ASM operation as defined by (4.3.31) and the SWP operations as defined by (4.3.23) on (7.5.14) and comparing the result to (7.5.15) yields:

$$\dot{\mathbf{Q}}_{11}^{-1} = \mathbf{Q}_{11}^{-1} + (\mathbf{F}_1\mathbf{Q}_{11}^{-1})'(\mathbf{I} - \mathbf{F}_1\mathbf{Q}_{11}^{-1}\mathbf{F}_1')^{-1}(\mathbf{F}_1\mathbf{Q}_{11}^{-1}), \tag{7.5.19}$$

$$\dot{\mathbf{H}}_{12} = \mathbf{H}_{12} - (\mathbf{F}_1\mathbf{Q}_{11}^{-1})'(\mathbf{I} - \mathbf{F}_1\mathbf{Q}_{11}^{-1}\mathbf{F}_1')^{-1}(\mathbf{F}_2 - \mathbf{F}_1\mathbf{H}_{12}), \tag{7.5.20}$$

$$\dot{\mathbf{Q}}_{22.1} = \mathbf{Q}_{22.1} - (\mathbf{F}_2 - \mathbf{F}_1\mathbf{H}_{12})'(\mathbf{I} - \mathbf{F}_1\mathbf{Q}_{11}^{-1}\mathbf{F}_1')^{-1}(\mathbf{F}_2 - \mathbf{F}_1\mathbf{H}_{12}), \tag{7.5.21}$$

$$\mathbf{F}_1\dot{\mathbf{Q}}_{11}^{-1} = (\mathbf{I} - \mathbf{F}_1\mathbf{Q}_{11}^{-1}\mathbf{F}_1')^{-1}(\mathbf{F}_1\mathbf{Q}_{11}^{-1}), \tag{7.5.22}$$

$$\mathbf{F}_2 - \mathbf{F}_1\dot{\mathbf{H}}_{12} = (\mathbf{I} - \mathbf{F}_1\mathbf{Q}_{11}^{-1}\mathbf{F}_1')^{-1}(\mathbf{F}_2 - \mathbf{F}_1\mathbf{H}_{12}), \tag{7.5.23}$$

and

$$(\mathbf{I} + \mathbf{F}_1\dot{\mathbf{Q}}_{11}^{-1}\mathbf{F}_1') = (\mathbf{I} - \mathbf{F}_1\mathbf{Q}_{11}^{-1}\mathbf{F}_1')^{-1}. \tag{7.5.24}$$

Following a similar procedure with (7.5.16) and (7.5.17) yields

$$\mathbf{Q}_{11}^{-1} = \dot{\mathbf{Q}}_{11}^{-1} - (\mathbf{F}_1\dot{\mathbf{Q}}_{11}^{-1})'(\mathbf{I} + \mathbf{F}_1\dot{\mathbf{Q}}_{11}^{-1}\mathbf{F}_1')^{-1}(\mathbf{F}_1\dot{\mathbf{Q}}_{11}^{-1}), \tag{7.5.25}$$

$$\mathbf{H}_{12} = \dot{\mathbf{H}}_{12} + (\mathbf{F}_1\dot{\mathbf{Q}}_{11}^{-1})'(\mathbf{I} + \mathbf{F}_1\dot{\mathbf{Q}}_{11}^{-1}\mathbf{F}_1')^{-1}(\mathbf{F}_2 - \mathbf{F}_1\dot{\mathbf{H}}_{12}), \tag{7.5.26}$$

and

$$\mathbf{Q}_{22.1} = \dot{\mathbf{Q}}_{22.1} + (\mathbf{F}_2 - \mathbf{F}_1\dot{\mathbf{H}}_{12})'(\mathbf{I} + \mathbf{F}_1\dot{\mathbf{Q}}_{11}^{-1}\mathbf{F}_1')^{-1}(\mathbf{F}_2 - \mathbf{F}_1\dot{\mathbf{H}}_{12}). \tag{7.5.27}$$

The analogues of (7.5.22) and (7.5.23) are formed by substituting (7.5.20) into (7.5.22) and (7.5.23).

Theorem 7.5 and the subsequent formulas are likely to be of most interest and use when $\mathbf{Q}$ is replaced by $\mathbf{Q}_{(+)}$, which means that the artificial variable V_0 is introduced as a $(p+1)$st variable. The important task is now to pass back and forth between $\text{SWP}[p+1, 1, 2, \ldots, s]\mathbf{Q}_{(+)}$ and $\text{SWP}[p+1, 1, 2, \ldots, s]\dot{\mathbf{Q}}_{(+)}$. The parts of these matrices are worth noting. From $\text{SWP}[1, 2, \ldots, s]$ applied to (7.5.4) it follows that

$$\text{SWP}[p+1, 1, 2, \ldots, s]\mathbf{Q}_{(+)}$$

$$= \begin{bmatrix} -\mathbf{T}_{11}^{-1} & \mathbf{J}_{12} & \mathbf{T}_{11}^{-1}\bar{\mathbf{X}}_1' \\ \mathbf{J}_{21} & \mathbf{T}_{22.1} & \bar{\mathbf{X}}_2' - \mathbf{J}_{21}\bar{\mathbf{X}}_1' \\ \bar{\mathbf{X}}_1\mathbf{T}_{11}^{-1} & \bar{\mathbf{X}}_2 - \bar{\mathbf{X}}_1\mathbf{J}_{12} & -1/n - \bar{\mathbf{X}}_1\mathbf{T}_{11}^{-1}\bar{\mathbf{X}}_1' \end{bmatrix}, \qquad (7.5.28)$$

where

$$\mathbf{J}_{12} = \mathbf{J}_{21}' = \mathbf{T}_{11}^{-1}\mathbf{T}_{12} \qquad (7.5.29)$$

and

$$\mathbf{T}_{22.1} = \mathbf{T}_{22} - \mathbf{J}_{21}\mathbf{T}_{12}. \qquad (7.5.30)$$

Similar formulas could be written down immediately for $\dot{\mathbf{Q}}_{(+)}$ referring to the reduced sample with r fewer individuals.

7.6 PRINCIPAL COMPONENT ANALYSIS

One possible attitude to multivariate statistics might hold that variable-space and individual-space are fundamentally affine spaces and should be regarded as Euclidean spaces only for inner products based on observed data. A different attitude says that it is difficult to contemplate any space of variables without at least some indefinite hints of Euclidean structure present. For example, a plot like Fig. 7.2.1 presumes related scales of measurement in different directions, so as to yield a picture comprehensible to the eye. Likewise, the visual impact of such a picture depends on the initial angle between the coordinate axes v_1 and v_2.

The user of a principal component analysis adopts the second of these two attitudes. In fact, he must promote vague feelings about scales and angles among $v_1, v_2, \ldots, v_p$ into a precise inner product. This inner product is not determined wholly by the sample data, and will be called here a *reference inner product*, where the term will be used either for an inner product over variable-space or for its dual over individual-space. The *principal component analysis* of a given sample relative to a given reference inner product over variable-space consists of finding the eigenvalues and eigenvectors of the sample covariance inner product relative to the reference inner product. The eigenvalues found in this way will be called *sample principal components of total variance relative to the chosen reference inner product* or, more briefly, *principal components*. The corresponding eigenvectors which form a basis of variable-space $\mathscr{E}$ will be called *principal variables*. This terminology is slightly different from and more general than that of Hotelling (1933, 1936) who first introduced the concept of a principal component analysis.

The following discussion first explores the properties of principal components and then finishes with a brief description of their statistical interpretation.

In practice, the reference inner product has usually been chosen in one of two ways. In both of these ways, the directly observed set of variables $V_1, V_2, \ldots, V_p$ is regarded as an orthogonal set, so that only the reference norms of $V_1, V_2, \ldots, V_p$ remain to be chosen. The first method chooses these norms independently of the sample data to represent some vague opinion of what should be comparable scales. Under the second method, the reference norms of $V_1, V_2, \ldots, V_p$ have been taken to be their sample standard deviations, so that the standardized variables $u_i = \text{var}\,(V_i)^{-1/2}V_i$ have unit norms according to the reference inner product. The second method depends on the sample data through the choice of reference norms but not as regards the orthogonality of $\mathbf{V}$, while the first method is entirely free of the particular sample outcomes. Other choices of a reference inner product may be reasonable, but it is clear that arbitrary dependence on the sample data cannot be allowed, since this would permit the user to produce completely arbitrary principal components and variables. Perhaps a reasonable restriction is to require the initial data free selection of *some* basis of variable-space to be an orthogonal basis for the reference inner product, where this basis need not consist of the basic observable variables; the reference norms for the basis may then be chosen in terms of the data.

In computing principal components and variables the first step is generally to find the sample covariance matrix relative to a basis $\mathbf{U}$ orthonormal according to the reference inner product. Calling this covariance matrix $\mathbf{S}^*$, it is clear that the eigenvalues of $\mathbf{S}^*$ are the principal components and the corresponding eigenvectors of $\mathbf{S}^*$ are the coordinate vectors relative to $\mathbf{U}$ of the principal variables. Recall the discussion of Section 5.4 at this point.

If the principal components are denoted by $\lambda_1, \lambda_2, \ldots, \lambda_p$, then

$$\sum_{i=1}^{p} \lambda_i = \text{tr}\, \mathbf{S}^*, \qquad (7.6.1)$$

where $\text{tr}\,\mathbf{S}^*$ denotes the sum of the diagonal elements of $\mathbf{S}^*$. Formula (7.6.1) may be proved by showing that $\text{tr}\,\mathbf{S}^* = \text{tr}\,\mathbf{C}\mathbf{S}^*\mathbf{C}'$ for any orthogonal matrix $\mathbf{C}$. Since $\text{tr}\,\mathbf{S}^*$ is a sum of variances, formula (7.6.1) explains the often used term *principal components of total variance*. The λ_i are themselves sample variances of the principal variables. Another relationship similar to (7.6.1) is

$$\prod_{i=1}^{p} \lambda_i = \det \mathbf{S}^*. \qquad (7.6.2)$$

Wilks (1932) introduced the term *generalized variance* for the quantity (7.6.2) as a general overall measure of the variability of the set of variables **U**. Neither the total variance (7.6.1) nor the generalized variance (7.6.2) tell the whole story, however, and the generalized variance in particular suffers the disadvantage of being close to zero when any λ_i is close to zero even though the remaining λ_j may be large.

Under the second method of choosing the reference inner product, the orthonormal basis **U** may be defined by $U_i = \text{var}\,(V_i)^{-1/2}V_i$ for $i = 1, 2, \ldots, p$. In this case the matrix $\mathbf{S}^*$ becomes the *sample correlation matrix* **R** whose (i, j) element is the sample correlation coefficient between V_i and V_j and whose diagonal elements are unity. (See (7.4.10).) In this case

$$\sum_{i=1}^{p} \lambda_i = p \tag{7.6.3}$$

and

$$\prod_{i=1}^{p} \lambda_i = \det \mathbf{R}. \tag{7.6.4}$$

The quantity (7.6.4) is sometimes called a *scatter coefficient* (Frisch, 1929). Since the scatter coefficient is the product of p nonnegative real numbers with a given sum, and since such a product is maximum if and only if the numbers are all equal, it follows that

$$0 \le \det \mathbf{R} \le 1. \tag{7.6.5}$$

The idea here is that the scatter coefficient represents a general measure of the degree of correlation among $V_1, V_2, \ldots, V_p$ where the smaller the value, the more correlation is present. Certainly, $\det \mathbf{R} = 1$ if and only if $V_1, V_2, \ldots, V_p$ are uncorrelated. Note, however, that no single index can describe the whole complex of **R** very well. For example, $\det \mathbf{R} = 0$ if any of the λ_i is zero, so that $\det \mathbf{R}$ does not differentiate among the different possible dimensions of scatter as measured by the rank of **R**.

Serious examples of principal component analyses are given in Examples 8.4, 9.1, and 10.3. Figure 7.6.1 shows the same concentration ellipse as Fig. 7.3.1 with the sample points omitted. When V_1 and V_2 are taken to be orthonormal according to the reference inner product, the principal components λ_1 and λ_2 are simply the squared lengths of the major and minor semi-axes of the ellipse, and these axes determine also the dual basis **w** of the principal variable basis **W**.

Finally, consider what meaning a principal component analysis might have. The aim of a principal component analysis is to provide a special basis of uncorrelated variables which provide a maximal range of importance. Usually the eigenvalues are ordered according to $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p$ and a subset $\lambda_1, \lambda_2, \ldots, \lambda_r$ for $r < p$ is selected as describing most of the variability in the sample. The quantity

$$\sum_{i=1}^{r} \lambda_i \Big/ \sum_{i=1}^{p} \lambda_i \tag{7.6.6}$$

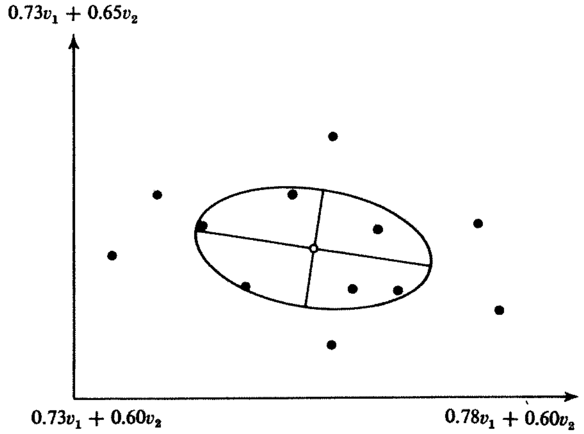

Fig. 7.6.1. The mean-centered concentration ellipse of Fig. 7.3.1 shown together with its major and minor axes which determine principal components of sample variance.

is commonly used to measure the fraction of total variance accounted for by the first r principal variables.

The prime intent of principal component analysis is therefore the attempt to simplify the study of multivariate samples by reducing their dimension in such a way as to lose as little information as possible. Usually it is hoped that the λ_i drop off very rapidly, so that an r of two or three may be selected and the sample represented in a greatly reduced individual-space of two or three dimensions.

There is a strong flavor of vagueness and arbitrariness about the technique of principal component analysis. The reference inner product and the measure (7.6.6) is somewhat arbitrary, and the means of choosing the reduced dimension r is not rationalized. Moreover, the nature of the importance of the first few principal variables is not defined. For example, it is mathematically possible that the last principal variable corresponding to the smallest λ_i should be the only one of use in predicting some separate but scientifically important variable. An empirical basis for the technique is thus seen to rest on whether the first few principal variables are those which are of value for predicting other variables. The data of Examples 8.4 and 9.1 tend to offer limited support for such an empirical basis. The evaluation of a particular sample principal component analysis is further complicated by the presence of sampling variation. Even if a population principal variable were closely aligned with some outside variable, the strength of the relation would be less apparent when the population principal was replaced by a sample analogue.

Principal variables are sometimes referred to as *factors* or *underlying factors*, the idea being that apart from minor disturbances all variables may be represented as linear functions of a few basic underlying variables. Accordingly, principal component analysis is a particular method of *factor analysis.* Attempts are sometimes made to regard the factors found in this way as hard well-defined variables, but such attempts deserve skeptical scrutiny. It may be that principal component analysis will some day be of use in locating hard underlying factors, such as genetically determined factors. At present, however, the uses are largely descriptive, explanatory, and empirical. Sampling theory and formal procedures for drawing inferences from samples to populations are in an underdeveloped and unsatisfactory state, both for principal component analysis and for methods of factor analysis generally. Lawley and Maxwell (1963) and Cattell (1965) are suggested as starting points for the reader wishing to pursue factor analysis further.

7.7 SEMI-DEFINITE SAMPLES

The terminology of Section 3.5 may be extended by defining a p-variate sample to be semi-definite if its sample covariance is a semi-definite inner product. More generally, a p-variate sample will be said to have rank f for $0 \leq f \leq p$ if its sample covariance has rank f.

> ***Theorem** 7.7.1. A p-variate sample of size n has rank f if and only if the n individuals all lie in a hyperplane of dimension f in $\mathscr{F}$ but do not lie in any hyperplane of dimension less than f.*

To assert that n individuals all lie in a hyperplane $u + \mathscr{U}$ in $\mathscr{F}$ is equivalent to asserting that the n individuals have identical observed values for all variables V in the subspace $\mathscr{V}$ in $\mathscr{E}$ dual to $\mathscr{U}$ in $\mathscr{F}$. Consequently the lowest-dimensional hyperplane containing all n individuals has dimension f if and only if the largest subspace of $\mathscr{E}$ on which all n individuals have identical values has dimension $p - f$. But the sample variance of a variable V is zero if and only if all n individuals have identical values on V, so that the subspace of such variables V has dimension $p - f$ if and only if the sample covariance has rank f, as required to prove the theorem.

It follows from the theory of Section 6.6 that the concentration of a semi-definite sample of rank f, being the dual of an inner product of rank f over $\mathscr{E}$, is a partial inner product over an f-dimensional subspace of $\mathscr{F}$. Since this subspace is the dual of the $(p - f)$-dimensional subspace of $\mathscr{E}$ over which the sample variance is zero, it is simply the hyperplane through the origin parallel to the hyperplane containing the sample. Thus, Theorem 7.7.2 which follows is proved.

> ***Theorem** 7.7.2. The concentration inner product of a semi-definite sample of rank f is a partial inner product defined over the f-dimensional subspace $\mathscr{U}$ of $\mathscr{F}$ which is parallel to the f-dimensional hyperplane $u + \mathscr{U}$ which contains the sample. The origin-centered ellipsoid of concentration lies in $\mathscr{U}$ and the mean-centered ellipsoid of concentration lies in $u + \mathscr{U}$.*

A set of n points in $\mathscr{F}$ must lie in some hyperplane of dimension at most $n - 1$. Consequently, if $n - 1 < p$, then the sample must be semi-definite of rank at most $n - 1 < p$. Usually, with observational data, it will turn out that the rank will achieve its maximal value, and, under this circumstance, a sample of size n is semi-definite of rank $n - 1$ if and only if $n - 1 < p$.

The theory of Section 7.3 holds just as well for samples of rank less than p. In particular, the shadow theory of Theorem 7.3 requires no change.

A p-variate sample of size n determines a linear transformation from $\mathscr{E}$ to $\mathscr{N}$ and also a linear transformation from $\mathscr{E}$ to $\mathscr{N}_{II}$ as discussed in Section 7.4. Conversely, the transformation from $\mathscr{E}$ to $\mathscr{N}$ determines the sample. Alternatively, the transformation from $\mathscr{E}$ to $\mathscr{N}_{II}$ together with the sample mean m determines the sample. From the foregoing it is clear that *the transformation from $\mathscr{E}$ to $\mathscr{N}_{II}$ has rank f if and only if the sample has rank f.* Moreover, two variables V and W in $\mathscr{E}$ transform into the same point in $\mathscr{N}_{II}$ if and only if $\operatorname{var}(V - W) = 0$.

Finally, principal component analysis has some special features in the case of a semi-definite sample. Suppose that the reference inner product over $\mathscr{E}$ asserts that $\mathbf{U}$ is an orthonormal basis. Then the principal component analysis simply requires that the eigenvalues and eigenvectors of the sample covariance matrix $\mathbf{S}$ relative to $\mathbf{U}$ shall be found. Alternatively, the matrix $\mathbf{T} = (n - 1)\mathbf{S}$ may be analyzed, since its eigenvalues are simply $(n - 1)$ times those of $\mathbf{S}$ and its eigenvectors are the same as those of $\mathbf{S}$. From (7.2.16)

$$\mathbf{T} = \mathbf{Z}'\mathbf{Z}, \tag{7.7.1}$$

where $\mathbf{Z}$ is the $n \times p$ matrix with rows $\mathbf{X}^{(i)} - \bar{\mathbf{X}}$ for $i = 1, 2, \ldots, n$, i.e., $\mathbf{Z}$ is the data matrix $\mathbf{X}$ with the mean vector $\bar{\mathbf{X}}$ subtracted from each row. From the theory of Section 6.5 it follows that the nonzero eigenvalues of $\mathbf{Z}'\mathbf{Z}$ are identical to the nonzero eigenvalues of $\mathbf{Z}\mathbf{Z}'$ and also that *the eigenvectors of $\mathbf{Z}\mathbf{Z}'$ are simply the sample values of the sample principal variables after the sample means are subtracted out.*

This result holds whether or not the sample is semi-definite, but its practical importance lies in the case where $n - 1 < p$ so that the sample must be semi-definite. In such a case the matrix $\mathbf{Z}\mathbf{Z}'$ is smaller than $\mathbf{Z}'\mathbf{Z}$ and so leads to a more manageable task of computing eigenvalues and eigenvectors. An illustration may be found in Example 10.3.

Semi-definite populations and, more specifically, semi-definite samples are less susceptible to analysis by contemporary multivariate methods than are samples where $n - 1 \geq p$, for reasons which will become apparent in later chapters. Thus, for example, while one might think that as much could be learned, in some rough scale of justice, from a 100-variate sample of size 30 as from a 30-variate sample of size 100, the statistical methodology for handling the former type of data is less well developed and on the whole unsatisfactory.

The method of principal component analysis may be an exception to the rule, but only as an empirical matter to be explored separately by fields of application.

7.8 EXERCISES

7.2.1 Suppose that a given sample $a_1, a_2, \ldots, a_n$ has mean m, and that this sample is translated into the sample $a_1^*, a_2^*, \ldots, a_n^*$ by the translation $a_i \rightarrow a_i^* = a_i + b$. Show that the mean of the second sample is given by $m^* = m + b$, but that the covariance inner product of the second example is the same as that of the first sample.

7.2.2 What is the condition on a sample which makes the sample covariance inner product function identically zero?

7.2.3 Show that an alternative to the two expressions (7.2.5) for cov (V, W) is given by

$$\text{cov}\,(V, W) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} (X^{(i)} - X^{(j)})(Y^{(i)} - Y^{(j)}).$$

What is the analogous vector version of (7.2.16)?

7.2.4 For what quantities do the elements of the sample mean vector $\bar{\mathbf{X}}$ and the sample covariance matrix $\mathbf{S}$ represent values?

7.3.1 Suppose that con (a_1, a_2) denotes the sample concentration of the sample individuals a_1 and a_2. Describe how to compute con (a_1, a_2) from the sample data $\mathbf{X}$. Compute con (a_1, a_2) for the data plotted in Fig. 7.3.1.

7.3.2 Make a graphic estimate of $m(V_1 + 2V_2)$ and var $(V_1 + 2V_2)$ from the concentration ellipse plotted in Fig. 7.3.1. Check the estimates by computation.

7.3.3 What is the distance from the origin to the sample mean in Fig. 7.3.1 according to the concentration inner product? For what variable V does this distance equal $m(V)/\text{var}\,(V)^{1/2}$?

7.3.4 Show that there exists a $(p - 1)$-dimensional subspace $\mathscr{V}_i$ of variable space $\mathscr{E}$ such that the sample values of individual a_i coincide with those of the sample mean individual. Show that the sample mean values and the sample mean corrected sum inner product both restricted to variables in $\mathscr{V}_i$ are not changed if a_i is removed from the sample. Show that the one-dimensional orthogonal complement $\mathscr{U}_i$ of $\mathscr{V}_i$ according to the sample covariance remains the orthogonal complement according to the sample covariance based on the sample with a_i removed. Thus, if U_i denotes a variable in $\mathscr{U}_i$, the sample mean and sample variance of U_i are altered by removing a_i but the other aspects of the first and second sample moments described above do not change. Use the foregoing theory to describe how the sample mean-centered concentration ellipsoid changes when a_i is removed from the sample.

7.4.1 Show that the subspace $\mathscr{N}_{II}$ defined in Section 7.4 consists of those vectors $\sum_{j=1}^{n} z_j N_j$ such that $\sum_{j=1}^{n} z_j = 0$.

7.4.2 Show that the decomposition $P = P_I + P_{II}$ defined by (7.4.4) and (7.4.5) is in fact the desired orthogonal decomposition.

7.4.3 Suppose that the range subspaces in $\mathscr{N}$ of the linear transformations (7.4.3) and (7.4.9) are denoted by $\mathscr{M}$ and $\mathscr{M}_{II}$. What is the maximal dimension of $\mathscr{M}$? Of $\mathscr{M}_{II}$?

7.5.1 Suppose that the $(p + n) \times (p + n)$ matrix

$$\begin{bmatrix} \mathbf{0} & \mathbf{X}' \\ \mathbf{X} & \mathbf{I} \end{bmatrix}$$

is formed from an $n \times p$ sample data matrix $\mathbf{X}$, where $\mathbf{0}$ is a $p \times p$ matrix of zeros and $\mathbf{I}$ is the $n \times n$ identity matrix. What is the result of applying SWP$[p + 1, p + 2, \ldots, p + n]$ to this matrix? Are the computations performed in this way inefficient?

7.5.2 Write out the special application of formulas (7.5.19) through (7.5.27) when $\mathbf{Q}$ is replaced by $\mathbf{Q}_{(+)}$ and the role of indices $1, 2, \ldots, s$ is played by $1, 2, \ldots, s, p + 1$. Note that the latter set of indices partitions naturally into $1, 2, \ldots, s$ and $p + 1$ and that the associated formulas should be further partitioned accordingly. Use the notation of (7.5.28).

7.5.3 Define the augmented data matrix $\mathbf{X}^{(+)}$ to consist of $\mathbf{X}$ with a *first* column of ones adjoined. Suppose that $\mathbf{X}^{(+)\prime} = \mathbf{D}^{(+)}\mathbf{G}$ defines the triangularization of $\mathbf{X}^{(+)\prime}$ as defined in Exercise 4.3.3. Show how to find $\bar{\mathbf{X}}$ and $\mathbf{T}$ from $\mathbf{D}^{(+)}$ and $\mathbf{G}$.

7.6.1 Prove formulas (7.6.1) and (7.6.3).

7.6.2 Show that the product of p positive numbers with a given sum is maximum when and only when the numbers are all equal. Deduce that the scatter coefficient det $\mathbf{R}$ is unity if and only if $\mathbf{R}$ is the identity matrix.

7.6.3 Consider a principal component analysis of the correlation matrix of the pair of variables V_1 and V_2. Show that the principal variables are

$$\frac{1}{\sqrt{\text{var}\,(V_1)}}\, V_1 \pm \frac{1}{\sqrt{\text{var}\,(V_2)}}\, V_2.$$

What are the corresponding principal components?

7.7.1 Show that the following three assertions are equivalent:

i) $V - W$ has an observation vector lying along $\mathscr{N}_I$ as defined in Section 7.4,

ii) the observations on V and W differ by a constant for all individuals in the sample, and

iii) var $(V - W) = 0$.

7.7.2 Construct an example of a 4-variate sample of size 3 which has rank 1. Describe simply the set of points lying on the ellipsoid of concentration of this sample.

7.7.3 Suppose that the mapping (7.4.3) from $\mathscr{E}$ to $\mathscr{N}$ has the range space $\mathscr{M}$ of dimension g. Show that the rank of the sample is either $g - 1$ or g depending on whether or not $\mathscr{M}$ contains $\mathscr{N}_I$ as a subspace.

7.7.4 Consider the reduced data matrix $\mathbf{Z}$ with sample means removed. Show that the sample has rank f if and only if $\mathbf{Z}$ has rank f.

7.7.5 Show that $\mathbf{ZZ}'$ can be computed directly from $\mathbf{XX}'$ and describe the required computations.

CHAPTER 8

ONE SAMPLE OF INDIVIDUALS: MULTIPLE REGRESSION AND CORRELATION ANALYSIS

8.1 INTRODUCTION

This chapter is concerned with the prediction of a value for a specified variable given the values of a different set of variables on the same individual. A related concern is with the nature of the covariation which makes such prediction possible. By convention, the variable to be predicted will be denoted by V_p and the variables used for prediction will be denoted by $V_1, V_2, \ldots, V_{p-1}$. In this context V_p will be called a *predictand* or *dependent variable* while $V_1, V_2, \ldots, V_{p-1}$ will be called *predictors* or *independent variables.* It is assumed that an individual a for which a prediction is desired is a member of the same population as that represented by a given sample $a_1, a_2, \ldots, a_n$ of n individuals observed on *all* of the variables $V_1, V_2, \ldots, V_p$. The prediction scheme is to be based on the given sample.

The particular approach to prediction taken here is quite simple. A single unknown value is to be predicted by a single predicted value rather than, for example, by a probability distribution over the possible unknown values. The single predicted value is to be based on a *linear predictor* $w_1V_1 + w_2V_2 + \cdots + w_{p-1}V_{p-1}$ chosen through the principle of *least squares.* The resulting analysis of the given sample will be called *multiple regression analysis* of V_p on $V_1, V_2, \ldots, V_{p-1}$.

The restriction to linearity is not in itself very important, for the variables called $V_1, V_2, \ldots, V_{p-1}$ may be arbitrary functions of any set of directly observable variables. A difficulty may arise, however, because the smaller the sample size n the smaller is the number of predictors which can be used with a given effectiveness. This issue can only be discussed inconclusively, for the art of guessing scientific laws is not governed by well-established rules. In this chapter the choice of $V_1, V_2, \ldots, V_{p-1}$ will be taken as given. Even so, there remain questions concerning which of a possible set $V_1, V_2, \ldots, V_{p-1}$ might

better be left out of a linear predictor, and some discussion of this fits naturally into Section 8.3.

The term *multiple regression* is largely a historical accident. Galton (1886) first used the word *regression* in connection with predicting the mature height of children from the heights of their parents. Galton corrected for the sex difference by multiplying all female heights by 1.08, and he used a single predictor variable taken to be the mean of the father's height and corrected mother's height. After some consideration of data it becomes apparent that the heights of children of parents whose height exceeds average by x inches will themselves, on the average, exceed average by less than x inches. In other words, the children *regress* in an average sense back to the mean. By a gradual metamorphosis, the term linear regression analysis came to mean the least squares prediction scheme when $p = 2$, and thence the term multiple regression came to mean the general case with a multiple battery of variables $V_1, V_2, \ldots, V_{p-1}$ available as predictors.

The history of the method, as opposed to that of its common statistical name, is quite different. According to Gauss (1809), he first used the method in 1795 in a different context and under the name *method of least squares.* The early history of the method of least squares is somewhat confused because Gauss did not publish his claim until 1809 and meanwhile Legendre (1806) had independently described the method. According to Eisenhart (1963), the method arose as a natural extension of the principle of averaging the results of several observations of the same quantity to reduce measurement error. It has been widely used in astronomy and the physical sciences since the time of Gauss. It is interesting that the basic computational ideas of Section 4.3 may be traced back to Gauss (1811) who derived them in connection with least squares analysis and illustrated them with the data which he used to identify the orbit of the asteroid Pallas from observations over the period 1803–1809. The Pallas data are used in Example 8.3.

On the history of the correlation coefficient Pearson (1896) wrote:

> The fundamental theorems of correlation were for the first time and almost exhaustively discussed by Bravais ("Analyse mathématique sur les probabilités des erreurs de situation d'un point," *Memoires par divers Savans*, T.IX., Paris, 1846, pp. 255–332) nearly half a century ago. He deals completely with the correlation of two and three variables. Forty years later Mr. J. D. Hamilton Dickson (*Proc. Roy. Soc.*, 1886, p. 63) dealt with a special problem proposed to him by Mr. Galton, and reached on a somewhat narrow basis* (*The coefficient of correlation was assumed to be the same for the arrays of all types, a result which really flows from the normal law of frequency.) some of Bravais' results for correlation of two variables. Mr. Galton at the same time introduced an improved notation which may be summed up in the "Galton function" or coefficient

> of correlation. This indeed appears in Bravais' work, but a single symbol is not used for it. It will be found of great value in the present discussion. In 1892 Professor Edgeworth, also unconscious of Bravais' memoir, dealt in a paper on "Correlated Averages" with correlation for three variables (*Phil. Mag.* **34,** 1892, pp. 194–204). He obtained results identical with Bravais', although expressed in terms of Galton's functions. He indicates also how the method may be extended to higher degrees of correlation. He starts by assuming a general form for the frequency of any complex of n organs each of given size. The form has been deduced on more or less legitimate assumptions by various writers. Several other authors, notably Schols, De Forest, and Czuber, have dealt with the same topic, although little of first-class importance has been added to the researches of Bravais. To Mr. Galton alone is due the idea of applying these results—usually spoken of as "the laws of error in the position of a point in space"—to the problem of correlation in the theory of evolution.

Karl Pearson had much to do with the popularity of the idea among statistical data analysts. See Walker (1931), Seal (1967), and Pearson (1967) for more historical detail.

This introduction concludes with some remarks on the concept of *cause.* The ability to predict one variable from another, which accompanies nontrivial correlation between the variables, is sometimes interpreted by saying that the predictor is having a causal effect on the predictand. It is clear, however, that causal effects should be attributed only with great caution. For example, height and weight will show positive correlation in many human samples. This indicates that either variable can help to predict the other, but it does not indicate that an increase in height causes an increase in weight, or vice versa. It would be more natural to interpret a correlation between height and weight as the result of a common causal factor.

The notion of cause appears to require belief in some mechanism whereby the causal factor is *acting* while the influenced factor is *reacting*. Thus it is a plausible hypothesis that a higher incidence of smoking causes a higher incidence of morbidity of various kinds, and observed correlations do provide evidence for this hypothesis. Such evidence may be challenged on the grounds that the influenced factor is reacting to other causal factors whose variation in the observed sample is not controlled in reaction to that of the alleged causal factor. Such counterarguments may sometimes be finessed in part by the well-known techniques of experimental design in the sense of Fisher (1966), i.e., by the collection of the right sort of data. Sometimes controlled experiments are possible, and sometimes not. The examples in Chapters 8, 9, and 10 are of the latter kind, while in Chapter 11 there are examples of the former kind.

In general, however, the notion of cause is relative and vague, with any cause being partly or wholly replaceable by something more fundamental or more controversial. Issues of causation as distinct from prediction are scarcely mentioned in the sequel.

8.2 BASIC DESCRIPTION

Suppose that $X_1, X_2, \ldots, X_p$ denote the values of $V_1, V_2, \ldots, V_p$ for an individual a, where $X_1, X_2, \ldots, X_{p-1}$ are known and X_p is unknown. The multiple regression analysis of V_p on $V_1, V_2, \ldots, V_{p-1}$ provides a predicted value $\hat{X}_p$ for the unknown value X_p of the form

$$\hat{X}_p = w + w_1X_1 + w_2X_2 + \cdots + w_{p-1}X_{p-1}. \qquad (8.2.1)$$

The coefficients $w, w_1, w_2, \ldots, w_{p-1}$ in (8.2.1) are determined from the data on a given p-variate sample of size n to minimize the sum of squares of the n prediction errors resulting from the application of (8.2.1) to the n sample individuals. In symbols, the criterion to be minimized is

$$\sum_{i=1}^{n}(X_p^{(i)} - \hat{X}_p^{(i)})^2, \qquad (8.2.2)$$

where

$$\hat{X}_p^{(i)} = w + w_1X_1^{(i)} + w_2X_2^{(i)} + \cdots + w_{p-1}X_{p-1}^{(i)}, \qquad (8.2.3)$$

and $X_j^{(i)}$ for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$ denotes as in Chapter 7 the (i, j) element of the $n \times p$ data matrix $\mathbf{X}$. The coefficients in (8.2.1), called *regression coefficients*, are chosen according to the *principle of least squares.* When the least squares regression coefficients are used to define $\hat{X}_p^{(i)}$ in (8.2.3) the differences $X_p^{(i)} - \hat{X}_p^{(i)}$ for $i = 1, 2, \ldots, n$ are referred to as *residuals* and the minimized value of the criterion (8.2.2), namely the sum of squares of the residuals, is commonly called the *residual sum of squares.*

In the language of variable-space, the prediction scheme defined above provides the *augmented best linear predictor*

$$\hat{V}_p = wV_0 + w_1V_1 + \cdots + w_{p-1}V_{p-1} \qquad (8.2.4)$$

based on the sample data $\mathbf{X}$, where V_0 refers to the artificial variable whose value is always unity. Suppose that $\mathbf{X}_{(+)}$ denotes the augmented data matrix $\mathbf{X}$ with a column of ones added and that $\mathbf{Q}_{(+)} = \mathbf{X}'_{(+)}\mathbf{X}_{(+)}$ denotes the corresponding augmented raw sum inner product matrix, as in (7.5.2). Then the criterion (8.2.2) may be written

$$(\mathbf{X}_{(+)}\mathbf{d}'_{(+)})'(\mathbf{X}_{(+)}\mathbf{d}'_{(+)}) = \mathbf{d}_{(+)}\mathbf{Q}_{(+)}\mathbf{d}'_{(+)}, \qquad (8.2.5)$$

where $\mathbf{d}_{(+)}$ is the $1 \times (p + 1)$ vector of coefficients

$$\mathbf{d}_{(+)} = [-w_1, -w_2, \ldots, -w_{p-1}, 1, -w]. \qquad (8.2.6)$$

Thus the least squares criterion is a squared length according to the sample raw sum inner product over the augmented variable-space spanned by $V_0, V_1,$

$V_2, \ldots, V_p$. Furthermore, minimizing this criterion is seen to be equivalent to choosing that variable in the subspace spanned by $V_0, V_1, \ldots, V_{p-1}$ which lies at minimum distance from V_p according to the raw sum inner product over the augmented variable-space. In other words, the *augmented best linear predictor $\hat{V}_p$ defined in* (8.2.4) *is the orthogonal projection of V_p into the subspace spanned by $V_0, V_1, \ldots, V_{p-1}$, and the residual $V_p - \hat{V}_p$ is the component of V_p orthogonal to each of $V_0, V_1, V_2, \ldots, V_{p-1}$, all in terms of the augmented variable-space and an associated sample raw sum inner product.*

The standard computational device for finding such an orthogonal projection is related to the process of successive orthogonalization as described in Chapter 4. Starting from the appropriate inner product matrix $\mathbf{Q}_{(+)}$, the desired computations are provided by SWP$[p+1, 1, 2, \ldots, p-1]\mathbf{Q}_{(+)}$. The off-diagonal elements of row p in the resulting $(p+1) \times (p+1)$ matrix are the coefficients of $V_1, V_2, \ldots, V_{p-1}, V_0$ in the orthogonal projection of V_p into the subspace spanned by $V_1, V_2, \ldots, V_{p-1}, V_0$, i.e., they are $w_1, w_2, \ldots, w_{p-1}, w$. The (p, p) diagonal element is the square of the raw sum norm of $V_p - \hat{V}_p$, i.e., it is the residual sum of squares.

The foregoing discussion defines the multiple regression analysis of V_p on $V_1, V_2, \ldots, V_{p-1}$, but with the disadvantage of being given largely in terms of the raw sum inner product, while statisticians are more accustomed to looking at sample means and corrected sum or covariance inner products. Consequently, the discussion will now be translated into the latter terms. The bridge is rather easy, since the analysis is provided by

$$\text{SWP}[p+1, 1, 2, \ldots, p-1]\mathbf{Q}_{(+)} = \text{SWP}[1, 2, \ldots, p-1]\text{SWP}[p+1]\mathbf{Q}_{(+)},$$

and SWP$[p+1]\mathbf{Q}_{(+)}$ is expressed in (7.5.4) in the desired terms.

For present purposes it is convenient to partition the rows and columns of the right side of (7.5.4) into $p-1$, 1, and 1 and to set

$$\text{SWP}[p+1]\mathbf{Q}_{(+)} = \begin{bmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} & \bar{\mathbf{X}}_1' \\ \mathbf{T}_{21} & t_{pp} & \bar{X}_p \\ \bar{\mathbf{X}}_1 & \bar{X}_p & -1/n \end{bmatrix} \tag{8.2.7}$$

in an obvious notation where, for example, t_{pp} denotes the (p, p) diagonal element of $\mathbf{T}$, and $\bar{\mathbf{X}}_1$ denotes the $1 \times (p-1)$ vector $[\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_{p-1}]$ of sample means. Performing the SWP$[1, 2, \ldots, p-1]$ operation on the right side of (8.2.7) yields

$$\text{SWP}[p+1, 1, 2, \ldots, p-1]\mathbf{Q}_{(+)}$$
$$= \begin{bmatrix} -\mathbf{T}_{11}^{-1} & \mathbf{T}_{11}^{-1}\mathbf{T}_{12} & \mathbf{T}_{11}^{-1}\bar{\mathbf{X}}_1' \\ \mathbf{T}_{21}\mathbf{T}_{11}^{-1} & t_{pp} - \mathbf{T}_{21}\mathbf{T}_{11}^{-1}\mathbf{T}_{12} & \bar{X}_p - \mathbf{T}_{21}\mathbf{T}_{11}^{-1}\bar{\mathbf{X}}_1' \\ \bar{\mathbf{X}}_1\mathbf{T}_{11}^{-1} & \bar{X}_p - \bar{\mathbf{X}}_1\mathbf{T}_{11}^{-1}\mathbf{T}_{12} & -1/n - \bar{\mathbf{X}}_1\mathbf{T}_{11}^{-1}\bar{\mathbf{X}}_1' \end{bmatrix}. \tag{8.2.8}$$

From the previous interpretation of row p of SWP$[p+1, 1, 2, \ldots, p-1]\mathbf{Q}_{(+)}$ it follows that

$$(w_1, w_2, \ldots, w_{p-1}) = \mathbf{T}_{21}\mathbf{T}_{11}^{-1} \tag{8.2.9}$$

and

$$w = \bar{X}_p - \mathbf{T}_{21}\mathbf{T}_{11}^{-1}\bar{\mathbf{X}}_1' = \bar{X}_p - w_1\bar{X}_1 - w_2\bar{X}_2 - \cdots - w_{p-1}\bar{X}_{p-1}, \tag{8.2.10}$$

while the residual sum of squares is given by

$$t_{pp} - \mathbf{T}_{21}\mathbf{T}_{11}^{-1}\mathbf{T}_{12}. \tag{8.2.11}$$

These results lead to the following alternatives to the first two paragraphs of Section 8.2.

The predicted value (8.2.1) may be written as

$$\hat{X}_p = \bar{X}_p + w_1(X_1 - \bar{X}_1) + w_2(X_2 - \bar{X}_2) + \cdots + w_{p-1}(X_{p-1} - \bar{X}_{p-1}), \tag{8.2.12}$$

where $w_1, w_2, \ldots, w_{p-1}$ are chosen to minimize the criterion

$$\sum_{i=1}^{n} [(X_p^{(i)} - \bar{X}_p) - w_1(X_1^{(i)} - \bar{X}_1)$$
$$- w_2(X_1^{(i)} - \bar{X}_1) - \cdots - w_{p-1}(X_{p-1}^{(i)} - \bar{X}_{p-1})]^2. \tag{8.2.13}$$

The criterion (8.2.13) may also be written

$$\mathbf{dTd}', \tag{8.2.14}$$

where

$$\mathbf{d} = [-w_1, -w_2, \ldots, -w_{p-1}, 1] \tag{8.2.15}$$

and $\mathbf{T}$ denotes the sample corrected sum inner product matrix for the variables $V_1, V_2, \ldots, V_p$. The variable

$$\dot{V}_p = w_1V_1 + w_2V_2 + \cdots + w_{p-1}V_{p-1} \tag{8.2.16}$$

will be called the *best linear predictor* for V_p in terms of $V_1, V_2, \ldots, V_{p-1}$. Note that $\dot{V}_p$ lies in the ordinary p-dimensional variable-space $\mathscr{E}$ in contrast to the augmented best linear predictor $\hat{V}_p$ defined in (8.2.4) which lies in augmented variable-space.

In order to use the best linear predictor for actual prediction it is necessary to know also the vector of sample means, i.e., to know $\hat{V}_p$. On the other hand, $\dot{V}_p$ has an advantage over $\hat{V}_p$ in that it belongs to the familiar variable-space $\mathscr{E}$ on which the notion of covariance is relevant and meaningful. *The best linear predictor $\dot{V}_p$ should be regarded as an orthogonal projection, for from* (8.2.9) *$\dot{V}_p$ and $V_p - \dot{V}_p$ are the components of V_p along and orthogonal to the subspace of $\mathscr{E}$ spanned by $V_1, V_2, \ldots, V_{p-1}$ where the inner product assigned to $\mathscr{E}$ is the* *sample corrected sum inner product. The above orthogonal components $\dot{V}_p$ and $V_p - \dot{V}_p$ are the same whether the sample corrected sum inner product or the sample covariance is used, since changes of scale of an inner product have no effect on an orthogonal decomposition.*

The decomposition (8.2.11) of t_{pp} into $\mathbf{T}_{21}\mathbf{T}_{11}^{-1}\mathbf{T}_{12} + (t_{pp} - \mathbf{T}_{21}\mathbf{T}_{11}^{-1}\mathbf{T}_{12})$ is sometimes described in *analysis of variance* terminology (cf. Scheffé, 1959) as the decomposition of the *total sum of squares about the grand mean* into the *fitted or explained sum of squares* plus the residual sum of squares.

In terms of familiar statistical quantities the computations of multiple regression analysis may be described as: (i) finding the sample mean vector $\bar{\mathbf{X}}$, (ii) finding the sample covariance matrix $\mathbf{S}$, and (iii) finding the regression coefficients $\mathbf{S}_{11}^{-1}\mathbf{S}_{12}$ and the residual variance from SWP$[1, 2, \ldots, p-1]\mathbf{S}$. Still, the original description in terms of finding $\mathbf{Q}_{(+)} = \mathbf{X}'_{(+)}\mathbf{X}_{(+)}$ and SWP$[p+1, 1, 2, \ldots, p]\mathbf{Q}_{(+)}$ is computationally more natural.

While both (8.2.1) and (8.2.12) produce identical predicted values, a modified scheme yielding different results is occasionally appropriate. There may sometimes be theoretical reasons for omitting the constant term w from (8.2.1), i.e., for thinking that a predictor of the form

$$X_p^* = w_1^* X_1 + w_2^* X_2 + \cdots + w_{p-1}^* X_{p-1} \tag{8.2.17}$$

may actually improve on the version (8.2.1). In this case the least squares criterion becomes

$$\sum_{i=1}^{n} (X_p^{(i)} - w_1^* X_1^{(i)} - w_2^* X_2^{(i)} - \cdots - w_{p-1}^* X_{p-1}^{(i)})^2 = \mathbf{d}^*\mathbf{Q}\mathbf{d}^{*\prime}, \tag{8.2.18}$$

where

$$\mathbf{d}^* = [-w_1^*, -w_2^*, \ldots, -w_{p-1}^*, 1] \tag{8.2.19}$$

and $\mathbf{Q}$ is the sample raw sum inner product matrix. The required coefficients together with the residual sum of squares here are given by the last row of SWP$[1, 2, \ldots, p-1]\mathbf{Q}$. This produces the *reduced best linear predictor*

$$V_p^* = V_p - w_1^* V_1 - w_2^* V_2 - \cdots - w_{p-1}^* V_{p-1}, \tag{8.2.20}$$

whose interpretation as an orthogonal projection analogous to $\hat{V}_p$ and $\dot{V}_p$ is left to the reader to describe.

Example 8.1. The following data were collected by the author in a kitchen experiment with very rough measuring equipment. The length L in cm, the width W in cm, and the volume V in cc were measured on a dozen grade A large eggs. From the directly observed variables, three transformed variables $V_1 = \log_{10} L$, $V_2 = \log_{10} W$, and $V_3 = \log_{10} (6/\pi)V$ were selected for analysis.

The 12 × 3 data matrix with 12 eggs as individuals and V_1, V_2, V_3 as variables is:

V_1	V_2	V_3
0.7659	0.6360	2.031
0.7353	0.6198	1.982
0.7416	0.6280	1.995
0.7600	0.6280	2.019
0.7861	0.6239	2.031
0.7539	0.6156	1.956
0.7747	0.6156	2.007
0.7718	0.6239	1.995
0.7889	0.6114	1.995
0.7659	0.6072	1.995
0.7689	0.6156	1.995
0.7478	0.6239	2.007

This first example is kept simple so that the reader may try to reproduce the analysis on a desk calculator. The example is not intended to be representative of statistical practice. Note that the sample values are quite discrete, belying the first impression of a glance.

The use of logarithms in defining V_1, V_2, V_3, and the factor $6/\pi$ in the expression for V_3 were suggested by the formula $V = (\pi/6)LW^2$ for the volume of an ellipsoid with two principal axes of length W (i.e., a circular cross-section) and one principal axis of length L. Thus, if the eggs were precisely ellipsoids with circular cross-section, and if the measurements had been made precisely without error, then $V_1 + 2V_2$ would be a perfect predictor for V_3. In the following computations, the predictor $V_1 + 2V_2$ is compared with the two least squares best linear predictors $\hat{V}_3 = wV_0 + w_1V_1 + w_2V_2$ and $V_3^* = w_1^*V_1 + w_2^*V_2$.

The computations begin by finding

$$\mathbf{Q}_{(+)} = \begin{bmatrix} 6.9964 & 5.6861 & 18.3292 & 9.1608 \\ 5.6861 & 4.6246 & 14.9038 & 7.4489 \\ 18.3292 & 14.9038 & 48.0368 & 24.0080 \\ 9.1608 & 7.4489 & 24.0080 & 12.0000 \end{bmatrix}.$$

The computer then applied the operators SWP[1], SWP[2], SWP[3], SWP[4], RSW[3], RSW[2], RSW[1], and RSW[4] and printed out the resulting 8 matrices:

$$\text{SWP}[1]\mathbf{Q}_{(+)} = \begin{bmatrix} -0.1429 & 0.8127 & 2.6198 & 1.3094 \\ 0.8127 & 0.003329 & 0.007217 & 0.003708 \\ 2.6198 & 0.007217 & 0.01756 & 0.008374 \\ 1.3094 & 0.003708 & 0.008374 & 0.005174 \end{bmatrix},$$

$$\mathbf{SWP}[1,2]\mathbf{Q}_{(+)} = \begin{bmatrix} -198.5754 & 244.1576 & 0.8577 & 0.4040 \\ 244.1576 & -300.4193 & 2.1682 & 1.1139 \\ 0.8577 & 2.1682 & 0.001910 & 0.0003347 \\ 0.4040 & 1.1139 & 0.0003347 & 0.001044 \end{bmatrix},$$

$$\mathbf{SWP}[1,2,3]\mathbf{Q}_{(+)} = \begin{bmatrix} -583.3215 & -728.5001 & 448.5989 & 0.2538 \\ -728.5001 & -2759.3479 & 1134.0810 & 0.7342 \\ 448.5989 & 1134.0810 & -523.0488 & 0.1751 \\ 0.2538 & 0.7342 & 0.1751 & 0.0009849 \end{bmatrix},$$

$$\begin{aligned}\mathbf{SWP}[1,2,3,4]\mathbf{Q}_{(+)} &= -\,\mathbf{Q}_{(+)}^{-1} \\ &= \begin{bmatrix} -648.7275 & -917.6850 & 403.4707 & 257.6735 \\ -917.6850 & -3306.5594 & 1003.5490 & 745.3128 \\ 403.4707 & 1003.5490 & -554.1859 & 177.7871 \\ 257.6735 & 745.3128 & 177.7871 & -1015.1307 \end{bmatrix},\end{aligned}$$

$$\mathbf{SWP}[1,2,4]\mathbf{Q}_{(+)} = \begin{bmatrix} -354.9839 & -187.0591 & 0.7280 & 387.1100 \\ -187.0591 & -1489.2800 & 1.8109 & 1067.2590 \\ 0.7280 & 1.8109 & 0.001803 & 0.3208 \\ 387.1100 & 1067.2590 & 0.3208 & -958.0952 \end{bmatrix},$$

$$\mathbf{SWP}[1,4]\mathbf{Q}_{(+)} = \begin{bmatrix} -331.4886 & -0.1256 & 0.5006 & 253.0584 \\ -0.1256 & 0.0006713 & 0.001216 & 0.7166 \\ 0.5006 & 0.001216 & 0.004005 & 1.6185 \\ 253.0584 & 0.7166 & 1.6185 & -193.2681 \end{bmatrix},$$

$$\mathbf{SWP}[4]\mathbf{Q}_{(+)} = \begin{bmatrix} 0.003017 & -0.0003790 & 0.001510 & 0.7634 \\ -0.0003790 & 0.0007189 & 0.001026 & 0.6207 \\ 0.001510 & 0.001026 & 0.004761 & 2.0007 \\ 0.7634 & 0.6207 & 2.0007 & -0.08333 \end{bmatrix},$$

$$\mathbf{Q}_{(+)} = \begin{bmatrix} 6.9909 & 5.6816 & 18.3148 & 9.1536 \\ 5.6816 & 4.6209 & 14.8921 & 7.4431 \\ 18.3148 & 14.8921 & 47.9990 & 23.9891 \\ 9.1536 & 7.4431 & 23.9891 & 11.9906 \end{bmatrix}.$$

All of this output or even all of this computation is not necessary, but it is shown to emphasize the repetitive nature of the calculation which the machine finds easy. The calculations were done carrying roughly 8 digits; the output shows 4 decimal places for numbers greater than 0.1 and 4 digits otherwise. The final $\mathbf{Q}_{(+)}$ resulting from 8 sweeping operations may be compared with the original $\mathbf{Q}_{(+)}$ to gain some idea of the effect of rounding error on the output. A rounding error of 5 in the fourth digit is roughly typical. The quantities which are interpreted statistically are based on fewer sweeping operations and should in general be correct to 3 digits at least.

From the third line of $\mathbf{SWP}[1,2,4]\mathbf{Q}_{(+)}$ it follows that

$$\hat{V}_3 = 0.3208V_0 + 0.7280V_1 + 1.8109V_2$$

while the residual sum of squares is 0.001803. From the third line of $\mathbf{SWP}[1,2]\mathbf{Q}_{(+)}$ it follows that

$$V_3^* = 0.8577V_1 + 2.1682V_2$$

while the residual sum of squares is 0.001910. Finally, a simple desk calculation on the original data matrix shows that the "theoretical" predictor

$$V_3^{**} = V_1 + 2V_2$$

has a residual sum of squares 0.002226.

The three predictors $\hat{V}_3$, V_3^*, and V_3^{**} when applied to the sample yield the following 3 columns of residuals, each calculated to 3 decimal places:

0.001	−0.005	−0.007
0.004	0.008	0.007
−0.003	−0.003	−0.003
0.008	0.006	0.003
0.008	0.004	−0.003
−0.028	−0.025	−0.029
0.007	0.008	0.011
−0.017	−0.020	−0.025
−0.007	−0.007	−0.017
0.017	0.022	0.015
0.000	0.001	−0.005
0.012	0.013	0.011

In considering these 3 vectors of residuals as points in the 12-dimensional Euclidean space $\mathscr{N}$, it should be remembered that the first is constrained to lie in the 9-dimensional subspace orthogonal to the vectors corresponding to V_0, V_1, and V_2, and similarly that the second is constrained to lie in the 10-dimensional subspace orthogonal to the vectors corresponding to V_1 and V_2. Consequently, the residual sums of squares 0.001803, 0.001910, and 0.002226 are squared lengths constrained to 9, 10, and 12 dimensions, respectively. To make them comparable, they are often divided by their associated dimension or *degree of freedom* number, leading to the three *residual mean squares* 0.0002003, 0.0001910, and 0.0001855. On this measure, the theoretical predictor V_3^{**} appears most accurate, although the differences are slight. Visual inspection of the three columns of residual does not turn up any striking differences in pattern.

Ignoring the crudeness of the data, the example illustrates a difficult scientific question. Is it better to use a theoretical predictor with given regression coefficients or a predictor from a wider model with fitted regression coefficients?

With finite sample sizes the sampling error from fitted regression coefficients may well exceed the actual error of the postulated theoretical regression coefficients, and whether or not this happens is unknown. Consequently, a real dilemma is posed. One imperfect solution to the dilemma is to prefer the theoretical model unless there are sufficient data to contradict the theoretical model in the sense of significance testing. Such significance tests will be discussed briefly in Section 14.2, but meanwhile the analysis should convey the feeling that fitting has not produced any clear cut improvement over the theoretical predictor, and correspondingly no denial of the ellipsoid model for these eggs.

In Example 8.1 use was made of the idea that variables may be represented by points in the n-dimensional space $\mathscr{N}$. For example, various residual sums of squares were interpreted as squared lengths in subspaces of $\mathscr{N}$. The dual geometric representation of multiple regression analysis in p-dimensional individual-space $\mathscr{F}$ is less obvious. Here the sample is represented either by the n sample individuals or by the sample mean-centered concentration ellipsoid. This description is dual to the description already given of $\dot{V}_p$ in $\mathscr{E}$ as the orthogonal projection of V_p into the subspace spanned by $V_1, V_2, \ldots, V_{p-1}$, in accordance with the sample covariance inner product. It is therefore clear at the outset that, whereas in $\mathscr{E}$ projection along a family of parallel lines was involved, in $\mathscr{F}$ projection along a family of parallel $(p-1)$-dimensional hyperplanes will be involved.

Consider the hyperplane in $\mathscr{F}$ consisting of the points $\mathbf{xv}$ where the coordinates $\mathbf{x}$ relative to the basis $\mathbf{v}$ dual to $\mathbf{V}$ in $\mathscr{E}$ satisfy the equation

$$x_p = w + w_1x_1 + \cdots + w_{p-1}x_{p-1}. \tag{8.2.21}$$

According to the criterion (8.2.2), multiple regression analysis may be regarded as the task of finding that hyperplane of the form (8.2.21) such that the sum of squares of the deviations of the sample individuals $a_1, a_2, \ldots, a_n$ from the hyperplane along a direction parallel to v_p is minimized. The resulting optimum hyperplane may be called the *sample regression hyperplane*. Since the sample regression hyperplane may also be expressed by the equation

$$x_p - \bar{X}_p = w_1(x_1 - \bar{X}_1) + w_2(x_2 - \bar{X}_2) + \cdots + w_{p-1}(x_{p-1} - \bar{X}_{p-1}), \tag{8.2.22}$$

it clearly passes through the sample mean point.

For any hyperplane such as (8.2.21), define αv_p to be its intersection with the axis defined by v_p, and define $\alpha_1 v_p, \alpha_2 v_p, \ldots, \alpha_n v_p$ to be the linear projections of $a_1, a_2, \ldots, a_n$ into the same axis along hyperplanes parallel to (8.2.21). Then the least squares criterion may be expressed as

$$\sum_{i=1}^{n} (\alpha_i - \alpha)^2. \tag{8.2.23}$$

Thus the task of multiple regression analysis is that of finding the hyperplane of the form (8.2.21) such that the projected sample $\alpha_1 v_p, \alpha_2 v_p, \ldots, \alpha_n v_p$ has the smallest clustering in the sense of (8.2.23). Note that for this smallest clustering

$$\alpha = \frac{1}{n} \sum_{i=1}^{n} \alpha_i \tag{8.2.24}$$

because the regression hyperplane is known to pass through the sample mean. Thus, the criterion (8.2.23) to be minimized may also be taken to be

$$\sum_{i=1}^{n} (\alpha_i - \bar{\alpha})^2. \tag{8.2.25}$$

The final geometric characterization dispenses with the points $a_1, a_2, \ldots, a_n$ and makes use only of the sample mean-centered concentration ellipsoid. Theorem 7.3 ensures that the mean-centered concentration ellipsoid of the projected sample $\alpha_1 v_p, \alpha_2 v_p, \ldots, \alpha_n v_p$ is the shadow cast by the mean-centered

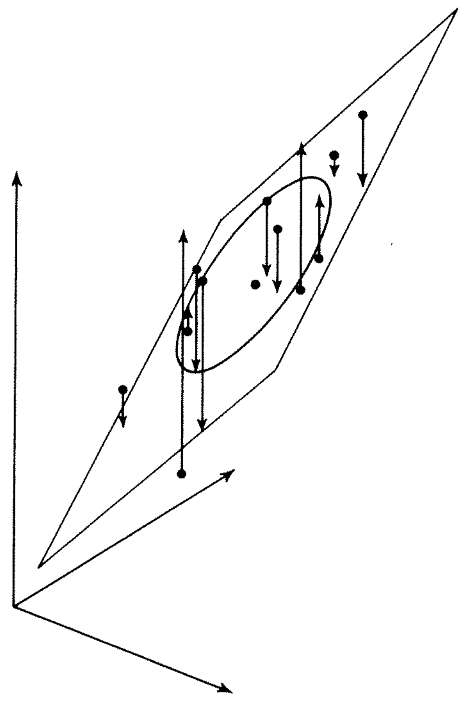

Fig. 8.2.1. The data of Example 8.1 as plotted in Fig. 7.3.2, showing in addition the fitted part of each point (marked by an arrowhead) and the best fitting hyperplane through the sample mean.

ellipsoid of concentration of the full sample under projection along the family of hyperplanes parallel to (8.2.21). It is easily checked that the mean-centered ellipsoid of concentration of the univariate sample $\alpha_1 v_p, \alpha_2 v_p, \ldots, \alpha_n v_p$ consists simply of the line segment joining the two points

$$\left\{\bar{\alpha} \pm \left[\sum_{i=1}^{n} (\alpha_i - \bar{\alpha})^2/(n-1)\right]^{1/2}\right\} v_p. \tag{8.2.26}$$

Consequently, minimizing the criterion (8.2.25) is equivalent to minimizing the length of the shadow (8.2.26). It is easy to see in geometric terms how this shadow is to be minimized. Consider the points $m \pm \beta v_p$ on the mean-centered ellipsoid of concentration of the full sample, i.e., the points where the line through the center of the ellipsoid parallel to v_p meets the ellipsoid. Clearly the shadow cast by the line segment from $m - \beta v_p$ to $m + \beta v_p$ lies in any shadow cast by the ellipsoid. Also, if $m \pm \beta v_p + \mathscr{V}$ denote the tangent hyperplanes of dimension $p - 1$ to the ellipsoid at $m \pm \beta v_p$, then projection along hyperplanes parallel to $\mathscr{V}$ casts a shadow identical to the shadow cast by the line segment from $m - \beta v_p$ to $m + \beta v_p$. It follows that $m + \mathscr{V}$ must be the desired regression hyperplane and, by comparing β with (8.2.26), that

$$\beta = \left[\sum_{i=1}^{n} (\alpha_i - \bar{\alpha})^2/(n-1)\right]^{1/2}, \tag{8.2.27}$$

where the right side of (8.2.27) is the sample standard deviation of the tightest projected sample.

The results of the last three paragraphs are illustrated in Fig. 8.2.1 which takes the data of Example 8.1 as plotted in Fig. 7.3.2 and adds the regression hyperplane.

8.3 REGRESSION COEFFICIENTS, CORRELATION COEFFICIENTS, AND THE MULTIPLE REGRESSION ANALYSIS OF V_p ON A SUBSET OF $V_1, V_2, \ldots, V_{p-1}$

Up to now only the regression analysis of V_p on the whole set $V_1, V_2, \ldots, V_{p-1}$ has been explicitly considered. The realities of data analysis often require or suggest several analyses of the same sample, so that it becomes advisable to understand certain relationships among the regression analyses of V_p on different subsets of $V_1, V_2, \ldots, V_{p-1}$. The relevant quantities here are regression coefficients and correlation coefficients.

In general, consider the multiple regression analysis of V_r on $V_{s_1}, V_{s_2}, \ldots, V_{s_t}$. The best linear predictor for V_r in this analysis will be denoted by

$$\sum_{i=1}^{t} w_{rs_i(s_1s_2\ldots s_t)} V_{s_i}, \tag{8.3.1}$$

where $w_{rs_i(s_1s_2\ldots s_t)}$ will be called the regression coefficient of V_r on V_{s_i} in the multiple regression analysis of V_r on $V_{s_1}, V_{s_2}, \ldots, V_{s_t}$. In this more general notation, the regression coefficients w_i appearing in (8.2.1), (8.2.4), or (8.2.16) are denoted by $w_{pi(12\ldots\overline{p-1})}$ for $i = 1, 2, \ldots, p - 1$.

A regression coefficient such as $w_{rs(s)}$, which refers to the regression analysis of V_r on V_s alone, may be called a *simple regression coefficient* and may be denoted simply by w_{rs}. Thus

$$w_{rs} = w_{rs(s)} = \frac{\operatorname{cov}(V_r, V_s)}{\operatorname{var}(V_s)}. \tag{8.3.2}$$

By contrast the general type of regression coefficient as appears in (8.3.1) with $t \geq 2$ may be called a *joint regression coefficient.* Suppose that $\mathbf{V}^* = \mathbf{AV}$ denotes the basis of variable-space $\mathscr{E}$ resulting from the successive orthogonalization of the basis $\mathbf{V} = [V_1, V_2, \ldots, V_p]'$, where the inner product is defined by the sample covariance. As in Sections 4.2 and 4.3, the matrix $\mathbf{A}$ is a triangular matrix with elements zero above the diagonal and unity along the diagonal. The remaining elements of $\mathbf{A}$ are all joint regression coefficients. In fact, $\mathbf{A}$ may be written

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ -w_{21(1)} & 1 & \cdots & 0 \\ -w_{31(12)} & -w_{32(12)} & \cdots & 0 \\ \cdot & \cdot & & \\ \cdot & \cdot & & \\ \cdot & \cdot & & \\ -w_{p1(12\ldots\overline{p-1})} & -w_{p2(12\ldots\overline{p-1})} & \cdots & 1 \end{bmatrix}, \tag{8.3.3}$$

for line r of $\mathbf{A}$ shows that $V_r - V_r^* = w_{r1(12\ldots\overline{r-1})}V_1 + w_{r2(12\ldots\overline{r-1})}V_2 + \cdots + w_{r\overline{r-1}(12\ldots\overline{r-1})}V_{r-1}$ where $V_r - V_r^*$ is the component of V_r along the subspace spanned by $V_1, V_2, \ldots, V_{r-1}$, i.e., $V_r - V_r^*$ is the reduced best linear predictor for V_r in terms of $V_1, V_2, \ldots, V_{r-1}$, as is required to demonstrate (8.3.3).

It is illuminating to introduce a terminology of partial regression coefficients, even though it will subsequently turn out that all *partial* regression coefficients are simply joint regression coefficients in disguise. After removing the components along $V_{s_1}, V_{s_2}, \ldots, V_{s_t}$ from each of $V_r, V_{q_1}, V_{q_2}, \ldots, V_{q_m}$, one may contemplate the regression analysis of $V_{r.s_1s_2\ldots s_t}$ on $V_{q_1.s_1s_2\ldots s_t}$, $V_{q_2.s_1s_2\ldots s_t}, \ldots, V_{q_m.s_1s_2\ldots s_t}$. Actually, since $V_r - V_{r.s_1s_2\ldots s_t}$ is orthogonal to the predictor variables $V_{q_1.s_1s_2\ldots s_t}, V_{q_2.s_1s_2\ldots s_t}, \ldots, V_{q_m.s_1s_2\ldots s_t}$, the regression analysis of V_r or $V_{r.s_1s_2\ldots s_t}$ on these predictor variables produces the same best linear predictor whose coefficients may be denoted by

$$w_{rq_i(q_1q_2\ldots q_m).s_1s_2\ldots s_t}, \tag{8.3.4}$$

which represents the general form of a *partial (joint) regression coefficient.* Examples of these are the elements of the matrix $\mathbf{B}$ produced by successive orthogonalization where $\mathbf{V} = \mathbf{B}\mathbf{V}^*$, namely

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ w_{21} & 1 & 0 & \cdots & 0 \\ w_{31} & w_{32.1} & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & & \\ w_{p1} & w_{p2.1} & w_{p3.12} & \cdots & 1 \end{bmatrix}. \tag{8.3.5}$$

Formula (8.3.5) follows by noting that column s on the right side of (8.3.5) may be determined from

$$\begin{aligned} w_{rs.12\ldots\overline{s-1}} &= w_{rs(s).12\ldots\overline{s-1}} \\ &= \frac{\operatorname{cov}(V_{r.12\ldots\overline{s-1}}, V_{s.12\ldots\overline{s-1}})}{\operatorname{var}(V_{s.12\ldots\overline{s-1}})} \end{aligned} \tag{8.3.6}$$

for $r = s + 1, s + 2, \ldots, p$, and the rule for computing column s of $\mathbf{B}$ given in Section 4.3.1 agrees with (8.3.6).

It is now time to remark that

$$w_{rq_i(q_1q_2\ldots q_m).s_1s_2\ldots s_t} = w_{rq_i(q_1q_2\ldots q_ms_1s_2\ldots s_t)}, \tag{8.3.7}$$

so that all partial regression coefficients may be interpreted simply as joint regression coefficients, and vice versa. Since dimensions may be relabeled and subscripts may be permuted, it will be sufficient to prove that

$$w_{p\overline{p-1}\,(\overline{s+1}\,\overline{s+2}\ldots\overline{p-1}).12\ldots s} = w_{p\overline{p-1}\,(12\ldots\overline{p-1})}.$$

To prove this, one need only follow through the computation of the right side by successively applying the operations SWP[1], SWP[2], $\ldots$, SWP[$p - 1$] in stages to the covariance matrix $\mathbf{S}$. One of the parts of SWP[$1, 2, \ldots, s$]$\mathbf{S}$ which results from the first s stages is the covariance matrix $\mathbf{S}_{(012\ldots s)}$ of the component variables $V_{\overline{s+1}.12\ldots s}$, $V_{\overline{s+2}.12\ldots s}, \ldots, V_{p.12\ldots s}$, and $w_{p\overline{p-1}(\overline{s+1}\,\overline{s+2}\ldots p).12\ldots s}$ may be calculated by further sweep operations on this $\mathbf{S}_{(012\ldots s)}$. The reader may check that these subsequent sweep operations on $\mathbf{S}_{(012\ldots s)}$ are actually included in the last $(p - s - 1)$ stages of the original $(p - 1)$ stages of sweep operations on $\mathbf{S}$ and consequently produce the same result for $w_{p\overline{p-1}(12\ldots p)}$ or $w_{p\overline{p-1}(\overline{s+1}\,\overline{s+2}\ldots p).12\ldots s}$.

In view of (8.3.7) it may be simpler always to use the joint regression coefficient notation, remembering, for example, that $w_{12(234)} = w_{12.34} = w_{12(23).4} = w_{12(24).3}$. It also follows from (8.3.7) together with (8.3.5) that

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ w_{21} & 1 & 0 & 0 & \cdots & 0 \\ w_{31} & w_{32(12)} & 1 & 0 & \cdots & 0 \\ w_{41} & w_{42(12)} & w_{43(123)} & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & \cdot & & \\ w_{p1} & w_{p2(12)} & w_{p3(123)} & w_{p4(1234)} & \cdots & 1 \end{bmatrix}. \tag{8.3.8}$$

From (8.3.8) and (8.3.3), together with $\mathbf{AB} = \mathbf{BA} = \mathbf{I}$, numerous identities involving regression coefficients may be deduced, but these are left for the interested reader to explore.

There is a class of correlation coefficients comparable to the class of regression coefficients relating a set of variables $V_1, V_2, \ldots, V_p$. Simple correlation coefficients have already been defined in (7.4.10). Thus, among $V_1, V_2, \ldots, V_p$ there are $p(p - 1)/2$ different simple correlation coefficients

$$\begin{aligned} r_{st} &= \operatorname{cor}(V_s, V_t) \\ &= \operatorname{cov}(V_s, V_t)/\operatorname{var}(V_s)^{1/2}\operatorname{var}(V_t)^{1/2} \\ &= r_{ts} \end{aligned} \tag{8.3.9}$$

for $1 \le s < t \le p$. These correlation coefficients together with $\operatorname{var}(V_s)$ for $s = 1, 2, \ldots, p$ determine the covariance matrix of $V_1, V_2, \ldots, V_p$. In geometric terms, the correlation coefficient r_{st} with the sample covariance inner product determines the angle between V_s and V_t in variable-space $\mathscr{E}$, as indicated by (7.4.12). Alternatively, r_{st} determines the angle between the corresponding pair of vectors in $\mathscr{N}_{II}$.

The correlation coefficient r_{st} may be regarded as the covariance between the *standardized variables* $U_s = \operatorname{var}(V_s)^{-1/2}V_s$ and $U_t = \operatorname{var}(V_t)^{-1/2}V_t$. Indeed, the sample correlation matrix $\mathbf{R}$ of $V_1, V_2, \ldots, V_p$ is simply the sample covariance matrix of the standardized basis $U_1, U_2, \ldots, U_p$. It follows easily that r_{st} may also be interpreted as the simple regression coefficient of either U_s on U_t or U_t on U_s.

The *multiple correlation coefficient* $r_{s(t_1t_2\ldots t_m)}$ between V_s and a set of m variables $V_{t_1}, V_{t_2}, \ldots, V_{t_m}$ is defined to be the simple correlation coefficient between V_s and the reduced best linear predictor for V_s in terms of $V_{t_1}, V_{t_2}, \ldots, V_{t_m}$. Denote this best linear predictor by $V_{s(t_1t_2\ldots t_m)}$, so that the orthogonal decomposition of V_s along and orthogonal to the subspace spanned by $V_{t_1}, V_{t_2}, \ldots, V_{t_m}$ is given by

$$V_s = V_{s(t_1t_2\ldots t_m)} + V_{s.t_1t_2\ldots t_m}. \tag{8.3.10}$$

Then

$$r_{s(t_1t_2\ldots t_m)} = \frac{\operatorname{cov}(V_s, V_{s(t_1t_2\ldots t_m)})}{\operatorname{var}(V_s)^{1/2}\operatorname{var}(V_{s(t_1t_2\ldots t_m)})^{1/2}}. \tag{8.3.11}$$

Since the angle between V_s and $V_{s(t_1t_2\ldots t_m)}$ is no greater than $\pi/2$,

$$0 \le r_{s(t_1t_2\ldots t_m)} \le 1. \tag{8.3.12}$$

Since $\operatorname{cov}(V_s - V_{s(t_1t_2\ldots t_m)}, V_{s(t_1t_2\ldots t_m)}) = 0$, or $\operatorname{cov}(V_s, V_{s(t_1t_2\ldots t_m)}) = \operatorname{var}(V_{s(t_1t_2\ldots t_m)})$, it follows that

$$r_{s(t_1t_2\ldots t_m)} = \operatorname{var}(V_{s(t_1t_2\ldots t_m)})^{1/2}/\operatorname{var}(V_s)^{1/2}. \tag{8.3.13}$$

Or, since var $(V_s) =$ var $(V_{s(t_1t_2\ldots t_m)}) +$ var $(V_{s.t_1t_2\ldots t_m})$, it follows that

$$1 - r^2_{s(t_1t_2\ldots t_m)} = \text{var}\,(V_{s.t_1t_2\ldots t_m})/\text{var}\,(V_s). \qquad (8.3.14)$$

Note that if $r_{s(t_1t_2\ldots t_m)}$ is regarded as $\cos\theta$, then (8.3.14) is $\sin^2\theta$.

Partial correlation coefficients are simply correlation coefficients among variables from which the same set of components has been removed. Thus $r_{t(q_1q_2\ldots q_u).s_1s_2\ldots s_m}$ is defined to be the multiple correlation coefficient between $V_{t.s_1s_2\ldots s_m}$ and the set $V_{q_1.s_1s_2\ldots s_m}, V_{q_2.s_1s_2\ldots s_m}, \ldots, V_{q_u.s_1s_2\ldots s_m}$. Unlike partial regression coefficients, partial correlation coefficients do not provide quantities directly expressible as multiple correlation coefficients already defined.

It is often convenient to think of a correlation coefficient r in terms of the quantity $1 - r^2$ which may always be regarded as a fraction of variance remaining after fitting a best linear predictor, as in (8.3.14). By considering successive reductions in variance from fitting linear predictors, one may immediately write down such identities as

$$1 - r^2_{t(q_1q_2\ldots q_us_1s_2\ldots s_m)} = (1 - r^2_{t(s_1s_2\ldots s_m)})(1 - r^2_{t(q_1q_2\ldots q_u).s_1s_2\ldots s_m}), \quad (8.3.15)$$

and

$$1 - r^2_{p(12\ldots\overline{p-1})} = \prod_{i=1}^{p-1}(1 - r^2_{pi.12\ldots\overline{i-1}}). \qquad (8.3.16)$$

The basic property of correlation coefficients which motivates their definition is that they are dimensionless, i.e., if $r =$ cor (V, W), then $r =$ cor $(\mu V, \nu W)$ for any $\mu \neq 0$ and $\nu \neq 0$. In other words, linear changes of scale do not affect a correlation coefficient. The appeal of correlation coefficients as a tool for interpreting data is closely tied to this invariance property.

Regression coefficients, on the other hand, are always measured in the units of a ratio of two variables. For example, if $\dot{V}_3 = w_1V_1 + w_2V_2$ is a predictor for V_3 where V_3 is height in inches and V_1 is weight in pounds, then w_1 must be measured in units of inches per pound. Generally, $w_{st_1(t_1t_2\ldots t_m)}$ is measured in units of V_s divided by V_{t_1}. This dependence on units must be remembered when regression coefficients are regarded as measures of association between variables, for a large regression coefficient may only reflect a particular choice of scale for the variables concerned. In this sense regression coefficients require more careful interpretation than correlation coefficients.

Individual regression coefficients also require careful interpretation because they can depend strongly on the set of variables included as predictors in the multiple regression analysis. For example, although $w_{p1}, w_{p1(12)}, \ldots, w_{p1(12\ldots p-1)}$ all have the same dimensions, they may still vary greatly. The safest attitude to assume toward an individual joint regression coefficient is to regard it as a simple partial regression coefficient, i.e., to regard $w_{rs_1(s_1s_2\ldots s_m)}$ as $w_{rs_1.s_2\ldots s_m}$. In this way $w_{rs_1(s_1s_2\ldots s_m)}$ is seen to be the weight applied to $V_{s_1.s_2\ldots s_m}$ as a single predictor for V_r. In other words, any joint regression coefficient may always be regarded as a weight applied to V_{s_1} after all the other predictors $V_{s_2}, \ldots, V_{s_m}$ have been taken into account.

The multiple regression analysis of V_p on $V_1, V_2, \ldots, V_{p-1}$ has the associated fractional reductions of variance $1 - r^2_{p(s_1s_2\ldots s_m)}$ where $s_1, s_2, \ldots, s_m$ is any subset of $1, 2, \ldots, p - 1$. There are $2^{p-1} - 1$ such nonempty subsets, and the corresponding set of $2^{p-1} - 1$ fractional reductions in variance provides a clear picture of the interaction of the variables $V_1, V_2, \ldots, V_{p-1}$ in their ability to jointly explain var (V_p). In statistical practice it often happens that a predictor for V_p is chosen which depends only on a subset of the available variables $V_1, V_2, \ldots, V_{p-1}$. The reasons for such a restricted predictor may be of two different sorts. First, for reasons of time, money, or effort it may be deemed impractical to expect anyone to make use of a predictor requiring the observation of the complete set $V_1, V_2, \ldots, V_{p-1}$. In such cases, a loss in prediction accuracy may be judged to be offset by increased practicability. Secondly, it is possible that deleting certain variables may result in an *increase* of prediction accuracy, because predictors based on finite samples find it hard to digest larger and larger numbers of independent variables. This issue was raised briefly in Example 8.1. To take another more extreme example, it may be noted that *multiple regression analysis is not even defined when* $p - 1 > n - 1$. For **S** has rank at most $n - 1$, and fitting $n - 1$ predictor variables is sufficient to reduce the residual variance of V_p to zero, so that no further fitting can be done. Theoretical understanding of this phenomenon of diminishing returns for variables introduced remains imperfect, while the phenomenon itself can be demonstrated empirically by making use of different predictors.

There are several standard methods for choosing a subset of the set of possible predictors. To simplify a discussion of these, make the unrealistic assumption that a decision has been made to include precisely k predictors. (This assumption is unrealistic because, for example, only one predictor might be of any value. Or, having chosen k predictors, it might be obvious that great benefits would accrue from the inclusion of a $(k + 1)$st predictor.) There are two popular methods of selecting k predictors. The first, which may be called the *forward method*, selects variables one at a time by the following rule:

i) Choose V_{s_1} out of $V_1, V_2, \ldots, V_{p-1}$ so that $r^2_{ps_1} \geq r^2_{pi}$ for $i = 1, 2, \ldots, p - 1$;
ii) Choose V_{s_2} out of the remaining V_i with $i \neq s_1$ so that $r^2_{ps_2.s_1} \geq r^2_{pi.s_1}$;
iii) Choose V_{s_3} out of the remaining V_i with $i \neq s_1, s_2$ so that $r^2_{ps_3.s_1s_2} \geq r^2_{pi.s_1s_2}$,

and so on until $V_{s_1}, V_{s_2}, \ldots, V_{s_k}$ have been chosen. In other words, the variables are chosen to yield the greatest reduction of residual sum of squares at each step of the introduction of a single predictor variable into the multiple regression analysis. The *backward method* begins from the complete regression analysis with predictor variables $V_1, V_2, \ldots, V_{p-1}$ and deletes one at a time from the analysis in such a way as to leave the minimum residual sum of squares at each stage or, equivalently, to leave the maximum multiple correlation between V_p and the predictor variables remaining at any stage. There are many variations on these methods. For example, the best pair of variables might be included at each stage in the forward method, and a similar modification could be made in the backward method. For a preselected k, it is plausible but computationally burdensome to look at all $\binom{p-1}{k}$ different sets of possible predictors and choose that with the smallest residual sum of squares.

It is nearly obvious that the different selection methods may give different results. For example, with three predictor variables V_1, V_2, and V_3, it may happen that $V_1 - V_2$ is a perfect predictor for V_4 while neither V_1 nor V_2 alone is as good as V_3 alone. To construct such an example suppose that U_1, U_2, U_3 are an orthonormal set of variables and define the set V_1, V_2, V_3, V_4 as follows: $V_4 = U_3$, $V_1 = (U_3 + U_1)/2$, $V_2 = (U_3 - U_1)/2$, and $V_3 = U_3 + U_2/4$. It follows easily that $r^2_{43} = \frac{16}{17}$ while $r^2_{42} = r^2_{41} = \frac{1}{2}$, and yet $r^2_{4(12)} = 1$ while $r^2_{4(13)} = r^2_{4(23)} = \frac{17}{18}$. This example illustrates the dilemma facing the user of either the forward or backward methods—by following down single chains of variables he must exclude examination of many pairs, triples, etc. which may have high predictive content. On the other hand, he may doubt the existence of such hidden combinations and be unwilling or unable to do the computations necessary to find them.

The forward scheme requires the least computing labor and is therefore the most used, especially for large p. Indeed the backward method may be computationally impractical for large p because it requires first carrying out the complete analysis. Theoretical considerations leading to a good method of selection remain generally undiscovered. In many examples, of course, various different methods of selection will produce effectively, if not exactly, the same result.

Example 8.2. This example is based on the data used by Cochran (1938) to illustrate the computations associated with the deletion or addition of a variable in multiple regression analysis. The following description is quoted from Cochran's paper:

> In a study of the effects of weather factors on the numbers of noctuid moths per night caught in a light trap, regressions were worked out on the minimum night temperature, the maximum temperature of the previous day, the average speed of the wind during the night and the amount of rain during the night. The dependent variable was log (number of moths + 1). This was found to be roughly normally distributed, whereas the numbers themselves had an extremely skew distribution. Further, a change in one of the weather factors was likely to produce the same *percentage* change at different times in the numbers of moths rather than the same *actual* change. Three years' data were included. These were grouped in blocks of nine consecutive days, so as to eliminate as far as possible the effects of the lunar cycle. After the removal of differences between blocks, 72 degrees of freedom remained for the regressions.

Interest centered on the effect of including night cloud cover as a fifth predictor variable. The basic data culled from Cochran's paper is a 6 × 6 corrected sum inner product matrix, where to correct is to subtract out the block means of each block of nine days. In this way certain dimensions in $\mathcal{N}$ were removed from the data vectors before the analysis started in order to eliminate the influence of factors not relevant to weather. This corrected sum inner product matrix is

```
T =
  0.14029E 02   0.56635E 01   0.19866E 01   0.27330E 01  −0.48670E 01   0.20744E 01
  0.56635E 01   0.14537E 02   0.12710E-00  −0.13470E 01   0.20600E-00   0.15747E 01
  0.19866E 01   0.12710E-00   0.20680E 01   0.29400E-00  −0.54460E 00  −0.64400E 00
  0.27330E 01  −0.13470E 01   0.29400E-00   0.17110E 02  −0.54200E 01   0.88500E 00
 −0.48670E 01   0.20600E-00  −0.54460E 00  −0.54200E 01   0.78700E 01  −0.19330E 01
  0.20744E 01   0.15747E 01  −0.64400E 00   0.88500E 00  −0.19330E 01   0.35520E 01
```

The numbers here are in "floating point" computer output where, for example, an exponent E 01 means that the given number should be multiplied by $10 = 10^1$ or an exponent E-02 means that the given number should be multiplied by $0.01 = 10^{-2}$. The 6 rows and columns of **T** refer to the variables V_1 = minimum night temperature, V_2 = maximum day temperature, V_3 = average night wind speed, V_4 = amount of night rainfall, V_5 = percentage of starlight obscured by clouds in a night sky camera, and V_6 = log (number of moths caught + 1).

The usual step-by-step process of finding the multiple regression analysis of V_6 on V_1, V_2, V_3, V_4, V_5 was carried out in a computer by finding

```
SWP[1]T =
 −0.71283E-01   0.40371E-00   0.14161E-00   0.19482E-00  −0.34693E-00   0.14787E-00
  0.40371E-00   0.12250E 02  −0.67491E 00  −0.24503E 01   0.21709E 01   0.73724E 00
  0.14161E-00  −0.67491E 00   0.17867E 01  −0.93022E-01   0.14462E-00  −0.93776E 00
  0.19482E-00  −0.24503E 01  −0.93022E-01   0.16578E 02  −0.44718E 01   0.48087E-00
 −0.34693E-00   0.21709E 01   0.14462E-00  −0.44718E 01   0.61815E 01  −0.12133E 01
  0.14787E-00   0.73724E 00  −0.93776E 00   0.48087E-00  −0.12133E 01   0.32453E 01
```

```
SWP[1, 2]T =
 −0.84587E-01   0.32955E-01   0.16385E-00   0.27557E-00  −0.41847E-00   0.12357E-00
  0.32955E-01  −0.81630E-01  −0.55093E-01  −0.20002E-00   0.17721E-00   0.60181E-01
  0.16385E-00  −0.55093E-01   0.17495E 01  −0.22802E-00   0.26422E-00  −0.89714E 00
  0.27557E-00  −0.20002E-00  −0.22802E-00   0.16087E 02  −0.40376E 01   0.62834E 00
 −0.41847E-00   0.17721E-00   0.26422E-00  −0.40376E 01   0.57968E 01  −0.13440E 01
  0.12357E-00   0.60181E-01  −0.89714E 00   0.62834E 00  −0.13440E 01   0.32009E 01
```

```
SWP[1, 2, 3]T =
−0.99933E-01   0.38115E-01   0.93657E-01   0.29692E-00  −0.44322E-00   0.20760E-00
 0.38115E-01  −0.83365E-01  −0.31491E-01  −0.20720E-00   0.18553E-00   0.31929E-01
 0.93657E-01  −0.31491E-01  −0.57159E 00  −0.13033E-00   0.15103E-00  −0.51280E 00
 0.29692E-00  −0.20720E-00  −0.13033E-00   0.16058E 02  −0.40032E 01   0.51141E 00
−0.44322E-00   0.18553E-00   0.15103E-00  −0.40032E 01   0.57569E 01  −0.12085E 01
 0.20760E-00   0.31929E-01  −0.51280E 00   0.51141E 00  −0.12085E 01   0.27408E 01
```

```
SWP[1, 2, 3, 4]T =
−0.10542E-00   0.41946E-01   0.96067E-01   0.18491E-01  −0.36920E-00   0.19814E-00
 0.41946E-01  −0.86039E-01  −0.33173E-01  −0.12904E-01   0.13387E-00   0.38528E-01
 0.96067E-01  −0.33173E-01  −0.57265E 00  −0.81166E-02   0.11853E-00  −0.50865E 00
 0.18491E-01  −0.12904E-01  −0.81166E-02  −0.62275E-01  −0.24930E-00   0.31848E-01
−0.36920E-00   0.13387E-00   0.11853E-00  −0.24930E-00   0.47589E 01  −0.10810E 01
 0.19814E-00   0.38528E-01  −0.50865E 00   0.31848E-01  −0.10810E 01   0.27245E 01
```

```
SWP[1, 2, 3, 4, 5]T =
−0.13407E-00   0.52332E-01   0.10526E-00  −0.84984E-03  −0.77581E-01   0.11428E-00
 0.52332E-01  −0.89805E-01  −0.36507E-01  −0.58905E-02   0.28131E-01   0.68938E-01
 0.10526E-00  −0.36507E-01  −0.57560E 00  −0.19071E-02   0.24908E-01  −0.48172E-00
−0.84984E-03  −0.58905E-02  −0.19071E-02  −0.75335E-01  −0.52386E-01  −0.24780E-01
−0.77581E-01   0.28131E-01   0.24908E-01  −0.52386E-01  −0.21013E-00  −0.22715E-00
 0.11428E-00   0.68938E-01  −0.48172E-00  −0.24780E-01  −0.22715E-00   0.24790E 01
```

These calculations were done to roughly 16 digit accuracy, and as a result, when RSW[1, 2, 3, 4, 5] was applied to SWP[1, 2, 3, 4, 5], the original **T** was reproduced exactly to the five digits shown in the output.

The output above provides the regression analysis of V_6 on each of the sets V_1 to V_i for $i = 1, 2, 3, 4, 5$. The main point of Cochran's paper was to illustrate the computations required to add V_5 to the predictor based on V_1, V_2, V_3, V_4 and to delete V_5 from the predictor based on V_1, V_2, V_3, V_4, V_5. These computations involve, in the language of this book, the operations SWP[5] applied to SWP[1, 2, 3, 4]**T** and RSW[5] applied to SWP[1, 2, 3, 4, 5]**T**. Note, however, that in Cochran's context it was necessary to assimilate V_5 before the 6 × 6 matrix SWP[1, 2, 3, 4]**T** was available.

The various simple and joint regression coefficients of V_6 on V_1 produced by this analysis are given by the (1, 6) elements of the above matrices:

$$\begin{aligned} w_{61} &= 0.14787 \\ w_{61(12)} &= 0.12357 \\ w_{61(123)} &= 0.20760 \\ w_{61(1234)} &= 0.19814 \\ w_{61(12345)} &= 0.11428. \end{aligned}$$

It is clear that the weight given to V_1 depends considerably on what other predictor variables are used.

The initial corrected sum of squares of V_6 is shown as the (6, 6) element of **T** to be 3.5520 and is seen to be reduced successively to 3.2453, 3.2009, 2.7408, 2.7245, and 2.4790 by successively adding V_1, V_2, V_3, V_4, and V_5 to the set of fitted variables. The corresponding fractions of residual variance to total variance are

$$\begin{aligned} 1 - r^2_{61} &= 3.2453/3.5520 = 0.9137 \\ 1 - r^2_{6(12)} &= 3.2009/3.5520 = 0.9012 \\ 1 - r^2_{6(123)} &= 2.7408/3.5520 = 0.7716 \\ 1 - r^2_{6(1234)} &= 2.7245/3.5520 = 0.7670 \\ 1 - r^2_{6(12345)} &= 2.4790/3.5520 = 0.6979. \end{aligned}$$

Various partial correlation coefficients may also be deduced, such as

$$\begin{aligned} 1 - r^2_{62.1} &= 3.2009/3.2453 = 0.9863 \\ 1 - r^2_{63.12} &= 2.7408/3.2009 = 0.8563 \\ 1 - r^2_{64.123} &= 2.7245/2.7408 = 0.9941 \\ 1 - r^2_{65.1234} &= 2.4790/2.7245 = 0.9099. \end{aligned}$$

To pursue the analysis of various correlation coefficients, the matrix **T** is reduced to the corresponding correlation matrix **R** where

```
R =
 0.10000E 01   0.39659E-00   0.36883E-00   0.17640E-00  −0.46320E-00   0.29387E-00
 0.39659E-00   0.10000E 01   0.23181E-01  −0.85410E-01   0.19259E-01   0.21914E-00
 0.36883E-00   0.23181E-01   0.10000E 01   0.49425E-01  −0.13499E-00  −0.23762E-00
 0.17640E-00  −0.85410E-01   0.49425E-01   0.10000E 01  −0.46708E-00   0.11352E-00
−0.46320E-00   0.19259E-01  −0.13499E-00  −0.46708E-00   0.10000E 01  −0.36560E-00
 0.29387E-00   0.21914E-00  −0.23762E-00   0.11352E-00  −0.36560E-00   0.10000E 01
```

The largest correlation coefficients in this array are those relating V_1 with V_2, V_3, and V_5 and those relating V_5 with V_4 and V_6. None of these attain 0.5 in absolute value, however, so that no variable explains as much as 25% of the variance of any other.

The computer then produced SWP[$i_1, i_2, \ldots, i_t$]**R** where $i_1, i_2, \ldots, i_t$ run over all subsets of 1, 2, 3, 4, 5. The (6, 6) elements of these matrices yield $1 - r^2_{6(i_1 i_2 \ldots i_t)}$ for these subsets as shown in Table 8.3.1.

It is now clear that the forward method of variable selection would choose in order V_5, V_3, V_1, V_2, V_4. Also the backward method would drop in succession from the complete set V_4, V_2, V_5, V_3, V_1, i.e., the order of importance given by the backward method is V_1, V_3, V_5, V_2, V_4, which differs from the forward method. If the best single predictor is desired, namely V_5, it is that given by the forward method but not by the backward method. On the other hand, among all ten pairs of predictors, the best pair V_1, V_3 agrees with the pair given by the backward method but not with the pair given by the forward method.

None of this indicates what predictor should be used, but it appears from Table 8.3.1 that very little is to be gained by including more than three, and

Table 8.3.1

$1 - r^2_{61} = 0.91364$	$1 - r^2_{6(2345)} = 0.72534$
$1 - r^2_{62} = 0.95198$	$1 - r^2_{6(1345)} = 0.71282$
$1 - r^2_{63} = 0.94354$	$1 - r^2_{6(1245)} = 0.81142$
$1 - r^2_{64} = 0.98711$	$1 - r^2_{6(1235)} = 0.70021$
$1 - r^2_{65} = 0.86634$	$1 - r^2_{6(1234)} = 0.76705$
$1 - r^2_{6(12)} = 0.90115$	$1 - r^2_{6(345)} = 0.77766$
$1 - r^2_{6(13)} = 0.77507$	$1 - r^2_{6(245)} = 0.81310$
$1 - r^2_{6(14)} = 0.90972$	$1 - r^2_{6(235)} = 0.72777$
$1 - r^2_{6(15)} = 0.84659$	$1 - r^2_{6(234)} = 0.87189$
$1 - r^2_{6(23)} = 0.89304$	$1 - r^2_{6(145)} = 0.84327$
$1 - r^2_{6(24)} = 0.93436$	$1 - r^2_{6(135)} = 0.71604$
$1 - r^2_{6(25)} = 0.81516$	$1 - r^2_{6(134)} = 0.77190$
$1 - r^2_{6(34)} = 0.92781$	$1 - r^2_{6(125)} = 0.81343$
$1 - r^2_{6(35)} = 0.78246$	$1 - r^2_{6(124)} = 0.89424$
$1 - r^2_{6(45)} = 0.86214$	$1 - r^2_{6(123)} = 0.77163$
$1 - r^2_{6(12345)} = 0.69792$	

V_1, V_3, V_5 appears to be the best triple on the scene. Oddly enough, however, V_2, V_3, V_5 is nearly as good as V_1, V_3, V_5. At the present time, the science of variable selection apparently can do no better than this.

8.4 A LEAST SQUARES EXAMPLE ILLUSTRATING THE DELETION OF INDIVIDUALS

Example 8.3. This example will review the original calculations of Gauss (1811) combining 12 observations on the asteroid Pallas to determine six parameters governing its orbit. The example will be described without going into details of the astronomical context in which least squares prediction first arose. The computations for deleting individuals will also be illustrated.

A set of observed quantities, say γ_i for $i = 1, 2, \ldots, 12$, is thought to be expressible as

$$\gamma_i = f_i(\theta_1, \theta_2, \ldots, \theta_6) + l_i.$$

Here l_i is the error of measurement in the observable quantity γ_i, where, if measurement could be perfect, physical theories would predict that γ_i should be a known function f_i of six quantities $\theta_1, \theta_2, \ldots, \theta_6$ whose values are unknown. The concept of measurement error arises here for the very practical reason that no values for $\theta_1, \theta_2, \ldots, \theta_6$ are sufficient to satisfy all 12 equations $\gamma_i = f_i(\theta_1, \theta_2, \ldots, \theta_6)$ with the observed γ_i. The problem therefore is to combine the 12 observations in order to come as close as possible to the unknown values $\theta_1, \theta_2, \ldots, \theta_6$ despite the known presence of error in the original observations.

The solution proposed by Gauss was to choose $\theta_1, \theta_2, \ldots, \theta_6$ to minimize $\sum_1^{12} [\gamma_i - f_i(\theta_1, \theta_2, \ldots, \theta_6)]^2$. The minimization problem was to be solved by beginning with an initial guess $\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_6^{(0)}$ which is known to be close to the desired answer, perhaps from physical understanding of the parameters. Then the functions f_i were to be approximated linearly by the first term Taylor series approximation

$$f_i(\theta_1, \theta_2, \ldots, \theta_6) = a_i + \sum_{j=1}^{6} (\theta_i - \theta_i^{(0)}) b_{ij},$$

where

$$a_i = f_i(\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_6^{(0)})$$

and

$$b_{ij} = \frac{\partial}{\partial \theta_j} f_i(\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_6^{(0)}).$$

The problem then became to find $\delta\theta_j = \theta_j - \theta_j^{(0)}$ for $j = 1, 2, \ldots, 6$ which minimized

$$\sum_{i=1}^{12} (\gamma_i - a_i - b_{i1}\,\delta\theta_1 - b_{i2}\,\delta\theta_2 - \cdots - b_{i6}\,\delta\theta_6)^2.$$

This is simply the least squares criterion (8.2.18) in a different notation, where $n = 12$ and $p = 7$ and where the role of the $n \times p$ data matrix $\mathbf{X}$ is played by

$$\begin{bmatrix} b_{11} & b_{12} & \cdots & b_{16} & \gamma_1 - a_1 \\ b_{21} & b_{22} & & b_{26} & \gamma_2 - a_2 \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ b_{\overline{12}1} & b_{\overline{12}2} & & b_{\overline{12}6} & \gamma_{\overline{12}} - a_{\overline{12}} \end{bmatrix}.$$

Gauss (1811) gave the above data matrix to be

$\mathbf{X} =$

0.79363	143.66	0.39493	0.95920	−0.18856	0.17387	183.93
−0.02658	46.71	0.02658	−0.20858	0.15946	1.25782	6.81
0.58880	358.12	0.26208	−0.85234	0.14912	0.17775	0.06
0.01318	28.39	−0.01318	−0.07861	0.91704	0.54365	3.09
1.73436	1846.17	−0.54603	−2.05662	−0.18833	−0.17445	0.02
−0.12606	−227.42	0.12606	−0.38939	0.17176	−1.35441	8.98
0.99584	1579.03	0.06456	1.99545	−0.06040	−0.33750	2.31
−0.08089	−67.22	0.08089	−0.09970	−0.46359	1.22803	−2.47
0.65311	1329.09	0.38994	−0.08439	−0.04305	0.34268	−0.01
0.69957	1719.32	0.12913	−1.38787	0.17130	−0.08360	317.73
−0.01315	−43.84	0.01315	0.02929	1.02138	−0.27187	−117.97
−0.00218	38.47	0.00218	−0.18710	0.47301	−1.14371	−38.12

Rows 10, 11, and 12 of Gauss's matrix are shown above as rows 12, 10, and 11 respectively. Gauss dropped his row 10 because he regarded the experiment yielding that observation as suspect. In other words he took his data matrix to be the first 11 rows of $\mathbf{X}$ shown above. If these first 11 rows are called $\mathbf{X}_{(-)}$, Gauss first calculated $\mathbf{Q}_{(-)} = \mathbf{X}'_{(-)}\mathbf{X}_{(-)}$. He then followed out the first layer of the elimination method of Section 4.3.1 which gave the residual sum of squares 96364.0 and essentially the triangular matrix $\mathbf{B}$ where $\mathbf{V} = \mathbf{BV}^*$. By solving back he got the last row of $\mathbf{A}$, where $\mathbf{V}^* = \mathbf{AV}$, which yielded

$$\begin{aligned}
\delta\theta_1 &= -3.06\\
\delta\theta_2 &= 0.054335\\
\delta\theta_3 &= 166.44\\
\delta\theta_4 &= -4.29\\
\delta\theta_5 &= -34.37\\
\delta\theta_6 &= -3.15.
\end{aligned}$$

Such was the origin of the basic computing device of Section 4.3.1.

The present analysis was carried out along similar lines, but doing several more steps for illustrative purposes. Twelve additional columns were added to the data matrix $\mathbf{X}$ corresponding to the indicator variables of the 12 individuals, thus giving the enlarged 12×19 data matrix

$$\mathbf{X}^* = [\mathbf{X}, \mathbf{I}],$$

where $\mathbf{I}$ denotes the 12×12 identity matrix. From this the 19×19 raw sum inner product matrix

$$\mathbf{Q}^* = \mathbf{X}^{*\prime}\mathbf{X}^*$$

was found, for use as the basis for further calculations. Note that

$$\mathbf{Q}^* = \begin{bmatrix} \mathbf{Q} & \mathbf{X}' \\ \mathbf{X} & \mathbf{I} \end{bmatrix},$$

where $\mathbf{Q} = \mathbf{X}'\mathbf{X}$ is the usual 7×7 raw sum inner product matrix.

The next step was to calculate SWP[1, 2, 3, 4, 5, 6]$\mathbf{Q}^*$. In its upper left 7×7 part, this has SWP[1, 2, 3, 4, 5, 6]$\mathbf{Q}$ which yields the least squares analysis based on the full sample of twelve individuals. The (7, 7) element gives the residual sum of squares 85850.82 and the elements (7, 1), (7, 2), . . . , (7, 6) give

$$\begin{aligned}
\delta\theta_1 &= -15.46409\\
\delta\theta_2 &= 0.0539589\\
\delta\theta_3 &= 216.1136\\
\delta\theta_4 &= -32.56353\\
\delta\theta_5 &= -55.26174\\
\delta\theta_6 &= -2.952751.
\end{aligned}$$

The remaining elements (7, 8), (7, 9), . . . , (7, 19) of row 7 of

$$\text{SWP}[1, 2, 3, 4, 5, 6]\mathbf{Q}^*$$

provide the residuals when the predictor is applied to the 12 individuals, namely

$$\begin{array}{r}
124.4295\\
3.8682\\
-85.7873\\
54.3330\\
-32.6659\\
-15.1287\\
-20.8007\\
-42.8145\\
-150.0130\\
171.8944\\
-62.0554\\
-24.0310.
\end{array}$$

Note that the observation which Gauss rejected does not appear to have a suspicious residual.

In order to eliminate the twelfth individual from $\mathbf{Q}$, recall from (7.5.2) that SWP[19]$\mathbf{Q}^*$ yields the desired reduced raw sum inner product matrix $\mathbf{Q}_{(-)}$ as its upper left 7×7 part. Consequently the further operation SWP[19] applied to SWP[1, 2, 3, 4, 5, 6]$\mathbf{Q}^*$ yields the least squares analysis based on the first 11 individuals. The resulting residual sum of squares is 85094.14 and the weights are

$$\begin{aligned}
\delta\theta_1 &= -15.5884\\
\delta\theta_2 &= 0.053991\\
\delta\theta_3 &= 218.4079\\
\delta\theta_4 &= -33.09147\\
\delta\theta_5 &= -51.19588\\
\delta\theta_6 &= -7.698775.
\end{aligned}$$

Again the residuals were found to be

$$\begin{array}{r}
125.7156\\
9.0136\\
-86.5398\\
53.1740\\
-32.4062\\
-22.7582\\
-21.1797\\
-35.3474\\
-149.1134\\
169.8026\\
-67.5134\\
-31.4876.
\end{array}$$

Note that the last element here, which is the (7, 19) element of

$$\text{SWP}[1, 2, 3, 4, 5, 6, 19]\mathbf{Q}^*,$$

is the residual when the predictor based on the first 11 individuals is applied to the twelfth individual.

Since Gauss's calculations produced results different from those given here, both sets of calculations were checked. Gauss's numbers are self-consistent in that the original data matrix produces the raw sum of products matrix except for two small discrepancies. However, much larger errors begin to appear in Gauss's computed values at the first stage of elimination. The change in the outcome of the analysis when the twelfth individual is dropped is not striking.

It is of some interest to examine the residuals for individual i with $i = 1, 2, \ldots, 12$ based on the predictor derived from the 11 individuals excluding individual i. The residual -31.4876 above is the special case $i = 12$. The full column of these residuals may be found directly from SWP[1, 2, 3, 4, 5, 6]**Q*** by dividing element $(7 + i, 7)$ by element $(7 + i, 7 + i)$ for $i = 1, 2, \ldots, 12$. These are

$$\begin{array}{r} 438.739 \\ 5.340 \\ -138.768 \\ 103.534 \\ -710.193 \\ -24.196 \\ -133.545 \\ -58.661 \\ -265.824 \\ 428.794 \\ -110.159 \\ -31.488. \end{array}$$

Since each of these residuals concerns an individual not included in the associated predictor, its square is an estimate of the squared error expected for predictors based on a sample of size 11. It is strikingly clear that these residuals are much larger than those resulting when the regression hyperplane is fitted directly to the twelve individuals. In fact the sum of squares of these residuals is 1,016,384, which is nearly 12 times the sum of squares of deviations from the fitted regression hyperplane. See the further discussion of these residuals in Chapter 14.

8.5 CORRELATION WITH A SINGLE CATEGORICAL VARIABLE

This section illustrates the use of regression and correlation techniques where the variable V_p to be predicted is a dichotomous variable. For simplicity, V_p will be assumed scaled to take the values zero or one.

In such a situation, the usefulness of a multiple regression analysis might be questioned. In particular, it makes little sense to use a predictor $\hat{V}_p$ taking continuous values to predict V_p taking values zero or one only. Still, a formal analysis leading to a multiple correlation coefficient between V_p and its best linear predictor can shed light on observed data. A correlation coefficient calculated for a pair of variables, one dichotomous and the other continuous, is sometimes called in the psychological literature a *biserial correlation coefficient* (cf. McNemar, 1962).

Consider a sample of $n = n_1 + n_2$ individuals where V_p takes the value zero for n_1 individuals and unity for n_2 individuals. Suppose that $X^{(11)}, X^{(12)}, \ldots, X^{(1n_1)}$ and $X^{(21)}, X^{(22)}, \ldots, X^{(2n_2)}$ denote the sample observations on another variable V where the $X^{(1i)}$ correspond to individuals having the value zero on V_p and the $X^{(2i)}$ correspond to individuals having the value unity on V_p. It is easily seen that the sample corrected sum inner product matrix of V and V_p is given by

$$\left[\begin{array}{c:c} \begin{array}{c} \sum_{i=1}^{n_1}(X^{(1i)} - \bar{X}^{(1)})^2 + \sum_{i=1}^{n_2}(X^{(2i)} - \bar{X}^{(2)})^2 \\ + \dfrac{n_1 n_2}{n}(\bar{X}^{(1)} - \bar{X}^{(2)})^2 \end{array} & \dfrac{n_1 n_2}{n}(\bar{X}^{(1)} - \bar{X}^{(2)}) \\ \hdashline \dfrac{n_1 n_2}{n}(\bar{X}^{(1)} - \bar{X}^{(2)}) & \dfrac{n_1 n_2}{n} \end{array}\right], \quad (8.5.1)$$

where

$$\bar{X}^{(1)} = \frac{1}{n_1}\sum_{i=1}^{n_1} X^{(1i)} \quad \text{and} \quad \bar{X}^{(2)} = \frac{1}{n_2}\sum_{i=1}^{n_2} X^{(2i)}.$$

It follows that the point biserial correlation coefficient r between V and V_p may be expressed as

$$r = \frac{(n_1 n_2/n)^{1/2}(\bar{X}^{(1)} - \bar{X}^{(2)})}{[\sum_{i=1}^{n_1}(X^{(1i)} - \bar{X}^{(1)})^2 + \sum_{i=1}^{n_2}(X^{(2i)} - \bar{X}^{(2)})^2 + (n_1 n_2/n)(\bar{X}^{(1)} - \bar{X}^{(2)})^2]^{1/2}}. \quad (8.5.2)$$

The particular form (8.5.2) shows the relation of r to several other quantities familiar to statisticians. In *analysis of variance* terminology the quantity

$$G = \frac{(n_1 n_2/n)(\bar{X}^{(1)} - \bar{X}^{(2)})^2}{\sum_{i=1}^{n_1}(X^{(1i)} - \bar{X}^{(1)})^2 + \sum_{i=1}^{n_2}(X^{(2i)} - \bar{X}^{(2)})^2} \quad (8.5.3)$$

is called a ratio of a *between sample mean square* on 1 degree of freedom to a *pooled within sample sum of squares* on $n - 2$ degrees of freedom. The two samples are of course $X^{(11)}, X^{(12)}, \ldots, X^{(1n_1)}$ and $X^{(21)}, X^{(22)}, \ldots, X^{(2n_2)}$, and

G clearly depends only on their sample means and variances. Statistics equivalent to G are

$$F = \frac{(n_1 n_2/n)(\bar{X}^{(1)} - \bar{X}^{(2)})^2}{[1/(n-2)][\sum_{i=1}^{n_1}(X^{(1i)} - \bar{X}^{(1)})^2 + \sum_{i=1}^{n_2}(X^{(2i)} - \bar{X}^{(2)})^2]} \tag{8.5.4}$$

where the denominator sum of squares has been reduced to a *mean square*, and

$$D = \frac{\bar{X}^{(1)} - \bar{X}^{(2)}}{\{[1/(n-2)][\sum_{i=1}^{n_1}(X^{(1i)} - \bar{X}^{(1)})^2 + \sum_{i=1}^{n_2}(X^{(2i)} - \bar{X}^{(2)})^2]\}^{1/2}} \tag{8.5.5}$$

which represents the ratio of the difference of the sample means to a pooled *root mean square.*

From these definitions it follows easily that

$$\frac{r^2}{1 - r^2} = G = \frac{1}{n-2} F = \frac{n_1 n_2}{n(n-2)} D^2 \tag{8.5.6}$$

so that any one of r^2, G, F, or D^2 determines the other three. Also, r and D have the same sign and so determine each other.

The foregoing quantities were defined for the pair of variables V and V_p where V_p is dichotomous. In particular, they may be defined where V is chosen to be the best linear predictor $\dot{V}_p$ for V_p in terms of $V_1, V_2, \ldots, V_{p-1}$. Note that $\dot{V}_p$ has the property of maximizing each of r^2, G, F, and D^2 among all V in the subspace spanned by $V_1, V_2, \ldots, V_{p-1}$. This maximum D^2 will be discussed further from the viewpoint of two multivariate samples in Section 10.2. The case of a categorical variable with more than two categories will be fully examined in Section 9.4.

Example 8.4 following also illustrates the use of principal component analysis in connection with multiple correlation analysis. The question being asked of the data is essentially: does a certain dichotomous variable exhibit any correlation with a set of sample principal variables?

Example 8.4. The data of this example were collected by Dr. Gene Smith in connection with his studies of the measurement of personality. The 264 individuals were freshman nursing students of whom 219 successfully completed their year while the remaining 45 either chose to leave or were asked to leave during or at the end of their freshman year. Fifteen dropouts were not included in the study since they left either to marry or for health or financial reasons.

Two batteries of personality tests were given to these 264 students before they entered nursing school. These resulted in 15 variables from the Edwards Personal Preference Schedule (Edwards, 1959) and 16 variables representing personality factors (Cattell, Saunders, and Stice, 1957). Two further variables measuring reading and verbal ability were also provided in the data. Suppose that V_1 and V_2 denote the reading and verbal ability variables, that $V_3, V_4, \ldots, V_{\overline{17}}$ denote the Edwards variables, that $V_{\overline{18}}, V_{\overline{19}}, \ldots, V_{\overline{33}}$ denote the Cattel variables, and that $V_{\overline{34}}$ denotes the dichotomous variable where *pass* is represented by a zero score and *fail* is represented by a score of unity.

Since the scales of measurement here have no absolute meaning, it is convenient to work as much as possible in terms of standardized variables, i.e., variables scaled to have unit variances. The corresponding covariances become correlation coefficients. Also, it was decided to remove from consideration all components of these variables which are correlated with V_1 and V_2 in order to escape at least partially the criticism that any correlation between success and personality is due to the common influence of ability on both scores.

Accordingly, the first step was to form the corrected sum inner product matrix of the variables $V_1, V_2, \ldots, V_{\overline{34}}$ which was then standardized to a correlation matrix. This correlation matrix was next subjected to SWP[1, 2] to remove components along V_1 and V_2. The remaining 32 × 32 inner product matrix was then standardized again to yield the correlation matrix of $V_{3.12}, V_{4.12}, \ldots, V_{\overline{34}.12}$.

The first standardization to a correlation matrix was strictly unnecessary, although it did help to show how the ability variables V_1 and V_2 are correlated with the rest. For example, it appeared that $r^2_{\overline{34}(12)} = 0.0477$, which is small but greater than could reasonably be attributed to sampling fluctuations in a sample of size 264 drawn from a normal population whose corresponding $\rho_{\overline{34}(12)} = 0$. (By chance one would have expected roughly $r^2_{\overline{34}(12)} \approx \frac{2}{263} = 0.0076$. See Section 14.2 for a discussion of the sampling distribution.) On the other hand, the correlations between V_1 or V_2 and the personality measures present much more nearly the aspect of chance fluctuations with the possible exception of the following subtable of correlation coefficients.

	V_5	V_{10}	V_{15}	V_{19}
V_1	−0.185	0.178	−0.189	0.330
V_2	−0.140	0.163	−0.170	0.297

A simple correlation of $\sqrt{1/263} = 0.062$ is a rough guide to the sampling fluctuation expected here. A high correlation with one of V_1 or V_2 is usually accompanied by a high correlation with the other. This is partly a reflection of the initial high correlation $r_{12} = 0.594$, but is also due to the fact that all of these variables measure somewhat similar attributes.

The main part of the analysis started from the 32 × 32 correlation matrix $V_{3.12}, V_{4.12}, \ldots, V_{\overline{34}.12}$. Denote this matrix by $\mathbf{R}$. The pair $[\mathbf{R}, \mathbf{I}]$ was subjected to the eigenvalue and eigenvector operations SDG[1, 2, 3, . . . , 15], followed by operations SDG[16, 17, . . . , 31] as defined in (5.4.21) in order to carry out principal component analysis on $V_{3.12}, V_{4.12}, \ldots, V_{\overline{17}.12}$ and $V_{\overline{18}.12}, V_{\overline{19}.12}, \ldots, V_{\overline{33}.12}$. Denote the outcome of these operations by $[\mathbf{Q}, \mathbf{K}]$.

Then $\mathbf{K}$ has the form

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{11} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_{22} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & 1 \end{bmatrix},$$

representing a partition of rows and columns into 15 + 16 + 1 where the off-diagonal matrices consist entirely of zeros. The rows of $\mathbf{K}_{11}$ express a set of principal variables $U_1, U_2, \ldots, U_{\overline{15}}$ in terms of $V^*_{3.12}, V^*_{4.12}, \ldots, V^*_{\overline{17}.12}$ and the rows of $\mathbf{K}_{22}$ express a set of principal variables $W_1, W_2, \ldots, W_{16}$ in terms of $V^*_{\overline{18}.12}, V^*_{\overline{19}.12}, \ldots, V^*_{\overline{33}.12}$ where $V^*_{i.12}$ denotes the standardized $V_{i.12}$.

$\mathbf{Q}$ provides the sample corrected sum inner product matrix of $U_1, U_2, \ldots, U_{\overline{15}}, W_1, W_2, \ldots, W_{\overline{16}}, V^*_{\overline{34}.12}$. It partitions into

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} & \mathbf{Q}_{13} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} & \mathbf{Q}_{23} \\ \mathbf{Q}_{31} & \mathbf{Q}_{32} & 1 \end{bmatrix},$$

where $\mathbf{Q}_{11}$ is a diagonal matrix of eigenvalues whose elements:

3.33, 1.86, 1.38, 1.19, 1.09, 0.91, 0.87
0.81, 0.74, 0.62, 0.60, 0.49, 0.48, 0.42, 0.001

are the principal components of variance for a principal component analysis of the reduced Edwards variables $V_{3.12}, V_{4.12}, \ldots, V_{\overline{17}.12}$. Similarly, $\mathbf{Q}_{22}$ is the diagonal matrix of eigenvalues:

2.93 1.76 1.35 1.22 1.10 1.00 0.88 0.85
0.80 0.74 0.68 0.67 0.59 0.47 0.40 0.32

representing a set of principal components from a principal component analysis of the reduced Cattell variables $V_{\overline{18}.12}, V_{\overline{19}.12}, \ldots, V_{\overline{33}.12}$.

The matrix $\mathbf{Q}$ was then standardized to provide the correlation matrix of $U_1, U_2, \ldots, U_{\overline{15}}, W_1, W_2, \ldots, W_{\overline{16}}, V_{\overline{34}.12}$. This is a convenient form for inspecting the three off-diagonal blocks of correlation coefficients which interrelate the Edwards variables, the Cattell variables, and the pass-fail variable.

Now the reason for the principal component analysis in the first place was to simplify the inspection of the overly large correlation matrix $\mathbf{R}$ by restricting consideration to the first few principal variables of each kind. *Visual inspection of the correlation matrix of $V_1, V_2, \ldots, V_{\overline{15}}, W_1, W_2, \ldots, W_{\overline{16}}, V_{\overline{34}.12}$ does indeed show a concentration of meaningful-appearing correlations in the restricted* *correlation matrix of $U_1, U_2, U_3, U_4, W_1, W_2, W_3, W_4, V_{\overline{34}.12}$ which is reproduced below.*

1.000	0	0	0	0.350	0.363	−0.185	−0.021	−0.155
0	1.000	0	0	0.203	−0.059	0.009	−0.116	0.005
0	0	1.000	0	−0.362	0.216	−0.113	0.102	0.076
0	0	0	1.000	0.113	−0.061	−0.049	−0.030	0.129
0.350	0.203	−0.362	0.113	1.000	0	0	0	−0.171
0.363	−0.059	0.216	−0.061	0	1.000	0	0	−0.091
−0.185	0.009	−0.113	−0.049	0	0	1.000	0	−0.056
−0.021	−0.116	0.102	−0.030	0	0	0	1.000	−0.136
−0.155	0.005	0.076	0.129	−0.171	−0.091	−0.056	−0.136	1

Again it may be remembered that 0.062 is a rough guide for a typical meaningless r value. There do appear to be meaningful relations among the two sets of principal variables. Still, these relationships are weak. The reader should always remember that weak relationships can be clearly demonstrated with large samples even though the relationships have little or no practical value for subsequent prediction. Even weaker are the relations between the pass-fail variable and the personality variables.

The foregoing example has some interest for a statistical theoretician because it demonstrates empirically that nontrivial correlation effects can be concentrated by means of a principal component analysis. In this way, attempts at a better theoretical understanding of principal component analysis may be encouraged.

No attempt has been made to draw conclusions for psychology from the example. In particular, the psychologist usually assigns suggestive names to his variables to give "meaning" to his analysis, and the discussion here does not attempt to penetrate this name-meaning approach to interpretations. The data are disappointing in that so little prediction capability appears to reside in the personality measures. Nor, perhaps, is the small apparent capability any more than might be expected from the leakage of ability measures into intended personality measures.

CHAPTER 9

ONE SAMPLE OF INDIVIDUALS: EXTENSIONS OF MULTIPLE REGRESSION ANALYSIS

9.1 JOINT PREDICTION OF A SET OF VARIABLES

Rather than predict a single variable from a set of predictor variables, it may be required to predict a battery of variables from a common set of predictor variables, the prediction method being based on a sample observed on all of the variables. Each member of the battery may be predicted separately, of course, by a multiple regression analysis, and the methods discussed in this chapter are essentially based on such an approach. Section 9.1 makes some introductory remarks on joint prediction.

Suppose that the predictor variables are denoted by $V_1, V_2, \ldots, V_s$ and that the variables to be predicted are denoted by $V_{s+1}, V_{s+2}, \ldots, V_p$. Suppose that a sample of size n on all p variables yields a $p \times p$ sample covariance matrix $\mathbf{S}$. Then the best linear predictors $\dot{V}_{s+1}, \dot{V}_{s+2}, \ldots, \dot{V}_p$ of $V_{s+1}, V_{s+2}, \ldots, V_p$ in terms of $V_1, V_2, \ldots, V_s$ may be described as the orthogonal projections of $V_{s+1}, V_{s+2}, \ldots, V_p$ into the subspace $\mathscr{E}_1$ spanned by $V_1, V_2, \ldots, V_s$ in Euclidean variable-space $\mathscr{E}$, where the inner product over $\mathscr{E}$ is the sample covariance.

In order to understand fully the extension of multiple regression analysis to the case of $p - s$ predicted variables, it is important to notice the following result. *If $\dot{V}_{s+1}, \dot{V}_{s+2}, \ldots, \dot{V}_p$ are the best linear predictors for $V_{s+1}, V_{s+2}, \ldots, V_p$ each in terms of $V_1, V_2, \ldots, V_s$, then the best linear predictor for $\alpha_{s+1}V_{s+1} + \alpha_{s+2}V_{s+2} + \cdots + \alpha_pV_p$ in terms of $V_1, V_2, \ldots, V_s$ is $\alpha_{s+1}\dot{V}_{s+1} + \alpha_{s+2}\dot{V}_{s+2} + \cdots + \alpha_p\dot{V}_p$ where $\alpha_{s+1}, \alpha_{s+2}, \ldots, \alpha_p$ are arbitrary real numbers.* The proof of this is trivial because the best linear predictors involved are all determined by a linear transformation, namely orthogonal projection into $\mathscr{E}_1$. Every linear transformation has by definition the property that the transform of a linear combination of vectors is the same linear combination of their transforms.

Suppose that the corrected sum inner product matrix $\mathbf{T}$ has been computed.

Then

$$\mathrm{SWP}[1, 2, \ldots, s]\mathbf{T} = \begin{bmatrix} -\mathbf{T}_{11}^{-1} & \mathbf{H}_{12} \\ \mathbf{H}_{21} & \mathbf{T}_{22.1} \end{bmatrix} \tag{9.1.1}$$

provides the multiple regression analyses of each of $V_{s+1}, V_{s+2}, \ldots, V_p$ on $V_1, V_2, \ldots, V_s$. The rows of $\mathbf{H}_{21}$ define the best linear predictors $\dot{V}_{s+1}, \dot{V}_{s+2}, \ldots, \dot{V}_p$ for $V_{s+1}, V_{s+2}, \ldots, V_p$ in terms of $V_1, V_2, \ldots, V_s$. The matrix $\mathbf{T}_{22.1}$ is the corrected sum inner product matrix of $V_t - \dot{V}_t = V_{t.12\ldots s}$ for $t = s + 1, s + 2, \ldots, p$.

In matrix terms, defining $\mathbf{V}_1 = [V_1, V_2, \ldots, V_s]'$, $\mathbf{V}_2 = [V_{s+1}, V_{s+2}, \ldots, V_p]'$ and $\dot{\mathbf{V}}_2 = [\dot{V}_{s+1}, \dot{V}_{s+2}, \ldots, \dot{V}_p]'$, one has

$$\dot{\mathbf{V}}_2 = \mathbf{H}_{21}\mathbf{V}_1, \tag{9.1.2}$$

and $\mathbf{T}_{22.1}$ is the inner product matrix of $\mathbf{V}_2 - \dot{\mathbf{V}}_2$. In addition, from the result italicized above, the best linear predictor of

$$\boldsymbol{\alpha}_2\mathbf{V}_2 = \alpha_{s+1}V_{s+1} + \alpha_{s+2}V_{s+2} + \cdots + \alpha_pV_p \tag{9.1.3}$$

is given by

$$\boldsymbol{\alpha}_2\dot{\mathbf{V}}_2 = (\boldsymbol{\alpha}_2\mathbf{H}_{21})\mathbf{V}_1, \tag{9.1.4}$$

and the corresponding residual sum of squares is $\boldsymbol{\alpha}_2\mathbf{T}_{22.1}\boldsymbol{\alpha}_2'$.

It should be remembered of course that the use of these predictors also requires knowledge of the sample mean vector, where such means are easily computed along with $\mathbf{T}$ from (7.5.4). Indeed the augmented predictors may be computed from $\mathrm{SWP}[1, 2, \ldots, s]\mathbf{Q}_{(+)}$ which includes (9.1.1) as a submatrix.

9.2 CANONICAL CORRELATION ANALYSIS

The analysis of Section 9.1 may be described as the determination of variables in the space $\mathscr{E}_1$ spanned by $\mathbf{V}_1 = [V_1, V_2, \ldots, V_s]'$ which are best linear predictors for variables in the space $\mathscr{E}_2$ spanned by $\mathbf{V}_2 = [V_{s+1}, V_{s+2}, \ldots, V_p]'$. A natural extension is the determination of that variable in $\mathscr{E}_2$ which is "most predictable" in terms of a variable in $\mathscr{E}_1$. The question here actually involves $\mathscr{E}_1$ and $\mathscr{E}_2$ symmetrically in that it seeks that pair of variables, one in $\mathscr{E}_1$ and one in $\mathscr{E}_2$, having maximal sample correlation coefficient. This question leads directly to the method of canonical correlation analysis proposed by Hotelling (1935).

The required theory has already been derived in Section 5.6 in terms of the relationships between a pair of subspaces of a Euclidean space, the subspaces here being $\mathscr{E}_1$ and $\mathscr{E}_2$ in variable-space $\mathscr{E}$ and the inner product being sample covariance. Using the theory of eigenvalues and eigenvectors relating a pair of inner products, one determines special orthogonal bases $\mathbf{W}_1 = [W_1, W_2, \ldots, W_s]'$ for $\mathscr{E}_1$ and $\mathbf{W}_2 = [W_{s+1}, W_{s+2}, \ldots, W_p]'$ for $\mathscr{E}_2$ with the property that

$$\operatorname{cov}(W_i, W_j) = 0 \tag{9.2.1}$$

for $i = 1, 2, \ldots, s$ and $j = s + 1, s + 2, \ldots, p$, except when $j = s + i$. In other words, only the pairs (W_i, W_{s+i}) for $i = 1, 2, \ldots, \min(s, p - s)$ may make angles θ_i different from $\pi/2$ among all pairs with one member in $\mathbf{W}_1$ and one member in $\mathbf{W}_2$. The scaling may be chosen and the subscripts arranged so that

$$0 \leq \theta_1 \leq \theta_2 \leq \cdots \leq \theta_{\min(s,p-s)}. \tag{9.2.2}$$

The corresponding correlation coefficients

$$r_i = \cos\theta_i = \operatorname{cor}(W_i, W_{s+i}), \tag{9.2.3}$$

$i = 1, 2, \ldots, \min(s, p - s)$ satisfy

$$1 \geq r_1 \geq r_2 \cdots \geq r_{\min(p,p-s)} \geq 0. \tag{9.2.4}$$

The pair W_i, W_{s+i} will be called *the ith pair of canonical variables* and the corresponding r_i will be called *the ith canonical correlation coefficient.*

It is clear that the pair W_1, W_{s+1} satisfies the original "most predictable" criterion of Hotelling. With this pair fixed, the pair W_2, W_{s+2} is the most predictable pair where W_2 is taken from the subspace of $\mathscr{E}_1$ orthogonal to W_1 and W_{s+2} is taken from the subspace of $\mathscr{E}_2$ orthogonal to W_{s+1}, and so on. The special covariance matrix of the basis $\mathbf{W} = [\mathbf{W}_1, \mathbf{W}_2]'$ implies that the best linear predictor in $\mathscr{E}_1$ for W_{s+i} in $\mathscr{E}_2$ is $[\operatorname{cov}(W_i, W_{s+i})/\operatorname{var}(W_i)]W_i$ and the best linear predictor in $\mathscr{E}_2$ for W_i in $\mathscr{E}_1$ is $[\operatorname{cov}(W_i, W_{s+i})/\operatorname{var}(W_{s+i})]W_{s+i}$ for $i = 1, 2, \ldots, \min[s, p - s]$. More generally, the best linear predictor in $\mathscr{E}_1$ for any variable $\sum_1^{p-s}\beta_{s+j}W_{s+j}$ in $\mathscr{E}_2$ is

$$\sum_{i=1}^{\min(s,p-s)} \beta_{s+i}\frac{\operatorname{cov}(W_i, W_{s+i})}{\operatorname{var}(W_{s+i})}W_i, \tag{9.2.5}$$

and the reader may easily supply the formula reversing the roles of $\mathscr{E}_1$ and $\mathscr{E}_2$. Note that $\mathbf{W}_1$ and $\mathbf{W}_2$ may be chosen to be orthonormal and in this case (9.2.5) takes the simple form

$$\sum_{i=1}^{\min(s,p-s)} \beta_{s+i}r_iW_i. \tag{9.2.6}$$

The canonical correlation coefficients are uniquely determined and the degree of uniqueness of the canonical variables may be deduced from Theorem 5.1.2. In general, with sample data the canonical correlation coefficients will all be distinct and the pairs of canonical variables will therefore be uniquely determined up to scale factors. The other case, which allows for sets of equal r_i, will not be described in detail. Note, however, that the shorter of the two bases $\mathbf{W}_1$ and $\mathbf{W}_2$ has a set of $|2s - p|$ variables at the end which are uncorrelated with all of the other canonical variables of either set and which may be replaced by any orthogonal basis of the subspace which they span.

Canonical correlation analysis has an obvious mathematical appeal; whether or not it is a statistically useful tool is less easily discovered. A similar question was raised when considering principal component analysis, which also involves eigenvalue and eigenvector analysis. The canonical correlation analysis method is less vague than the principal component method in the sense that it is free of an arbitrary choice of a reference inner product, but the question of meaning and usefulness of the artificial canonical variables remains. The hope is that in a many-variable situation the first few canonical variables will prove to be the important ones and thus provide a means for reducing the number of variables under consideration to more easily comprehensible dimensions.

The computations required for canonical correlation analysis afford good illustrations of the SWP, MST, and SDG operators. Consider an initial position given the sample corrected sum inner product matrix

$$\mathbf{T} = \begin{bmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{T}_{21} & \mathbf{T}_{22} \end{bmatrix} \tag{9.2.7}$$

of the basis $\mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2]'$, where the partitions refer as usual to $p = s + (p - s)$.

According to the theory of Section 5.6 the canonical variables $\mathbf{W}_2 = [W_{s+1}, W_{s+2}, \ldots, W_p]'$ are eigenvectors of an inner product π_1 relative to an inner product π_2. The roles of $\mathscr{U}$ and $\mathscr{V}$ in Section 5.6 are played here by $\mathscr{E}_2$ and $\mathscr{E}_1$, respectively. Using the sample corrected sum inner product to make $\mathscr{E}$ Euclidean, the inner product matrix relative to $\mathbf{V}_2$ for π_2 is $\mathbf{T}_{22}$. Since π_1 refers to the inner products among the components of $\mathbf{V}_2$ in $\mathscr{E}_1$, the inner product matrix relative to $\mathbf{V}_2$ for π_1 is $\mathbf{T}_{22} - \mathbf{T}_{22.1}$. The corresponding eigenvalues are interpreted in (5.6.1) as $\cos^2\theta$, or r^2 in correlation coefficient terms. It is computationally more convenient to deal with $\mathbf{T}_{22.1}$ rather than $\mathbf{T}_{22} - \mathbf{T}_{22.1}$, and thence to find the eigenvalues and eigenvectors of $\pi_2 - \pi_1$ relative to π_2. The eigenvectors are the same, of course, while the eigenvalues $\cos^2\theta$ or r^2 are replaced by $\sin^2\theta$ or $1 - r^2$.

The calculations proceed in three steps each of which may be regarded as a computer subroutine:

$$\operatorname{MST}[s + 1, s + 2, \ldots, p][\mathbf{T}, \mathbf{I}] = \left[\begin{bmatrix} \mathbf{T}_{11} & \dot{\mathbf{T}}_{12} \\ \dot{\mathbf{T}}_{21} & \mathbf{I} \end{bmatrix}, \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_{22} \end{bmatrix}\right], \tag{9.2.8}$$

$$\operatorname{SWP}[1, 2, \ldots, s]\begin{bmatrix} \mathbf{T}_{11} & \dot{\mathbf{T}}_{12} \\ \dot{\mathbf{T}}_{21} & \mathbf{I} \end{bmatrix} = \begin{bmatrix} -\mathbf{T}_{11}^{-1} & \mathbf{T}_{11.}^{-1}\ {}_{12} \\ \dot{\mathbf{T}}_{21}\mathbf{T}_{11}^{-1} & \end{bmatrix}, \tag{9.2.9}$$

and

$$\operatorname{SDG}[s + 1, s + 2, \ldots, p]\left[\begin{bmatrix} -\mathbf{T}_{11}^{-1} & \mathbf{T}_{11}^{-1}\dot{\mathbf{T}}_{12} \\ \dot{\mathbf{T}}_{21}\mathbf{T}_{11}^{-1} & \dot{\mathbf{I}} \end{bmatrix}, \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_{22} \end{bmatrix}\right]$$

$$= \left[\begin{bmatrix} -\mathbf{T}_{11}^{-1} & \mathbf{T}_{11}^{-1}\ddot{\mathbf{T}}_{12} \\ \ddot{\mathbf{T}}_{21}\mathbf{T}_{11}^{-1} & \ddot{\mathbf{I}} \end{bmatrix}, \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{22}\mathbf{K}_{22} \end{bmatrix}\right]. \tag{9.2.10}$$

The step (9.2.8) in effect replaces the basis $\mathbf{V}_2$ of $\mathscr{E}_2$ by the basis $\mathbf{U}_2 = \mathbf{K}_{22}\mathbf{V}_2$ whose π_2 inner product matrix is $\mathbf{I}$. The step (9.2.9) completes the preparation for the eigenvalue analysis by finding the $\pi_2 - \pi_1$ inner product matrix for the basis $\mathbf{U}_2$ which is denoted in (9.2.9) by $\dot{\mathbf{I}} = \mathbf{I} - \dot{\mathbf{T}}_{21}\mathbf{T}_{11}^{-1}\dot{\mathbf{T}}_{12}$. The third computing step (9.2.10) produces the diagonal matrix $\ddot{\mathbf{I}}$ whose diagonal elements are eigenvalues of the form $1 - r_i^2$ where r_i is a canonical correlation coefficient. The eigenvectors $(\mathbf{C}_{22}\mathbf{K}_{22})\mathbf{V}_2$ form a basis of canonical variables $\mathbf{W}_2$ in $\mathscr{E}_2$; a corresponding set of canonical variables in $\mathscr{E}_1$ is given by $(\ddot{\mathbf{T}}_{21}\mathbf{T}_{11}^{-1})\mathbf{V}_1$.

The only assertion here which is not obvious is that concerning the canonical variables in $\mathscr{E}_1$. From (9.2.9) it is clear that $(\dot{\mathbf{T}}_{21}\mathbf{T}_{11}^{-1})\mathbf{V}_1$ provides the best linear predictors for $\mathbf{U}_2$ in terms of $\mathbf{V}_1$, and it is easily checked that the operations (9.2.10) modify $\dot{\mathbf{T}}_{21}\mathbf{T}_{11}^{-1}$ into $\ddot{\mathbf{T}}_{21}\mathbf{T}_{11}^{-1}$ by row operations in such a way that $(\ddot{\mathbf{T}}_{21}\mathbf{T}_{11}^{-1})\mathbf{V}_1$ provides the best linear predictors for $\mathbf{W}_2$ in terms of $\mathbf{V}_1$. But, from (9.2.2), these best linear predictors are simply specially scaled versions of the corresponding canonical variables in $\mathscr{E}_1$, as claimed above.

Some details deserve further explanation. For convenience of notation suppose that the diagonal elements of $\ddot{\mathbf{I}}$ in (9.2.10) are arranged in *increasing* order. If these diagonal elements are denoted by $1 - r_1^2, 1 - r_2^2, \ldots, 1 - r_{p-s}^2$ then $r_1, r_2, \ldots, r_{\min(s,p-s)}$ are the ordered canonical correlation coefficients satisfying (9.2.4). If $s < p - s$, then only the first s of the diagonal elements $1 - r_1^2, 1 - r_2^2, \ldots, 1 - r_{p-s}^2$ correspond to nontrivial canonical correlation coefficients while the remaining $p - 2s$ are simply unity, corresponding to zero correlations. Also, if $s < p - s$ then only the first s elements of $(\ddot{\mathbf{T}}_{21}\mathbf{T}_{11}^{-1})\mathbf{V}_1$ are different from $\emptyset$ and these s non-$\emptyset$ elements define the basis $\mathbf{W}_1$ of canonical variables in $\mathscr{E}_1$. On the other hand, if $s \geq p - s$, there will in general be no unit-valued elements of $\ddot{\mathbf{I}}$, but only the first $p - s$ elements of $\mathbf{W}_1$ will be provided by $(\ddot{\mathbf{T}}_{21}\mathbf{T}_{11}^{-1})\mathbf{V}_1$. The remaining $2s - p$ elements of $\mathbf{W}_1$ may be chosen to form any orthogonal basis of the subspace of $\mathscr{E}_1$ orthogonal to $\mathscr{E}_2$, but they are not provided by the above computations.

It is of interest to understand the choice of scale implied by the given computations. Since the SDG operation modifies the $(\pi_2 - \pi_1)$-orthonormal basis $\mathbf{V}_2$ into another $(\pi_2 - \pi_1)$-orthonormal basis $\mathbf{W}_2$, and since π_2 is the sample corrected sum inner product, it follows that the canonical variables $\mathbf{W}_2 = (\mathbf{C}_{22}\mathbf{K}_{22})\mathbf{V}_2$ are scaled to have unit norm according to the sample corrected sum inner product. The scaling of the corresponding $(\ddot{\mathbf{T}}_{22}\mathbf{T}_{11}^{-1})\mathbf{V}_1$ may be deduced from their interpretation as best linear predictors for $\mathbf{W}_2$. In fact, if the ith element of $(\ddot{\mathbf{T}}_{21}\mathbf{T}_{11}^{-1})\mathbf{V}_1$ were to be rescaled by dividing by r_i, then it too would have unit norm according to the sample corrected sum inner product while in its given form it has norm r_i.

A final observation on the computations is that the roles of $\mathscr{E}_1$ and $\mathscr{E}_2$ may be interchanged. Thus, there is a choice between an s-dimensional or $(p - s)$-dimensional eigenvalue calculation, and for some purposes it may be better to choose the smaller dimension.

9.3 AN EXAMPLE ILLUSTRATING EXPLORATORY USE OF CANONICAL CORRELATION ANALYSIS

Example 9.1. The data analyzed here were supplied by Dr. Gene Smith. They consist of the scores of 221 nursing students on two sets of personality measures. The first set provides the 16 personality variables of Cattell, Saunders, and Stice (1957) which also appeared in Example 8.4, while the second set provides 31 variables devised by Dr. Smith. The purpose of analysis is to try to throw light on the nature and extent of the covariation between the two sets of variables.

The first analysis which comes to mind is to find the $(16 + 31) \times (16 + 31)$ sample correlation matrix $\mathbf{R}$ of the two sets of variables combined. Then the $16 \times 31 = 496$ correlation coefficients relating the two sets of variables may be examined with reference to the name-meanings of the individual variables to see if large correlations appear where expected. For example, variables intended to measure something like extroversion in the different sets should show large correlations. The signs of the correlation coefficients are also subject to interpretation. For example, a variable measuring extroversion should be positively correlated with another variable measuring extroversion but negatively correlated with a variable measuring introversion.

By and large, this type of analysis is extrastatistical and subjective in that it depends on the meaning which the psychologist attributes to his variables. It can be very satisfying to see sets of nontrivial correlation coefficients (e.g. in these data 0.3, 0.4, or occasionally 0.5) appearing in places where they seem to have natural interpretations. On the other hand the initial impact of such correlation coefficients may need re-assessment because they are often partial reflections of one another, i.e., if V_1 and U_1 are correlated, then one may expect V_2 and U_2 to be correlated if V_1 is correlated with V_2 and U_1 with U_2. Thus there remains a need for judging the significance of observed correlations. Strictly speaking, this should mean displaying all the sets of interrelated correlation coefficients which could have any psychologically sensible interpretation, and then trying to determine how many of these sets of possible interpretations are meaningfully supported by the sample data. Even more, there is a need to assess in quantitative terms the strength of those relations which are deemed meaningful.

All this is difficult if not impossible, at least in the current state of the art. What is usually done is to carry out a significance test of the null hypothesis of no correlation whatsoever between the two sets. If this null hypothesis is rejected, then the psychologist will feel that at least the worst did not happen and he will use his own judgment to make as many interpretations as he thinks the data will support. Such interpretations are then used to deepen understanding of what the given psychological measures are providing and to suggest new measuring instruments.

Analyses directly involving the name-meanings of the variables are not discussed in this book, which is not to say that they are unimportant, but only to admit that the science of statistics is not yet able to be of much assistance in that area. Instead, more statistical explorations are carried out which ignore the opportunities and difficulties of psychological interpretation.

The most obvious analysis is simply to find the sample canonical correlation coefficients and corresponding canonical variables. This was done beginning computationally from the 47 × 47 correlation matrix **R** whose direct interpretation was discussed above. Applying the operations (9.2.6), (9.2.7), and (9.2.8) to **R** instead of **T** means only that a standardized basis is used in place of the original basis. The canonical correlation coefficients are the same, but the canonical variables are expressed in terms of the standardized variables rather than the original variables. To save space only the set of $1 - r_i^2$ together with r_i and not the coefficients defining the canonical variables are reproduced in Table 9.1.1. Judged as single simple correlation coefficients these sample canonical correlation coefficients appear quite large. It should be remembered, however, that each may also be interpreted as a multiple correlation coefficient relating a canonical variable chosen from one set of variables with the other complete set of variables. A typical meaningless multiple r^2 with 31 variables is 31 times a typical meaningless simple r^2.

Table 9.1.1

$1 - r_i^2$ AND r_i FOR THE 16 PAIRS OF SAMPLE CANONICAL VARIABLES RELATING THE 16 CATTELL VARIABLES AND THE 31 SMITH VARIABLES

$1 - r_i^2$	r_i
0.4352	0.7515
0.5742	0.6525
0.6547	0.5876
0.7063	0.5420
0.7418	0.5081
0.7728	0.4767
0.7932	0.4547
0.8132	0.4322
0.8237	0.4199
0.8539	0.3822
0.8852	0.3389
0.8996	0.3169
0.9261	0.2719
0.9453	0.2339
0.9489	0.2283
0.9724	0.1662

An additional analysis was carried out on the same 47 × 47 correlation matrix **R**. To begin, a principal component analysis was carried out on the Cattell and Smith variables separately by computing

$$\mathrm{SDG}[1, 2, \ldots, 16]\mathrm{SDG}[17, 18, \ldots, 47][\mathbf{R}, \mathbf{I}] = \left[\begin{bmatrix}\mathbf{R}_{11}^* & \mathbf{R}_{12}^*\\ \mathbf{R}_{21}^* & \mathbf{R}_{22}^*\end{bmatrix}, \begin{bmatrix}\mathbf{K}_{11}^* & \mathbf{0}\\ \mathbf{0} & \mathbf{K}_{22}^*\end{bmatrix}\right],$$

where $\mathbf{R}_{11}^*$ is the diagonal matrix of principal components for the Cattell variables and $\mathbf{R}_{22}^*$ is the same for the Smith variables. $\mathbf{K}_{11}^*$ expresses the 16 Cattell principal variables in terms of the original standardized Cattell variables, and $\mathbf{K}_{22}^*$ does the same for the Smith variables. If, as would normally be the case, the original standardized variables were regarded as having unit *sample variance*, then the sample principal components are also *sample variances* for the corresponding principal variables and the elements of $\mathbf{R}_{12}^*$ are the *sample covariances* between the two sets of principal variables. The two sets of principal components are shown in Table 9.1.2. Note that the Cattell components drop off more slowly than the Smith components. For example, to achieve 95% of the total of 16.0000 requires 14 of the 16 Cattell variables, while to achieve 95% of the total 31.0000 requires only 16 of the 31 Smith variables. This suggests a greater redundancy in the Smith variables than in the Cattell variables.

Table 9.1.2

PRINCIPAL COMPONENTS OF THE 16 CATTELL VARIABLES AND THE 31 SMITH VARIABLES

Cattell components	Smith components	
3.7618	11.4583	0.1726
2.2444	6.5800	0.1478
1.7615	4.6448	0.1368
1.2207	2.1822	0.1343
1.0256	1.1778	0.1211
0.9461	0.6819	0.1000
0.8869	0.4348	0.0973
0.7499	0.3997	0.0937
0.5944	0.3646	0.0918
0.5379	0.3164	0.0845
0.4991	0.2415	0.0783
0.4475	0.2191	0.0705
0.4249	0.2091	0.0671
0.3638	0.2079	0.0553
0.2945	0.1915	0.0551
0.2410	0.1834	
Total 16.0000		Total 31.0000

The next stage of analysis is to reduce the sample covariance matrix of the principal variables to a correlation matrix, i.e., to alter

$$\begin{bmatrix} \mathbf{R}_{11}^* & \mathbf{R}_{12}^* \\ \mathbf{R}_{21}^* & \mathbf{R}_{22}^* \end{bmatrix} \quad \text{to} \quad \begin{bmatrix} \mathbf{I} & \mathbf{R}_{12}^{**} \\ \mathbf{R}_{21}^{**} & \mathbf{I} \end{bmatrix}$$

by dividing through each row and column by the square root of its diagonal elements. The 16 × 31 matrix $\mathbf{R}_{12}^{**}$ provides the set of 496 correlation coefficients between the two sets of principal variables. In accordance with the hope that only the first few principal variables of each set contain important variation, these correlation coefficients were looked at in two groups, the first group of 25 being the correlation coefficients relating the first 5 principal variables of each kind, and the second group being the remaining 471 = 496 − 25 correlation coefficients. The two distributions of the absolute values of these correlation coefficients corresponding to the set of 25 and the set of 471 are shown in Table 9.1.3. Two points should be kept in mind when considering Table 9.1.3. The first is that these are correlations among two sets of variables where correlations within each set are all zero. Consequently, no correlation in the table is contaminated by another in the sense of relating intracorrelated pairs of variables. The second point is that although an expected frequency is given for each observed frequency under the hypothesis of no correlation at all between Cattell and Smith variables, it is obviously not appropriate to compare the observed and expected frequencies by a χ^2 goodness-of-fit test because the observed correlation coefficients are not independently drawn from a population. Still the agreement between observed and expected on the left side is quite striking except for the two most extreme categories. On the right side the last 4 categories considerably exceed their null expectations. On the left side the three values exceeding 0.2 are 0.243 between the first Smith principal variable and the fifteenth Cattell principal variable, 0.227 between the twenty-second Smith and the eleventh Cattell, and 0.205 between the thirty-first Smith and the fifth Cattell. It is difficult to believe that these three correlations mean anything more specific than that some small residual tendency for correlation remains after the 25 more promising elements have been removed. The four values exceeding 0.2 on the right are 0.227, 0.296, 0.360, and 0.557 as may be seen from Table 9.1.4.

Table 9.1.3

THE DISTRIBUTION OF THE 25 CORRELATION COEFFICIENTS AMONG THE FIRST 5 PRINCIPAL VARIABLES OF EACH SET AND OF THE REMAINING 471 CORRELATION COEFFICIENTS AMONG PRINCIPAL VARIABLES. THE EXPECTED FREQUENCIES ARE CALCULATED ACCORDING TO THE $\beta(\frac{1}{2}, \frac{219}{2})$ DISTRIBUTION FOR r^2 (see Section 14.2.)

	Group of 471		Group of 25	
$\lvert r \rvert$	Observed frequency	Null expected frequency	Observed frequency	Null expected frequency
0–0.025	127	136.6	4	7.3
0.025–0.050	126	119.2	5	6.3
0.050–0.075	77	90.9	2	4.8
0.075–0.100	67	60.8	3	3.2
0.100–0.125	34	34.8	0	1.9
0.125–0.150	22	17.4	4	0.9
0.150–0.175	6	7.5	2	0.4
0.175–0.200	9	2.7	1	0.1
0.200+	3	1.1	4	0.1
	471	471.0	25	25.0

Table 9.1.4

THE SAMPLE CORRELATION COEFFICIENTS AMONG THE FIRST FIVE PRINCIPAL VARIABLES OF EACH KIND

		Smith principal variables				
		1	2	3	4	5
Cattell principal variables	1	−0.227	−0.175	−0.041	0.017	0.138
	2	−0.296	0.557	0.087	−0.061	0.086
	3	0.140	0.040	−0.170	−0.134	0.021
	4	−0.042	−0.058	0.043	0.176	0.013
	5	0.005	0.029	−0.079	0.360	0.128

The fact that the largest value 0.557 relates the second principal variables in each set is striking evidence that variables pulled out by a principal component analysis have a tendency to be good variables for prediction purposes as well. The other large value 0.360 shows that this tendency persists down to the fourth and fifth principal variables.

Next reconsider canonical correlation analysis. If the analysis applied to **R** were applied to

$$\begin{bmatrix} \mathbf{I} & \mathbf{R}_{12}^* \\ \mathbf{R}_{21}^* & \mathbf{I} \end{bmatrix}$$

instead, then exactly the same canonical correlation coefficients as those in Table 9.1.1 would result, for these quantities are not dependent on the choice of basis in $\mathscr{E}_1$ or $\mathscr{E}_2$. It is instructive to note how an array like $\mathbf{R}_{12}^*$ which Table 9.1.3 shows to be so close to a null distribution can produce canonical correlations as healthy-appearing as those in Table 9.1.1. This in turn suggests doing the canonical correlation analysis on smaller subsets of principal variables in the hope of

Table 9.1.5

CANONICAL CORRELATION ANALYSES BASED ON TABLE 9.1.4

	$1 - r_i^2$	r_i
First 5 Cattell principal variables versus first 5 Smith principal variables.	0.5699	0.6558
	0.8094	0.4366
	0.8783	0.3489
	0.9658	0.1819
	0.9969	0.0559
First 2 Cattell principal variables versus first 2 Smith principal variables.	0.6028	0.6303
	0.9209	0.2812

getting more meaningful sample canonical correlation coefficients. Two such reduced canonical correlation analyses were done, the first relating the first five principal variables from each set and the second relating the first two principal variables in each set. The results are shown in Table 9.1.5.

9.4 THE CASE OF A SINGLE CLASSIFICATION VARIABLE REPRESENTED BY $V_{s+1}, V_{s+2}, \ldots, V_p$

A *classification variable with $p - s$ classes* is defined by a rule such that each individual is assigned to one of $p - s$ classes. Such a variable could be reduced to a standard real-valued variable by assigning a real value to each of the $p - s$ classes, but information is lost by doing so. A means of retaining all of the sample information is to set up an indicator variable for each of the $p - s$ classes; more specifically, to define V_{s+i} to be a variable taking the value unity for a variable in class i and zero otherwise, for $i = 1, 2, \ldots, p - s$. The discussion of this section follows out special cases of the analyses described earlier in this chapter which arise when such a classification variable is related to s measured variables $V_1, V_2, \ldots, V_s$. The still more special case of $p - s = 2$ was introduced in Section 8.5. It will be seen that the notation and terminology of multivariate analysis of variance is helpful in the present context.

A few remarks about classification variables may help clarity and perspective. A sample of male children of a certain age might be measured on height, weight, and eye color, where eye color is reported in one of five categories. Such a sample belongs to the type considered here where V_1 and V_2 refer to height and weight while V_3, V_4, V_5, V_6, and V_7 denote indicator variables for eye color. If a second categorical variable were adjoined, say hair color classified into four categories, then the pair of classification variables together define a cross classification into $4 \times 5 = 20$ categories, and again the present analyses could be carried out. Note, however, that the standard linear and quadratic analyses are not suggested here for analyzing covariation between such a pair of categorical variables. Nor are cross-classifications explicitly discussed except briefly in Example 10.3.

The hypothetical sample on height, weight, and eye color may be compared with another hypothetical situation providing height and weight for samples of male children of a given age from five different nationalities. The classification in the latter case would normally be conceptualized rather differently: rather than regard nationality as one of many varying characteristics in a unitary population, one would regard the nationality groups as five different populations from which five different samples were available. The line between these two attitudes is not always firmly fixed. The former is adopted in this chapter, but the terminology of analysis of variance developed for the latter attitude will be introduced. Chapter 10 illustrates the separate population attitude, while Chapter 11 illustrates a class of situations where it is appropriate to regard the data as representing a single sample on some variables and several samples on other variables.

Suppose that n_i sample individuals fall in category i for $i = 1, 2, \ldots, p - s$ where $\sum_1^{p-s} n_i = n$. Suppose that the scores on V_r for the n_i individuals in category i are denoted by $X_r^{(i,j)}$ for $j = 1, 2, \ldots, n_i$, $i = 1, 2, \ldots, p - s$, and $r = 1, 2, \ldots, s$. The n sample individuals may be ordered so that the data matrix $\mathbf{X}$ takes the form

$$\mathbf{X} = \begin{bmatrix} X_1^{(1,1)} & \cdots & X_s^{(1,1)} & 1 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ X_1^{(1,n_1)} & \cdots & X_s^{(1,n_1)} & 1 & 0 & \cdots & 0 \\ X_1^{(2,1)} & \cdots & X_s^{(2,1)} & 0 & 1 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ X_1^{(2,n_2)} & \cdots & X_s^{(2,n_2)} & 0 & 1 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ X_1^{(p-s,n_{p-s})} & \cdots & X_s^{(p-s,n_{p-s})} & 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{9.4.1}$$

The sample means and inner products for $V_1, V_2, \ldots, V_s$ have the general forms given in Section 7.2 while the remaining means and inner products have the special forms

$$\begin{aligned} \bar{X}_{s+j} &= n_j/n, \\ (V_i, V_{s+j})_Q &= n_j\bar{X}_i^{(j)}, \\ (V_i, V_{s+j})_T &= n_j(\bar{X}_i^{(j)} - \bar{X}_i), \\ (V_{s+j}, V_{s+k})_Q &= 0, \\ (V_{s+j}, V_{s+k})_T &= -n_jn_k/n, \\ (V_{s+j}, V_{s+j})_Q &= n_j, \end{aligned} \tag{9.4.2}$$

and

$$(V_{s+j}, V_{s+j})_T = n_j(n - n_j)/n,$$

for $i = 1, 2, \ldots, s$, and $j \neq k = 1, 2, \ldots, p - s$, where

$$\bar{X}_i^{(j)} = (1/n_j)\sum_{l=1}^{n_j} X_i^{(j,l)}. \tag{9.4.3}$$

Best linear predictors may be sought either for the V_{s+j} in terms of V_1, $V_2, \ldots, V_s$ or for the V_i in terms of $V_{s+1}, V_{s+2}, \ldots, V_p$. The latter will be explored here. Such multiple regression analyses may be visualized geometrically in the n-dimensional space $\mathscr{N}$ introduced in Section 7.4 where every variable V is represented by a point $P(V)$. The regression analysis of V_i on $V_{s+1}, V_{s+2}, \ldots, V_p$ is clearly related to the orthogonal projection of $P(V_i)$ into the subspace spanned by $P(V_0), P(V_{s+1}), \ldots, P(V_p)$. Usually this subspace would have dimension $p - s + 1$, but here V_0 and $V_{s+1} + V_{s+2} + \cdots + V_p$ are both variables which take the value unity for all individuals; consequently, $P(V_0)$ lies in the subspace spanned by $P(V_{s+1}), P(V_{s+2}), \ldots, P(V_p)$ which has dimension $p - s$. Thus one of the variables $V_0, V_{s+1}, \ldots, V_p$ is redundant.

In computing terms, the desired orthogonal projection is carried out via any $p - s$ of the $p - s + 1$ operations SWP[$s + 1$], SWP[$s + 2$], . . . , SWP[p], SWP[$p + 1$] applied to $\mathbf{Q}_{(+)}$. (The remaining sweeping operation cannot be carried out because there is a zero in the corresponding diagonal position.) By carrying out the first $p - s$ operations one finds that the augmented best linear predictor for V_i is

$$\hat{V}_i = \sum_{j=1}^{p-s} \bar{X}_i^{(j)} V_{s+j} \tag{9.4.4}$$

for $i = 1, 2, \ldots, s$. But of course substitutions may be made by identifying V_0 with $V_{s+1} + \cdots + V_p$ to obtain equivalent expressions.

Similarly the (nonaugmented) best linear predictor may be determined by orthogonal projection of $P_{II}(V_i)$ into the subspace spanned by $P_{II}(V_{s+1})$, $P_{II}(V_{s+2}), \ldots, P_{II}(V_p)$ which has dimension $p - s - 1$ since $P_{II}(V_{s+1} + V_{s+2} + \cdots + V_p) = \emptyset$ in $\mathscr{N}$. The reader may check that the best linear predictor may also be expressed by (9.4.4).

The fitted part of the sample corrected sum norm of V_i may be written

$$(\dot{V}_i, \dot{V}_i)_T = \sum_{j=1}^{p-s} n_j(\bar{X}_i^{(j)} - \bar{X}_i)^2, \tag{9.4.5}$$

and the residual part may be written

$$(V_i - \dot{V}_i, V_i - \dot{V}_i)_T = \sum_{j=1}^{p-s}\sum_{l=1}^{n_j}(X_i^{(j,l)} - \bar{X}_i^{(j)})^2, \tag{9.4.6}$$

so that the decomposition $(V_i, V_i)_T = (\dot{V}_i, \dot{V}_i)_T + (V_i - \dot{V}_i, V_i - \dot{V}_i)_T$ is expressed here by the identity

$$\sum_{j=1}^{p-s}\sum_{l=1}^{n_j}(X_i^{(j,l)} - \bar{X}_i)^2 = \sum_{j=1}^{p-s} n_j(\bar{X}_i^{(j)} - \bar{X}_i)^2 + \sum_{j=1}^{p-s}\sum_{l=1}^{n_j}(X_i^{(j,l)} - \bar{X}_i^{(j)})^2. \tag{9.4.7}$$

More generally, in the spirit of Section 9.1 consider the joint prediction of each of $V_1, V_2, \ldots, V_s$ in terms of $V_{s+1}, V_{s+2}, \ldots, V_p$. The sample corrected sum inner product of $V_1, V_2, \ldots, V_s$ decomposes accordingly into the sum of an inner product associated with the fitted parts of $V_1, V_2, \ldots, V_s$ and an inner product associated with the residual parts of $V_1, V_2, \ldots, V_s$. The (i, h) elements of the matrices of the latter two inner products are respectively

$$(V_i, V_h)_A = \sum_{j=1}^{p-s} n_j(\bar{X}_i^{(j)} - \bar{X}_i)(\bar{X}_h^{(j)} - \bar{X}_h), \tag{9.4.8}$$

and

$$(V_i, V_h)_W = \sum_{j=1}^{p-s}\sum_{l=1}^{n_j}(X_i^{(j,l)} - \bar{X}_i^{(j)})(X_h^{(j,l)} - \bar{X}_h^{(j)}), \tag{9.4.9}$$

where

$$(V_i, V_h)_T = (V_i, V_h)_A + (V_i, V_h)_W \tag{9.4.10}$$

for i and $h = 1, 2, \ldots, s$.

The quantities appearing above are all familiar to users of *analysis of variance* ideas. The particular analysis of variance under consideration is the simple case of *a one-way classification into* $p - s$ *groups*. Expression (9.4.5) may be called the *among group sum of squares* for V_i while the term (*pooled*) *within group sum of squares* is used for expression (9.4.6). Considering V_1, $V_2, \ldots, V_s$ jointly, rather than just a single V_i, the terminology and concepts of analysis of variance are generalized to *multivariate analysis of variance*. Thus the among group sum of squares is replaced by the *among group sum inner product* defined by (9.4.8) and the *within group sum inner product* defined by (9.4.9) and the analysis of variance decomposition (9.4.7) generalizes to the multivariate analysis of variance decomposition (9.4.10).

Note that knowledge of the analysis of variance decomposition for every variable in the space spanned by $V_1, V_2, \ldots, V_s$ is equivalent to the knowledge of the multivariate analysis of variance decomposition (9.4.10). In other words, multivariate analysis of variance may be described as the determination of the whole complex of univariate analyses of variance for all of the variables of a subspace.

Recall, however, the remark made earlier that the concern of Section 9.4 is with classifications which are more like *bona fide* variables rather than classifications which mark the individuals into separate samples. Since the terminology of analysis of variance is more often used in the latter situation, its uses here should be compared with the uses discussed in Chapters 10 and 11. Here the concern is more with best linear prediction from one subspace to another and with correlation analyses relating the two subspaces.

If the multiple correlation coefficient between V_i and the set $V_{s+1}, V_{s+2}, \ldots, V_p$ is denoted by t_i, then

$$t_i^2 = (V_i, V_i)_A/(V_i, V_i)_T = \sum_{j=1}^{p-s} n_j(\bar{X}_i^{(j)} - \bar{X}_i)^2 \Big/ \sum_{j=1}^{p-s}\sum_{l=1}^{n_j}(X_i^{(j,l)} - \bar{X}_i)^2 \quad (9.4.11)$$

for $i = 1, 2, \ldots, s$. The equivalent ratio

$$H_i = (V_i, V_i)_A/(V_i, V_i)_W = t_i^2/(1 - t_i^2) \quad (9.4.12)$$

is more familiar in analysis of variance contexts. The canonical correlation analysis relating the sets $V_1, V_2, \ldots, V_s$ and $V_{s+1}, V_{s+2}, \ldots, V_p$ produces sets of canonical variables $W_1, W_2, \ldots, W_s$ and $W_{s+1}, W_{s+2}, \ldots, W_p$. Formulas like (9.4.11) and (9.4.12) could also be written for the canonical correlation coefficients r_i and their associated

$$G_i = r_i^2/(1 - r_i^2). \quad (9.4.13)$$

This set of G_i generalizes the single ratio G defined in (8.5.3).

Certain details need to be filled in. Since $P_{II}(V_{s+1}), \ldots, P_{II}(V_p)$ span a subspace of dimension $p - s - 1$, the among group sum inner product is said to have $p - s - 1$ *degrees of freedom*. Similarly, the within group sum inner product is said to have $n - (p - s)$ degrees of freedom. The rank of the among group sum inner product is the dimension of the subspace spanned by the components of $P(V_1), P(V_2), \ldots, P(V_s)$ in the $(p - s - 1)$-dimensional subspace spanned by $P_{II}(V_{s+1}), \ldots, P_{II}(V_p)$, and so this rank is at most $\min(s, p - s - 1)$. Similarly, the maximum rank of the within group sum inner product is $\min(s, n - p + s)$. These maxima are generally attained with sample data, since precise linear relations among $V_1, V_2, \ldots, V_s$ are rare.

In the general theory of canonical correlation analysis of Section 9.2, the number of nonzero canonical correlation coefficients is at most $\min(s, p - s)$. Here, because the among group sum inner product has rank at most $\min(s, p - s - 1)$, the number of nonzero canonical correlation coefficients is at most $\min(s, p - s - 1)$. Another way to see this is to note that the sample corrected sum inner product of $V_{s+1}, V_{s+2}, \ldots, V_p$ has rank $p - s - 1$ and so is a semidefinite inner product over a $(p - s)$-dimensional space. The last canonical variable in the set $W_{s+1}, W_{s+2}, \ldots, W_p$ has variance zero, i.e., is $V_{s+1} + V_{s+2} + \cdots + V_p$ apart from a scale factor, and meaningful correlation coefficients may be associated only with $W_{s+1}, W_{s+2}, \ldots, W_{p-1}$.

9.5 THE FORWARD METHOD OF SELECTING PREDICTOR VARIABLES

The discussion in Section 8.2 of various methods for selecting a subset of a set of available predictor variables is extended here to cover the selection of a subset of $V_1, V_2, \ldots, V_s$ for use in predicting the set $V_{s+1}, V_{s+2}, \ldots, V_p$. For brevity only the *forward* method will be discussed, but a similar discussion could easily be supplied for the *backward* method or for other variants.

Having selected $V_{i_1}, V_{i_2}, \ldots, V_{i_{t-1}}$ the forward method of Section 8.2 next chooses V_{i_t} to maximize the multiple correlation coefficient between V_p and the selected predictors. When V_p is generalized to $V_{s+1}, V_{s+2}, \ldots, V_p$, there is unfortunately no single multiple correlation coefficient to be maximized. Instead there is such a multiple correlation coefficient between each variable in the subspace spanned by $V_{s+1}, V_{s+2}, \ldots, V_p$ and the set of selected predictors.

If the problem of choosing V_{i_t} is regarded in a coordinate-free way, then the aim is to make the subspace spanned by $V_{i_1}, V_{i_2}, \ldots, V_{i_t}$ as close as possible to the subspace spanned by $V_{s+1}, V_{s+2}, \ldots, V_p$, where "close" is to be measured in terms of the sample covariance inner product. Any set of quantities possessing this degree of invariance is determined by the set of canonical correlation coefficients. These may be denoted by $r_1, r_2, \ldots, r_{p-s}$, some of which may be zero. Thus the problem is reduced to the specification of a criterion $C = C(r_1, r_2, \ldots, r_{p-s})$ to be optimized by the choice of V_{i_t}. Some of the criteria which have been proposed are

$$C_1 = \prod_{i=1}^{p-s}(1 - r_i^2),$$

$$C_2 = \sum_{i=1}^{p-s} r_i^2, \quad \text{and} \quad C_3 = \sum_{i=1}^{p-s}\frac{r_i^2}{1 - r_i^2}. \quad (9.5.1)$$

These criteria may also be used for testing the null hypothesis that the two sets of variables are uncorrelated. Note that large values of C_2 and C_3 are desirable, while small values of C_1 are hoped for. Another criterion sometimes proposed for testing is $C_4 = \max(r_1, r_2, \ldots, r_{p-s}) = r_1$, but this seems less desirable for a selection criterion since it concentrates on a single dimension and ignores possibly important ability to predict in other dimensions. C_4 also has the disadvantage of requiring that the canonical correlations be computed at each stage, while this computation may be circumvented under C_1, C_2, or C_3. In some circumstances, one might abandon the coordinate-free approach and specify a nonnegative quadratic form in the prediction errors to be minimized. Such a quadratic loss would be equivalent to specifying a new basis $U_{s+1}, U_{s+2}, \ldots, U_p$ of the space spanned by $V_{s+1}, V_{s+2}, \ldots, V_p$ and maximizing the sum of the squares of the multiple correlation coefficients of $U_{s+1}, U_{s+2}, \ldots, U_p$ with the space spanned by $V_{i_1}, V_{i_2}, \ldots, V_{i_t}$.

Having decided on a criterion, the selection procedure is defined, except for a rule saying when to stop. Such a stopping rule would be based on the change in C at any stage, i.e., if the selection of a further predictor does not bring about sufficient improvement in C, then the selection procedure is stopped. A measure of what constitutes sufficient improvement may be deduced in a rough and logically imperfect way by a consideration of significance tests.

Having selected a subset of predictors $V_{i_1}, V_{i_2}, \ldots, V_{i_t}$ there are two further reductions which need consideration. The real concern is with the best linear predictors of $V_{s+1}, V_{s+2}, \ldots, V_p$ in terms of $V_{i_1}, V_{i_2}, \ldots, V_{i_t}$ and, if $p - s < t$, these best linear predictors span only a proper subspace of the space spanned by $V_{i_1}, V_{i_2}, \ldots, V_{i_t}$. Consequently one need only retain this subspace. Moreover, within this subspace one may isolate a set of canonical predictors with associated canonical correlation coefficients given by the nonzero subset of $r_1, r_2, \ldots, r_{p-s}$. Again it may be judged, using a tenuous significance testing argument, that only a smaller subset of $r_1, r_2, \ldots, r_{p-s}$ are large enough to represent more than meaningless sampling fluctuations. If so, the space of effective predictors may be further reduced to include only those canonical variables having apparently adequate r. Note that the second stage of reduction here actually includes the first, for a decision to retain all of the canonical variables which correspond to nonzero canonical correlations automatically restricts consideration to the appropriate space of best linear predictors.

To understand the computations required for forward variable selection, it is necessary to have a convenient form for the chosen criterion C and then to find a direct way to compute the revised C when a new variable enters the system. Since the criteria (9.5.1) are symmetric between the two sets $V_{i_1}, V_{i_2}, \ldots, V_{i_t}$ and $V_{s+1}, V_{s+2}, \ldots, V_p$ the following description may be applied with the two sets interchanged. However, the prescription for revising C when a new variable is added is different depending on whether the first or second set is swept. The description here follows the latter path, but the reader may wish to work out the former and compare the two.

Suppose that

$$\mathbf{T} = \begin{bmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{T}_{21} & \mathbf{T}_{22} \end{bmatrix} \tag{9.5.2}$$

denotes the corrected sum inner product matrix of the two sets $V_{i_1}, V_{i_2}, \ldots, V_{i_t}$ and $V_{s+1}, V_{s+2}, \ldots, V_p$. The criteria C_1 and C_2 are, respectively, the product of the eigenvalues of $\mathbf{T}_{11.2} = \mathbf{T}_{11} - \mathbf{T}_{12}\mathbf{T}_{22}^{-1}\mathbf{T}_{21}$ relative to $\mathbf{T}_{11}$ and the sum of the eigenvalues of $\mathbf{T}_{11} - \mathbf{T}_{11.2}$ relative to $\mathbf{T}_{11}$. These quantities are simply expressible when $V_{i_1}, V_{i_2}, \ldots, V_{i_t}$ is replaced by an orthonormal basis $U_1, U_2, \ldots, U_t$ so that the corrected sum inner product matrix of $U_1, U_2, \ldots, U_t, V_{s+1}, V_{s+2}, \ldots, V_p$ has the form

$$\dot{\mathbf{T}} = \begin{bmatrix} \mathbf{I} & \dot{\mathbf{T}}_{12} \\ \dot{\mathbf{T}}_{21} & \mathbf{T}_{22} \end{bmatrix}. \tag{9.5.3}$$

Then C_1 is the product of the eigenvalues of $\mathbf{I} - \dot{\mathbf{T}}_{12}\mathbf{T}_{22}^{-1}\dot{\mathbf{T}}_{21}$ relative to $\mathbf{I}$, or

$$C_1 = \det\,(\mathbf{I} - \dot{\mathbf{T}}_{12}\mathbf{T}_{22}^{-1}\dot{\mathbf{T}}_{21}), \tag{9.5.4}$$

and, similarly,

$$\begin{aligned} C_2 &= \operatorname{tr}\,(\dot{\mathbf{T}}_{12}\mathbf{T}_{22}^{-1}\dot{\mathbf{T}}_{21}) \\ &= t - \operatorname{tr}\,(\mathbf{I} - \dot{\mathbf{T}}_{12}\mathbf{T}_{22}^{-1}\dot{\mathbf{T}}_{21}). \end{aligned} \tag{9.5.5}$$

Note that $\dot{\mathbf{T}}$ may be defined from $\mathbf{T}$ using MST$[1, 2, \ldots, t]$, while $\mathbf{I} - \dot{\mathbf{T}}_{12}\mathbf{T}_{22}^{-1}\dot{\mathbf{T}}_{21}$ follows from $\mathbf{T}$ using SWP$[t + 1, t + 2, \ldots, t + p - s]$. (One of these sweeping operations must be omitted if $V_{s+1}, V_{s+2}, \ldots, V_p$ represent a categorical variable.) Finally det $(\mathbf{I} - \dot{\mathbf{T}}_{12}\mathbf{T}_{22}^{-1}\dot{\mathbf{T}}_{21})$ is a by-product of the successive sweepings SWP$[1, 2, \ldots, t]$ applied to $\mathbf{I} - \dot{\mathbf{T}}_{12}\mathbf{T}_{22}^{-1}\dot{\mathbf{T}}_{21}$. The criterion C_3 is similar to C_2 except that the eigenvalues being summed are those of $\mathbf{T}_{11} - \mathbf{T}_{11.2}$ relative to $\mathbf{T}_{11.2}$ and different orthogonalizations are important.

Finally, consider the situation when $V_{i_1}, V_{i_2}, \ldots, V_{i_t}$ have been selected and a further $V_{i_{t+1}}$ is contemplated. Suppose that SWP$[t + 1, t + 2, \ldots, t + p - s]\mathbf{T}$ and SWP$[t + 1, t + 2, \ldots, t + p - s,\ 1, 2, \ldots, t]\mathbf{T}$ are both in hand. Suppose that $V_{i_{t+1}}$ is assimilated into the system and that $\mathbf{T}_{(+)}$ denotes the $(t + p - s + 1) \times (t + p - s + 1)$ extension of $\mathbf{T}$ with $V_{i_{t+1}}$ put into row and column $t + 1$. The ASM operator of Section 4.3.3 may be made to produce SWP$[t + 2, t + 3, \ldots, t + p - s + 1]\mathbf{T}_{(+)}$ and SWP$[t + 2, t + 3, \ldots, t + p - s + 1, 1, 2, \ldots, t]\mathbf{T}_{(+)}$. Denoting the $(t + 1, t + 1)$ elements of these two matrices by A_{t+1} and B_{t+1}, respectively, it is easily checked that B_{t+1}/A_{t+1} is the factor by which C_1 must be multiplied when $V_{i_{t+1}}$ is adjoined. The reader may supply corresponding descriptions for C_2 and C_3, which are somewhat easier. It should be stressed that the forward method requires that these assimilation procedures be tried for all possible choices of $V_{i_{t+1}}$ at each stage. This means that sums of products of each such $V_{i_{t+1}}$ with each of $V_{i_1}, V_{i_2}, \ldots, V_{i_t}, V_{s+1}, V_{s+2}, \ldots, V_p, V_0$ must be found. However, the sums of products for each pair of unselected variables are not required, which can result in considerable saving of computation when the number of unselected variables is large, as in the following example.

9.6 AN EXAMPLE ILLUSTRATING PREDICTION OF A CATEGORICAL VARIABLE AND FORWARD SELECTION OF PREDICTOR VARIABLES

Example 9.2. The following discussion summarizes an analysis of meteorological data by Dr. Robert Miller (1961, 1962). The objective of the analysis was to provide a method of forecasting short term ceiling conditions at an air force base. The specific data refer to McGuire Air Force Base, Wrightstown, New Jersey. The forecasts were to be provided two hours in the future for the categorical variable determined by the five ceiling conditions:

1. Closed: ceiling 0–200 feet.
2. Low instrument: ceiling 200–500 feet.
3. High instrument: ceiling 500–1500 feet.
4. Low open: ceiling 1500–5000 feet.
5. High open: ceiling 5000 feet and up.

Any particular forecast is based on the values of 75 quantities measuring 15 meteorological variables at each of five weather stations at the time of forecasting. The five weather stations consist of McGuire Air Force Base and four surrounding stations at Philadelphia, Newark, Lakehurst Naval Air Station, and Atlantic City. The 15 meteorological variables are:

1. Height of the lowest cloud layer.
2. Height of the second cloud layer.
3. Amount of lowest cloud layer.
4. Amount of second cloud layer.
5. Height of ceiling.
6. Visibility.
7. Three hour change in ceiling height.
8. Three hour change in visibility.
9. Three hour change in pressure.
10. Temperature-dewpoint depression/temperature.
11. East-west wind component.
12. North-south wind component.
13. Three hour change in wind direction.
14. Total cloud cover.
15. Three hour change in temperature.

The past data to be used in setting up the forecasting method consist of 1874 time points falling at consecutive three hour intervals during the winter months of 1954–55 and 1955–56. A further 926 time points from the winter months of 1956–57 were held in reserve for checking the forecasts on independent data. At each of these 1874 + 926 time points the values of 75 variables are given, and the corresponding ceiling category two hours later is also given. In the original data the numbers n_i in ceiling category i for $i = 1, 2, 3, 4, 5$ are 49, 84, 158, 228, 1355. In the independent data the corresponding numbers are 35, 76, 118, 124, 573. Evidently, the flying weather was worse in the winter of 1956–57 than in the average of the two preceding winters.

Miller's basic calculations with the original sample of size 1874 were those described in Section 9.4 and 9.5. Miller used the terminology of multivariate analysis of variance rather than the equivalent approach in terms of artificial variables $V_{s+1}, V_{s+2}, \ldots, V_p$ and his computer programs were written directly rather than in terms of the operators described in this book. The forward method of selection was used, the criterion being C_3 in (9.5.1). Note that

$$C_3 = \sum_{i=1}^{p-s} G_i,$$

using the G measure (9.4.9) in place of the r measure. Miller reports that the first five predictor variables chosen in order are:

1. Height of ceiling at Philadelphia.
2. Height of ceiling at McGuire Air Force Base.
3. East-west component at McGuire Air Force Base.
4. Height of ceiling at Newark.
5. Total cloud cover at Newark.

The corresponding values of C_3 after selecting t predictors for $t = 1, 2, 3, 4, 5$ are 2.369, 2.994, 3.171, 3.318, 3.419. Miller ceased selecting at this point because he judged that the next few variables selected resulted in an insufficient increase in the value of C_3. The basis for this judgment has a tenuous connection with significance testing, but it is doubtful whether the subsequent analysis would have been much altered by the addition of a small number of additional predictors. Miller had found from experience that the inclusion of a large number of predictors is often harmful in that forecasts based on independent data become less accurate.

The next step was to carry out the canonical correlation analysis of the five selected predictors against the artificial variables representing the classification. This results in four nontrivial canonical correlation coefficients r_i and the four corresponding linear combinations of the five selected predictor variables. Miller reports the values of $G_i = r_i^2/(1 - r_i^2)$ to be

$$\begin{aligned} G_1 &= 3.312 \\ G_2 &= 0.087 \\ G_3 &= 0.020 \\ G_4 &= 0.001 \end{aligned}$$

and the corresponding canonical predictors to be

$$\begin{aligned} W_1 &= -7.246V_1 - 6.276V_2 + 1.898V_3 - 3.008V_4 + 1.000V_5 \\ W_2 &= 0.251V_1 - 0.271V_2 - 0.589V_3 + 0.319V_4 + 1.000V_5 \\ W_3 &= -2.353V_1 + 2.664V_2 + 0.980V_3 + 0.971V_4 + 1.000V_5 \\ W_4 &= 7.105V_1 - 10.880V_2 + 4.427V_3 + 7.372V_4 + 1.000V_5, \end{aligned}$$

where V_1, V_2, V_3, V_4, V_5 represent the five selected predictors listed above. Note the scaling of each W_i has been fixed arbitrarily by choosing the coefficient of V_5 to be unity.

The canonical correlation analysis provides four different potential sets of predictors, i.e., W_1 alone *or* W_1 and W_2 *or* W_1, W_2, and W_3 *or* W_1, W_2, W_3, and W_4. Miller again presents a tenuous significance-testing argument suggesting that only W_1 need be taken into account. Later he shows empirically that the best forecasts on the independent data result from the use of three or four canonical predictors.

The analysis was also made to produce the sample mean vectors and sample covariance matrices of the five selected predictors for each of the five subsamples corresponding to the five ceiling categories. From these the sample mean vectors and sample covariances of the four canonical predictors were found, again for each of the five subsamples. Rather than reproduce these numbers, a graphical analysis of the first two canonical variables is reproduced. Suppose that W_1^* and W_2^* denote rescaled versions of W_1 and W_2, chosen to have pooled within sample mean squares of 1 and G_2/G_1, respectively. The corresponding dual two-dimensional individual-space is represented in the following graphs where w_1^* and w_2^* are drawn to appear orthonormal. Figure 9.6.1 shows the mean-centered concentration ellipses of the five subsamples, except that the radii have been rescaled by the factor 1.18, which is the factor required to rescale the ellipse of concentration of a bivariate normal probability distribution so that half of the probability is contained within the ellipse. In other words, it is the factor such that roughly half of the sample points should lie inside the ellipse if the subsample distributions resemble bivariate normal distributions. The subsequent Figs. 9.6.2, 9.6.3, 9.6.4, 9.6.5, and 9.6.6 show the actual sample individuals in the five subsamples, including both those in the original data and those in the independent data.

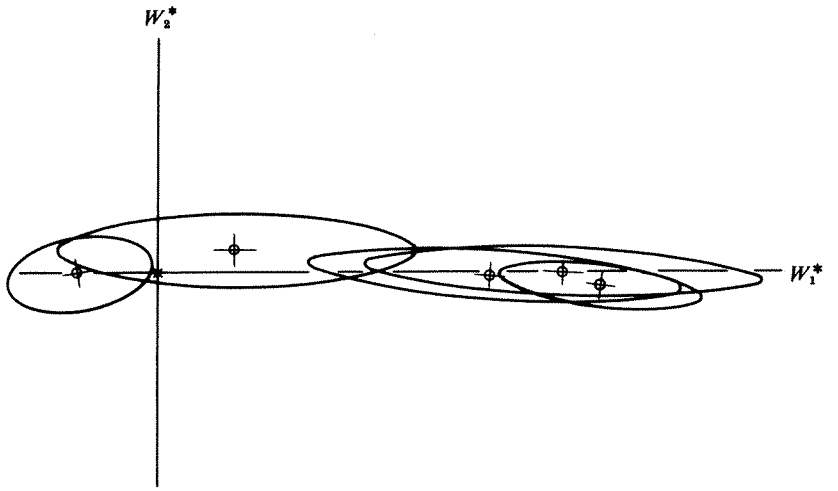

Fig. 9.6.1. Concentration ellipses (after scaling by the factor 1.18) corresponding to the five weather categories, and defined by the canonical variables W_1^*, W_2^* and the original sample of 1874 time points.

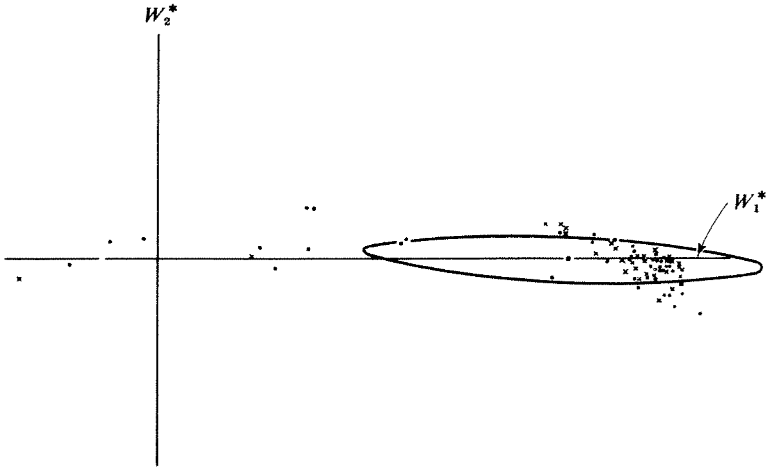

Fig. 9.6.2. Same as Fig. 9.6.1, showing the concentration ellipse of the first weather category only, and showing the original sample points (dots) and independent sample points (crosses) of the first weather category.

Miller remarks that these pictures show clearly that the subsamples do not follow bivariate normal distributions. He also notes that the two sets of data coincide fairly well except possibly for ceiling category 2, and points out the

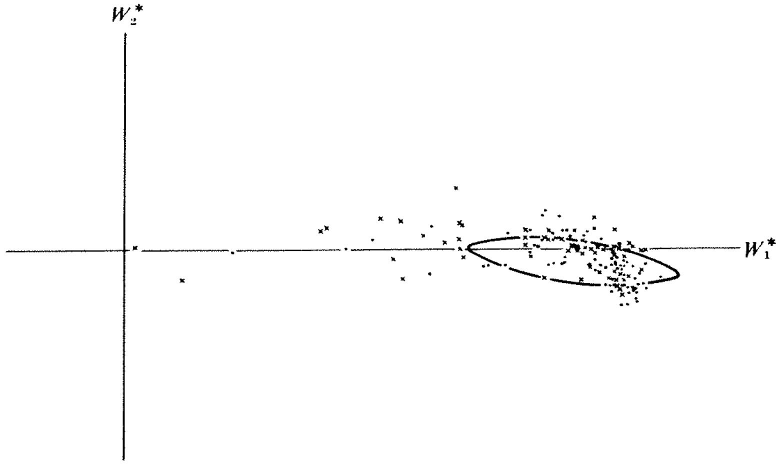

Fig. 9.6.3. Same as Fig. 9.6.2, replacing the first weather category by the second.

double clustering in category 5, apparently due to discreteness in the predictors.

For a detailed discussion of how these preliminary stages of data reduction lead to actual forecasts the reader is referred to Miller's monograph, but the general approach is clear from Fig. 9.6.1. Each time a forecast is desired, the values of the canonical predictors W_1^* W_2^* are computed and a point in the plane of Fig. 9.6.1 is located. One must then decide how probable it is that this point belongs in each of the five subsamples. Roughly speaking this is to be based on

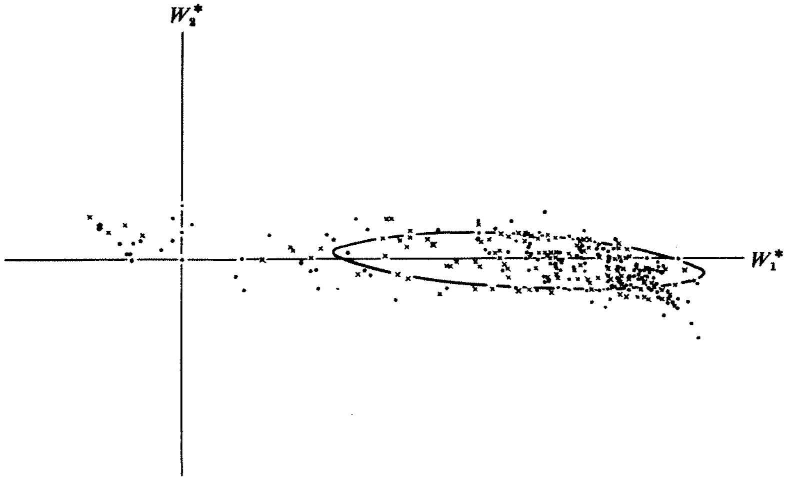

Fig. 9.6.4. Same as Fig. 9.6.2, replacing the first weather category by the third.

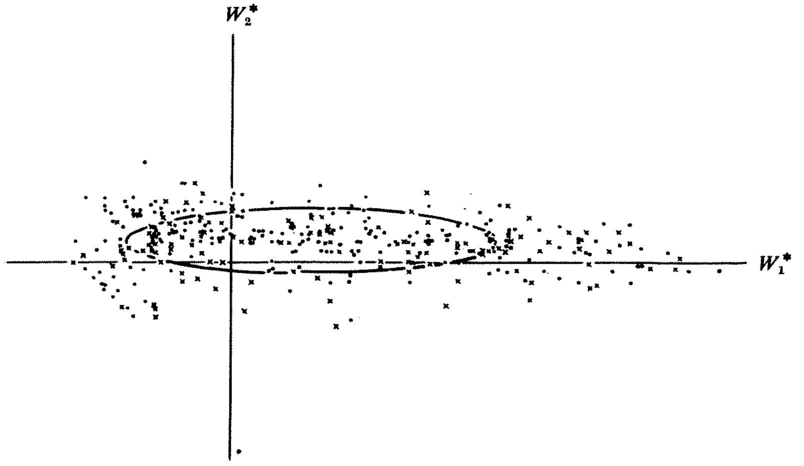

Fig. 9.6.5. Same as Fig. 9.6.2, replacing the first weather category by the fourth.

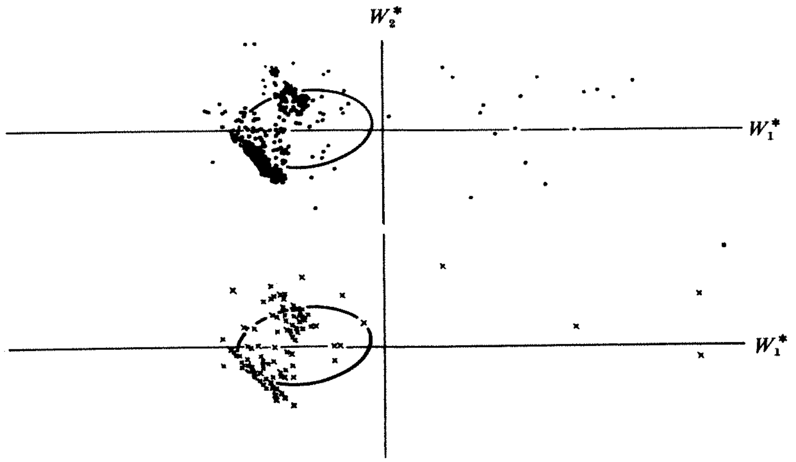

Fig. 9.6.6. Same as Fig. 9.6.2, replacing the first weather category by the fifth, and separating the original and independent data.

distance from subsample means. The scaling of W_1^* and W_2^* to have sample variances 1 and G_2/G_1 was chosen so that the distance measure reflects the relative importance of the canonical variables as measured by G. Miller had some empirical backing for the specific choice of 1 and G_2/G_1, i.e., other choices have produced less accurate forecasts in a variety of examples.

Corresponding to the canonical variables W_1, W_2, W_3, W_4 is another set of canonical variables which are linear combinations of the artificial variables which represent the ceiling categories. These corresponding canonical variables are determined up to scale changes to be the best linear predictors of W_1, W_2, W_3, W_4 in terms of the artificial variables, and the form of these best linear predictors is clear from (9.4.5), i.e., the subsample means of the canonical predictors provide the other set of canonical variables. These subsample means for at least W_1 and W_2 may be seen in Fig. 9.6.1. Drawing a rough curve through these means suggests that the first canonical predictor may be thought of as discriminating between high and low ceilings while the second canonical predictor separates the middle range from the two extremes.

The fact that the sample means in Fig. 9.6.1 do not lie in order on a simple smooth curve may plausibly be regarded as a consequence of sampling variation. Recall that the ceiling categories 1 and 2 have relatively few individuals. This sampling variation is a very complex phenomenon for several reasons. For one, the effect of the procedure for selecting variables on sampling variation is very difficult to understand. In addition, the individuals in these data are successive time points at only three hour intervals. Since weather conditions often change little over such periods, there is a sense in which there are many fewer than the apparent number of sample individuals, and correspondingly greater sampling variation could be expected than for random samples from a population of the same size.

Much as with the psychological examples 8.4 and 9.1, it is tempting to try to interpret the selected predictors and canonical predictors in terms of physical or meteorological theories. For example, the selection of the Philadelphia ceiling and the east-west wind component suggests the importance of looking west for future weather. No direct tie with a physical theory seems possible, however.

CHAPTER 10

TWO OR MORE SAMPLES OF INDIVIDUALS

10.1 INTRODUCTION

A set of k samples on a common set of p variables may be determined by their k data matrices $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(k)}$. If the sample sizes are $n_1, n_2, \ldots, n_k$, then the data matrices have dimensions $n_1 \times p, n_2 \times p, \ldots, n_k \times p$, respectively. The ith row of $\mathbf{X}^{(l)}$ representing the ith individual in sample l will be denoted by $\mathbf{X}^{(l,i)}$ for $i = 1, 2, \ldots, n_l$ and $l = 1, 2, \ldots, k$.

Each of these samples yields the basic statistics which were described in Section 7.2 for a single sample, namely

$$\bar{\mathbf{X}}^{(l)} = \sum_{i=1}^{n_l} \mathbf{X}^{(l,i)}/n_l, \tag{10.1.1}$$

$$\mathbf{Q}^{(l)} = \sum_{i=1}^{n_l} \mathbf{X}^{(l,i)\prime}\mathbf{X}^{(l,i)}, \tag{10.1.2}$$

$$\mathbf{T}^{(l)} = \mathbf{Q}^{(l)} - n_l\bar{\mathbf{X}}^{(l)\prime}\bar{\mathbf{X}}^{(l)}, \tag{10.1.3}$$

and

$$\mathbf{S}^{(l)} = \mathbf{T}^{(l)}/(n_l - 1), \tag{10.1.4}$$

for $l = 1, 2, \ldots, k$.

If the k samples are combined to form a single sample of size

$$n = \sum_{l=1}^{k} n_l, \tag{10.1.5}$$

then there is a *total mean* or *grand mean* vector

$$\bar{\mathbf{X}} = \sum_{l=1}^{k}\sum_{j=1}^{n_l} \mathbf{X}^{(l,j)}/n_l \tag{10.1.6}$$

and a corresponding *total sum of products corrected for the grand mean* inner product matrix

$$\mathbf{T} = \sum_{l=1}^{k}\sum_{j=1}^{n_l} (\mathbf{X}^{(l,j)} - \bar{\mathbf{X}})'(\mathbf{X}^{(l,j)} - \bar{\mathbf{X}}), \tag{10.1.7}$$

and thence a *total sample covariance* matrix

$$\mathbf{S} = \mathbf{T}/(n-1). \tag{10.1.8}$$

It is often unnatural to pool k samples into a single sample, and correspondingly the statistics $\bar{\mathbf{X}}$, $\mathbf{T}$, and $\mathbf{S}$ may not have much direct appeal. They do, however, relate closely to the analysis of variance formulation which was introduced in Section 9.4. Thus $\mathbf{T}$ decomposes into

$$\mathbf{T} = \mathbf{T}_A + \mathbf{T}_W, \tag{10.1.9}$$

where

$$\mathbf{T}_A = \sum_{l=1}^{k} n_l(\bar{\mathbf{X}}^{(l)} - \bar{\mathbf{X}})'(\bar{\mathbf{X}}^{(l)} - \bar{\mathbf{X}}) \tag{10.1.10}$$

and

$$\mathbf{T}_W = \sum_{l=1}^{k}\sum_{j=1}^{n_l}(\mathbf{X}^{(l,j)} - \bar{\mathbf{X}}^{(l)})'(\mathbf{X}^{(l,j)} - \bar{\mathbf{X}}^{(l)})$$
$$= \sum_{l=1}^{k}\mathbf{T}^{(l)}. \tag{10.1.11}$$

This is mathematically natural in the sense of representing the inner products of components in special orthogonal subspaces in $\mathscr{N}$ which are defined in terms of the indicator variables of the k groups. The theory was given in Section 9.4, where s was used in place of p and $p - s$ in place of k, and will not be repeated. Here $\mathbf{T}_A$ defines the among group inner product on $k - 1$ degrees of freedom with rank at most min $(p, k-1)$ while $\mathbf{T}_W$ defines the within group inner product on $n - k$ degrees of freedom with rank at most min $(p, n-k)$. In the special case $k = 2$, $n\bar{\mathbf{X}} = n_1\bar{\mathbf{X}}^{(1)} + n_2\bar{\mathbf{X}}^{(2)}$ and substitution in (10.1.10) leads to the inner product matrix

$$\mathbf{T}_A = (\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}}^{(2)})'(\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}}^{(2)})/(1/n_1 + 1/n_2) \tag{10.1.12}$$

with rank unity (or zero if $\bar{\mathbf{X}}^{(1)} = \bar{\mathbf{X}}^{(2)}$).

Analysis of variance considerations also suggest the weighted average

$$\mathbf{S}_W = \mathbf{T}_W/(n-k), \tag{10.1.13}$$

which may be written in the form

$$\mathbf{S}_W = \sum_{l=1}^{k} w_l\mathbf{S}^{(l)}\Big/\sum_{l=1}^{k} w_l, \tag{10.1.14}$$

where

$$w_l = n_l - 1 \tag{10.1.15}$$

for $l = 1, 2, \ldots, k$. The particular weights (10.1.15) are suggested by the sampling theory of statistical inference for circumstances where it may be assumed that the k populations being sampled have essentially the same covariance inner products.

In general, the grand mean $\bar{\mathbf{X}}$ defined in (10.1.6) or the pooled within sample covariance $\mathbf{S}_W$ defined in (10.1.13) are not directly interesting quantities unless the underlying population means or population covariances, respectively, are common. Both $\bar{\mathbf{X}} = (n_1\bar{\mathbf{X}}^{(1)} + n_2\bar{\mathbf{X}}^{(2)} + \cdots + n_k\bar{\mathbf{X}}^{(k)})/n$ and $\mathbf{S}_W$ weight the contributions from each sample roughly in proportion to sample size. But the variation in sample size may be accidental and irrelevant to the subject of investigation. Equal weightings of $\bar{\mathbf{X}}^{(l)}$ and $\mathbf{S}^{(l)}$ could be interesting, as could be weightings which reflect population sizes. For any set of constants $c_1, c_2, \ldots, c_k$, the weighting $w_l = c_l^2/n_l$ in (10.1.14) would lead to an $\mathbf{S}_W$ appropriate for judging the sampling variation of $(c_1\bar{\mathbf{X}}^{(1)} + c_2\bar{\mathbf{X}}^{(2)} + \cdots + c_k\bar{\mathbf{X}}^{(k)})/(c_1 + c_2 + \cdots + c_k)$.

10.2 TWO-SAMPLE ANALYSIS

Consideration will be given first to the comparison of a pair of sample means $m^{(1)}$ and $m^{(2)}$ and later to the more complex task of describing similarities and differences between a pair of sample covariance inner products. Viewed as two points in individual-space $\mathscr{F}$, with no reference to any coordinate system or to any inner product, the relative positions of $m^{(1)}$ and $m^{(2)}$ cannot be compared in any meaningful way. Consequently an inner product π_d over $\mathscr{F}$, or equivalently its dual inner product π over $\mathscr{E}$, will be assumed. The dual pair π and π_d will be assumed to have full rank unless otherwise stated, but their source need not be specified exactly. They might be sample-determined in various ways, or they might be reference inner products determined apart from the sample data. The initial aim is to study the information provided by $m^{(1)}$, $m^{(2)}$, and π_d.

Two important concepts arise here, namely that of *distance* which was introduced by Mahalanobis (1936) and that of *best linear discriminator* which was independently introduced by Fisher (1936, 1938). The close relations between these concepts, to be explored below, quickly became clear.

The *distance D* between $m^{(1)}$ and $m^{(2)}$ may be defined simply to be

$$D = (m^{(1)} - m^{(2)}, m^{(1)} - m^{(2)})_d^{1/2}, \tag{10.2.1}$$

where $(a, a)_d^{1/2}$ denotes the norm of a under π_d. A more general approach is to define a *distance* $D(\mathscr{W})$ *for every subspace* $\mathscr{W}$ *of variable-space* $\mathscr{E}$, so that D defined by (10.2.1) is the special case $D(\mathscr{E})$. $D(\mathscr{W})$ is defined in the same way as $D(\mathscr{E})$ in terms of the same sample means $m^{(1)}$ and $m^{(2)}$ and the same inner product π_d considered as linear functionals and an inner product over the restricted variable-space $\mathscr{W}$. If $\mathscr{W}$ has dimension unity, then $D(\mathscr{W})$ will also be denoted by $D(W)$ for any W different from $\emptyset$ in $\mathscr{W}$, and in this case (10.2.1) may be written

$$D(W) = |m^{(1)}(W) - m^{(2)}(W)|/(W, W)^{1/2}. \tag{10.2.2}$$

Note that the substitution of αW for W, with $\alpha \neq 0$, does not alter the right side of (10.2.2).

A *best linear discriminator* may be defined to be any variable V in $\mathscr{E}$ with maximum $D(V)$, the idea being that such a variable makes the sample means appear as far apart as any variable can. In fact, as shown in Theorem 10.2, such a variable has as much separating power in the sense of distance as the whole space $\mathscr{E}$ of variables.

Theorem 10.2. *The best linear discriminator V defined by $m^{(1)}$, $m^{(2)}$, and π_d is unique up to a scale factor. The unique one-dimensional subspace of best linear discriminators is the dual space of the orthogonal complement of the one-dimensional subspace of $\mathscr{F}$ spanned by $m^{(1)} - m^{(2)}$. For any best linear discriminator V,*

$$D(V) = D(\mathscr{E}), \tag{10.2.3}$$

and, for any variable U in the subspace of $\mathscr{E}$ orthogonal to V,

$$D(U) = 0. \tag{10.2.4}$$

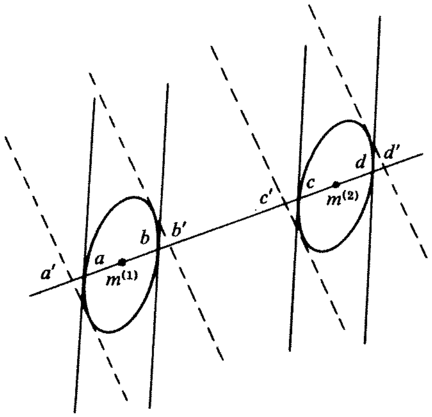

Fig. 10.2.1. Mean-centered concentration ellipses and associated tangent lines ($p = 2$) as described in the text.

The theorem follows simply because there is only a small amount of mathematical structure present, which may be characterized geometrically in individual-space $\mathscr{F}$ as in Fig. 10.2.1, i.e., by the pair of points $m^{(1)}$ and $m^{(2)}$ and the pair of ellipsoids consisting of points at distance unity or less from $m^{(1)}$ and $m^{(2)}$ according to π_d. Define $\mathscr{U}_d$ to be the one-dimensional subspace of $\mathscr{F}$ spanned by $m^{(1)} - m^{(2)}$. Define $\mathscr{V}_d$ to be the orthogonal complement of $\mathscr{U}_d$ in $\mathscr{F}$. Note that the family of parallel $(p - 1)$-dimensional hyperplanes including $\mathscr{V}_d$ also includes the four tangent hyperplanes to the two ellipsoids where the line joining $m^{(1)}$ and $m^{(2)}$ meets these ellipsoids, as illustrated in Fig. 10.2.1. The subspaces $\mathscr{U}$ and $\mathscr{V}$ in $\mathscr{E}$ dual to $\mathscr{U}_d$ and $\mathscr{V}_d$ in $\mathscr{F}$ are those which Theorem 10.2

claims to consist of variables U such that $D(U) = 0$ and of best linear discriminators V such that $D(V) = D(\mathscr{E})$, respectively.

The claim about $\mathscr{U}$ is the more obvious of the two. Each U in $\mathscr{U}$ defines a family of parallel $(p - 1)$-dimensional hyperplanes in $\mathscr{F}$ on which U takes constant values. The duality of $\mathscr{U}$ and $\mathscr{U}_d$ means that the family includes a hyperplane containing the line joining $m^{(1)}$ and $m^{(2)}$. Consequently $m^{(1)}(U) = m^{(2)}(U)$, and, from (10.2.2), $D(U) = 0$.

Consider now any variable W as a candidate for a best linear discriminator. Such a W has a unique expression as $V + U$ with V in $\mathscr{V}$ and U in $\mathscr{U}$. Now $m^{(1)}(W) - m^{(2)}(W) = m^{(1)}(V) - m^{(2)}(V)$ because $m^{(1)}(U) = m^{(2)}(U)$, and $(W, W) \geq (V, V)$ because V and U are orthogonal. Consequently from (10.2.2)

$$D(W) \leq D(V) = D(\mathscr{V}), \tag{10.2.5}$$

which shows that $\mathscr{V}$ is a space of best linear discriminators. Furthermore, the inequality in (10.2.5) is strict unless W belongs to $\mathscr{V}$, so that $\mathscr{V}$ is the unique space of best linear discriminators. Finally, $D(\mathscr{E}) = D(\mathscr{U}) + D(\mathscr{V})$ because $\mathscr{U}$ and $\mathscr{V}$ are orthogonal complements, and $D(\mathscr{U}) = 0$ because $D(U) = 0$ for all U in $\mathscr{U}$, so that $D(V) = D(\mathscr{V}) = D(\mathscr{E})$, as required to complete the proof.

An alternative proof could proceed more directly in terms of the geometry illustrated in Fig. 10.2.1. If the line joining $m^{(1)}$ and $m^{(2)}$ intersects the ellipsoid centered at $m^{(1)}$ in a and b, and the ellipsoid centered at $m^{(2)}$ in c and d, then $D = D(\mathscr{E})$ is the ratio of length of the line segment $m^{(1)}m^{(2)}$ to the length of any of the equal line segments (semi-axes) $m^{(1)}a$, $m^{(1)}b$, $m^{(2)}c$, and $m^{(2)}d$. Any variable W is characterized by the family of parallel $(p - 1)$-dimensional hyperplanes on which it takes constant values, and this family includes four tangents to the two ellipsoids, as illustrated in Fig. 10.2.1. Projection along the family of parallel hyperplanes into the line joining $m^{(1)}$ and $m^{(2)}$ carries individuals having common values on W into a single individual, so that the line becomes a particular representation of the one-dimensional individual-space of W. The two samples still have means $m^{(1)}$ and $m^{(2)}$ after projection, so that $m^{(1)}$ and $m^{(2)}$ represent the sample means in the individual-space of W (and do so for every choice of W). From Theorem 6.6, the shadows $a'b'$ and $c'd'$ cast on the line joining $m^{(1)}$ and $m^{(2)}$ by the projection represent the one-dimensional ellipsoids of points having distance at most unity from $m^{(1)}$ and $m^{(2)}$ according to the inner product dual to the inner product induced by π on the one-dimensional space spanned by W. This means that $D(W)$ is the ratio of lengths of the line segment $m^{(1)}m^{(2)}$ to any of the line segments $m^{(1)}a'$, $m^{(1)}b'$, $m^{(2)}c'$, $m^{(2)}d'$.

The geometric proof of Theorem 10.2 is now obvious. Clearly the ratio $D(W)$ is maximized when $a = a'$, $b = b'$, $c = c'$, $d = d'$, which is achieved by making W that variable V whose corresponding hyperplanes are tangent at a, b, c, and d. Also $D(V) = D(\mathscr{E})$ for such a V. Moreover, the line segment $m^{(1)}a'$ has infinite length when W is any U orthogonal to V, so that $D(U) = 0$ for such U.

Some readers may object that the foregoing discussion does not give simple formulas or computing rules for distances or best linear discriminators. But such concrete descriptions are trivial, and not in themselves very illuminating. If $\mathbf{V}$ and $\mathbf{v}$ are dual bases of $\mathscr{E}$ and $\mathscr{F}$, if $m^{(1)} = \bar{\mathbf{X}}^{(1)}\mathbf{v}$ and $m^{(2)} = \bar{\mathbf{X}}^{(2)}\mathbf{v}$, and if π is represented by $\mathbf{Q}$ relative to $\mathbf{V}$, then

$$D^2 = (\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}}^{(2)})\mathbf{Q}^{-1}(\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}}^{(2)})', \tag{10.2.6}$$

and

$$V = [(\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}}^{(2)})\mathbf{Q}^{-1}]\mathbf{V} \tag{10.2.7}$$

defines a best linear discriminator. The reader should check directly that $D(V)^2 = D^2$ for V defined in (10.2.7), and that this particular choice of V is scaled to have norm D^2. To compute D^2 and V, given $\bar{\mathbf{X}}^{(1)}$, $\bar{\mathbf{X}}^{(2)}$ and $\mathbf{Q}$, a convenient approach is to set up a $(p + 1) \times (p + 1)$ matrix

$$\begin{bmatrix} \mathbf{Q} & (\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}}^{(2)})' \\ \bar{\mathbf{X}}_1^{(1)} - \bar{\mathbf{X}}^{(2)} & 0 \end{bmatrix},$$

and to apply SWP$[1, 2, \ldots, p]$. The last row of the result yields $(\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}}^{(2)})\mathbf{Q}^{-1}$ and $-D^2$. Moreover, the successive stopping points SWP$[1, 2, \ldots, s]$ along the way for $s = 1, 2, \ldots, p$ provide best linear discriminators and distances for subsets of $\mathbf{V}$, making it possible to see how D increases as variables are successively put into the system.

It may sometimes be useful to produce a full orthogonal basis whose first member is a best linear discriminator and whose remaining members span the orthogonal space $\mathscr{U}$. From Theorem 10.2, such a basis may be characterized as a basis of eigenvectors of the among sample inner product (10.1.12) relative to π. The reader may check that the corresponding eigenvalues are $(1/n_1 + 1/n_2)^{-1}D^2, 0, 0, \ldots, 0$.

Sample-based inner products will now be considered more fully, first as a means of choosing π in the analyses just described, and then as sample properties which merit comparison in their own right.

The most common choice for π is the pooled within sample covariance with inner product matrix $\mathbf{S}_W$. In this case $(1/n_1 + 1/n_2)^{-1}D(W)^2$ is immediately seen from (10.2.2) and (10.1.12) to be the familiar ratio of among sample mean square to pooled within sample mean square (or F statistic) as used in the analysis of variance. The alternative choices for π defined by the inner product matrices $\mathbf{S}_W$, $\mathbf{T}_W$, $\mathbf{S}$, or $\mathbf{T}$ are all closely related. By expressing the data in terms of a basis of eigenvectors of $\mathbf{T}_A$ relative to $\mathbf{S}_W$ it is easily seen that all four choices for π lead to the same best linear discriminators while the corresponding distances D, D', D'', and D''' are related by

$$D^2 = (n_1 + n_2 - 2)D'^2, \tag{10.2.8}$$

$$D''^2 = (n_1 + n_2 - 1)D'''^2, \tag{10.2.9}$$

and

$$D'''^2 = D'^2/[1 + D'^2/(1/n_1 + 1/n_2)]. \tag{10.2.10}$$

The various distances D, D', D'', and D''' are all in the tradition of analysis of variance thinking. If it is presumed that the samples come from populations with essentially different covariance structures, then other distances may be interesting. For example, the use of the individual sample covariances $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ determines a pair of *directed distances* $D^{(1)}$ and $D^{(2)}$. The directed distance $D^{(1)}$ measures how far away the mean of sample 2 appears from the center of sample 1 when viewed in relation to the spread of sample 1, and so measures the extent to which $\bar{\mathbf{X}}^{(2)}$ might or might not appear a plausible member of sample 1. A similar observation applies to $D^{(2)}$. Note that $\bar{\mathbf{X}}^{(1)}$ could easily appear a plausible member of sample 2 while $\bar{\mathbf{X}}^{(2)}$ did not appear a plausible member of sample 1, or vice versa. More generally, weightings as in (10.1.14) for various w_i could sometimes be of interest. In particular, for judging the significance of the difference $\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}}^{(2)}$ irrespective of the difference in covariance matrices, one would be interested in D_Δ defined from

$$\mathbf{S}_\Delta = \frac{\mathbf{S}^{(1)}/n_1 + \mathbf{S}^{(2)}/n_2}{1/n_1 + 1/n_2}. \tag{10.2.11}$$

Direct comparison of $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ may be approached in several ways. Simple direct comparison of corresponding elements will give first indications. Another possibility would be to examine the eigenvalues and eigenvectors of $\mathbf{S}^{(1)}$ relative to $\mathbf{S}^{(2)}$. Note that these eigenvectors are also orthogonal with respect to any weighted combination of $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$, such as $\mathbf{S}_W$ and $\mathbf{S}_\Delta$. A third approach relating the apparent differences to $m^{(1)}$ and $m^{(2)}$ is discussed below.

Four possible relative positions and shapes of a pair of mean-centered concentration ellipsoids are illustrated in Fig. 10.2.2, for $p = 2$, as cases a, b, c, d. In case a, $\mathbf{S}^{(1)} = \mathbf{S}^{(2)}$, while in cases b, c, and d increasingly general differences between the two covariance structures appear. The tangent lines to the ellipses in Fig. 10.2.2 represent $(p - 1)$-dimensional hyperplanes in $\mathscr{F}$ on which the corresponding best linear discriminator is constant, while, by the shadow theory of Theorem 6.6, the semi-axes of the ellipsoids along the line joining $m^{(1)}$ and $m^{(2)}$ represent the sample standard deviations when the individual-space of the best linear discriminator is represented as the line joining $m^{(1)}$ and $m^{(2)}$. Case b refers to a situation where the best linear discriminators defined by $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ are the same and even have the same standard deviations, while the covariance structures differ in other respects. Under case c, the two best linear discriminators are again the same, but this common best linear discriminator has different sample variances in the two samples, as is shown by the different lengths of the semi-axes of the ellipses. Finally, case d represents a general situation where the sample best linear discriminators defined by $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ are different.

Although case d will almost always occur with actual sample data, it is of interest to ascertain how close to situations a, b, or c the data come. Several statistics which may partly serve this purpose will now be defined. Suppose that the origin-centered concentration ellipsoids defined by $\mathbf{S}^{(1)}$, $\mathbf{S}^{(2)}$, and $\mathbf{S}_W$ are denoted by $\pi^{(1)}$, $\pi^{(2)}$, and π_W respectively. Suppose that the line joining ø and $m^{(1)} - m^{(2)}$ cuts these ellipsoids in $s^{(1)}$, $s^{(2)}$, and s_W. Two further points s_1 and s_2 on this line may be defined to be the intersections of the line with the pair of $(p - 1)$-dimensional hyperplanes tangent to $\pi^{(1)}$ and $\pi^{(2)}$ and parallel to the hyperplane which is tangent to π_W at s_W. This situation is pictured in Fig. 10.2.3 for the general case d. Note that Fig. 10.2.3 includes the information of the kind given by case d of Fig. 10.2.2 along with additional information.

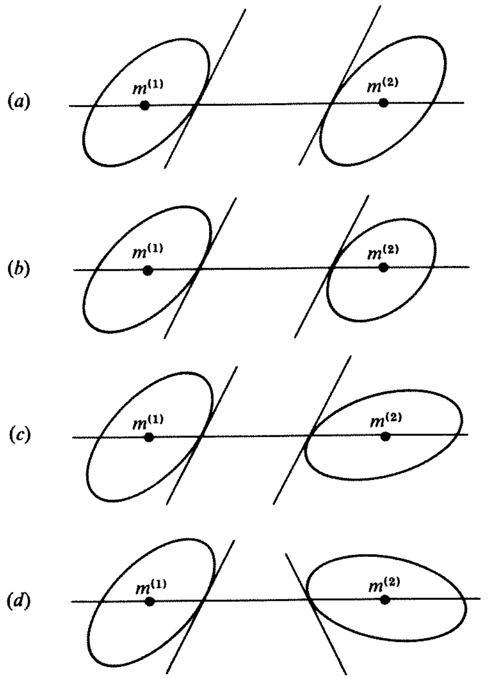

Fig. 10.2.2. Four possible relative positions of a pair of concentration ellipsoids.

Under case a described above, the three ellipsoids $\pi^{(1)}$, $\pi^{(2)}$, and π_W are identical, and consequently the five points $s^{(1)}$, $s^{(2)}$, s_W, s_1, and s_2 are likewise identical. Under case b, the three ellipsoids are no longer identical, but they still define the same best linear discriminator and assign it the same variance, so that $s^{(1)}$, $s^{(2)}$, s_W, s_1, and s_2 are still the same. Under case c ambiguity begins to appear, not in the concept of best linear discriminator which is the same for $\mathbf{S}^{(1)}$, $\mathbf{S}^{(2)}$, and any weighting thereof, but in the variance to be assigned to this best linear discriminator; for $s^{(1)} = s_1$ and $s^{(2)} = s_2$, while the ratio of the lengths of the line segments $øs_1$ and $øs_2$ deviates from unity and is therefore a measure of deviation from case b towards case c. These lengths are proportional to the standard deviations of the common best linear discriminator under $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$, i.e., proportional to the distance-like quantities D_1 and D_2 expressible as

$$D_i^2 = [(\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}}^{(2)})\mathbf{S}_W^{-1}]\mathbf{S}^{(i)}[(\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}}^{(2)})\mathbf{S}_W^{-1}]' \qquad (10.2.12)$$

for $i = 1, 2$, and a suggested measure is

$$K = (øs_1)^2/(øs_2)^2 = D_1^2/D_2^2. \qquad (10.2.13)$$

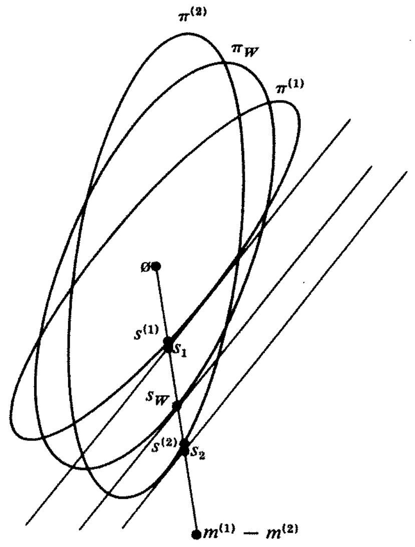

Fig. 10.2.3. Two origin-centered sample concentration ellipses and their associated pooled within sample ellipse, together with a mean difference vector and related points and tangent lines.

The ratio of the lengths of the line segments $øs_1$ and $øs^{(1)}$ is unity when case c holds, but greater than unity under case d, and similarly for $øs_2$ and $øs^{(2)}$. Consequently the ratios $øs_1/øs^{(1)}$ and $øs_2/øs^{(2)}$ may serve to indicate deviations from case c toward case d. By noting that D_1, D_2, and D are directly proportional to the lengths of $øs_1$, $øs_2$, and $øs$, and also that D, $D^{(1)}$, and $D^{(2)}$ are inversely proportional to the lengths of the line segments $øs$, $øs^{(1)}$, and $øs^{(2)}$ one may deduce that

$$K_i = (øs_i)^2/(øs^{(i)})^2 = D^{(i)^2} D_i^2/D^4 \qquad (10.2.14)$$

for $i = 1, 2$, thus relating the suggested measures K_1 and K_2 to easily computed distance measures.

10.3 EXAMPLES OF TWO-SAMPLE ANALYSIS

Example 10.1. This example is based on data collected by Tagiuri (1965) in a study of the values of research *scientists* in industry, the managers of scientists in industry called here *research managers*, and *executives* of a more traditional kind. Values are defined quantitatively as the scores obtained on the "Study of values" questionnaire of Allport, Vernon, and Lindzey (1960). Each questionnaire yields six numerical scores intended to measure the relative importance for the individual of *theoretical*, *economic*, *aesthetic*, *social*, *political*, and *religious* values. Each of the individuals in a sample from each of the three categories of individuals filled out the questionnaire in the normal way to reflect their own values. In addition, subsamples of individuals from each category filled out the questionnaire a second time to rate their conception of a typical individual in a different category. In this fashion executives rated a typical research manager, research managers rated both a typical scientist and a typical executive, and scientists rated research managers. Here, only the self-rating scores of 204 scientists and the self-rating scores of 236 research managers will be discussed but see Section 11.3 for further aspects of these data.

The Allport-Vernon-Lindzey questionnaire has the special feature that the six scores yielded by each questionnaire are constrained to have a constant sum of 240 points, implying that an elevated score on one variable is necessarily accompanied by an average depressing effect on the scores arising from the remaining variables. The same phenomenon applies to sample averages. Also, the sample covariance matrix is necessarily of less than full rank because the sum of the variables, having a constant value, necessarily has variance zero. The absence of full rank covariances causes only a minor complication for the standard multivariate analysis to be described, because the variable-space $\mathscr{E}$ may be considered from the beginning to have dimension five instead of six, and the sample covariances do have rank 5. Thus, while it is convenient to display means, variances, and covariances for all six variables, the analyses involving inversion, such as distance and best linear discriminator computations, will be carried out using only the first five variables. The choice of a variable for omission is arbitrary but has no effect on the resulting analyses which are coordinate-free in $\mathscr{E}$ and so may be based on any five linearly independent basis variables.

An examination of the basic linear and quadratic statistics yields much useful information. The mean vector for the self-rating data on the 204 scientists is

$$[51.21, 40.73, 38.11, 34.27, 40.43, 35.26],$$

and for the 236 research managers is

$$[48.97, 43.67, 37.62, 32.14, 42.11, 35.50].$$

The same pair of samples in the same order yielded sample covariance matrices

$$\begin{bmatrix} 47.7 & -1.2 & 5.8 & -7.0 & -3.2 & -42.0 \\ -1.2 & 82.7 & -38.5 & -31.2 & 8.2 & -20.0 \\ 5.8 & -38.5 & 82.1 & 1.7 & -13.8 & -37.3 \\ -7.0 & -31.2 & 1.7 & 54.0 & -14.9 & -2.5 \\ -3.2 & 8.2 & -13.8 & -14.9 & 46.1 & -22.5 \\ -42.0 & -20.0 & -37.3 & -2.5 & -22.5 & 124.4 \end{bmatrix}$$

and

$$\begin{bmatrix} 40.4 & 2.2 & -4.2 & -16.3 & 3.1 & -25.3 \\ 2.2 & 77.5 & -33.0 & -29.8 & 19.9 & -36.8 \\ -4.2 & -33.0 & 80.0 & 7.7 & -20.8 & -29.8 \\ -16.3 & -29.8 & 7.7 & 51.5 & -16.5 & 3.4 \\ 3.1 & 19.9 & -20.8 & -16.5 & 41.0 & -26.8 \\ -25.3 & -36.8 & -29.8 & 3.4 & -26.8 & 115.2 \end{bmatrix}.$$

The standard deviations computed from the two covariance matrices are respectively

$$[6.91, 9.09, 9.06, 7.35, 6.79, 11.15]$$

and

$$[6.36, 8.80, 8.94, 7.18, 6.40, 10.73].$$

Similarly the corresponding correlation matrices are

$$\begin{bmatrix} 1 & -0.019 & 0.093 & -0.138 & -0.068 & -0.546 \\ -0.019 & 1 & -0.468 & -0.467 & 0.133 & -0.197 \\ 0.093 & -0.468 & 1 & 0.026 & -0.224 & -0.369 \\ -0.138 & -0.467 & 0.026 & 1 & -0.299 & -0.031 \\ -0.068 & 0.133 & -0.224 & -0.299 & 1 & -0.281 \\ -0.546 & -0.197 & -0.369 & -0.031 & -0.281 & 1 \end{bmatrix}$$

and

$$\begin{bmatrix} 1 & 0.039 & -0.074 & -0.357 & 0.076 & -0.371 \\ 0.039 & 1 & -0.419 & -0.472 & 0.353 & -0.390 \\ -0.074 & -0.419 & 1 & 0.120 & -0.363 & -0.311 \\ -0.357 & -0.472 & 0.120 & 1 & -0.359 & 0.044 \\ 0.076 & 0.353 & -0.363 & -0.359 & 1 & -0.390 \\ -0.371 & -0.390 & -0.311 & 0.044 & -0.390 & 1 \end{bmatrix}.$$

The questionnaire is designed so that a score of 40 on each value-scale should be roughly typical. Thus both the scientists and research managers show high average scores for theoretical values, with the scientists slightly higher. The other five value scales may be discussed in similar terms. In general the self-rating average scores of the two groups differ from 40 in directions which might have been expected, and differ between groups as might have been expected from *a priori* judgments about the tastes and attitudes of the two sets of subjects.

The various distance quantities defined in Section 10.2 were computed for the chosen pair of samples. The pooled within sample covariance matrix was found to be

$$\mathbf{S}_W = \begin{bmatrix} 43.77 & 0.62 & 0.44 & -12.00 & 0.20 & -33.04 \\ 0.62 & 79.89 & -35.52 & -30.48 & 14.50 & -29.01 \\ 0.44 & -35.52 & 80.09 & 4.91 & -17.53 & -33.31 \\ -12.00 & -30.48 & 4.91 & 52.66 & -15.75 & 0.66 \\ 0.20 & 14.50 & -17.53 & -15.75 & 43.36 & -24.79 \\ -33.04 & -29.01 & -33.31 & 0.66 & -24.79 & 119.48 \end{bmatrix},$$

whose first five rows and columns have inverse

$$\begin{bmatrix} 0.0250 & 0.0030 & 0.0012 & 0.0080 & 0.0023 \\ 0.0030 & 0.0208 & 0.0086 & 0.0122 & 0.0009 \\ 0.0012 & 0.0086 & 0.0171 & 0.0054 & 0.0060 \\ 0.0080 & 0.0122 & 0.0054 & 0.0300 & 0.0090 \\ 0.0023 & 0.0009 & 0.0060 & 0.0090 & 0.0284 \end{bmatrix}.$$

The vector of mean differences is

$$[2.24, -2.94, -0.49, 2.13, -1.69, -0.24],$$

whose first five elements multiplied by the above 5 × 5 inverse covariance matrix yielded the standard best linear discriminator

$$0.06092V_1 + 0.02575V_2 + 0.01285V_3 - 0.03339V_4 + 0.02338V_5.$$

Equivalent expressions omitting a different V_i may be obtained by dropping a different row and column from the matrix inversion. Multiplying the first five elements of the mean difference vector by the vector of best linear discriminator coefficients yielded

$$D^2 = 0.318 \qquad \text{or} \qquad D = 0.564.$$

Thus the sample means on the best linear discriminator differ by 0.564 standard deviations where standard deviation is defined in terms of the pooled within sample covariance structure.

An alternative distance analysis is provided by

$$\mathbf{S}_\Delta = \begin{bmatrix} 44.30 & 0.37 & 1.17 & -11.33 & -0.26 & -34.26 \\ 0.37 & 80.28 & -35.92 & -30.59 & 13.64 & -27.79 \\ 1.17 & -35.92 & 81.15 & 4.47 & -17.02 & -33.85 \\ -11.33 & -30.59 & 4.47 & 52.84 & -15.63 & 0.23 \\ -0.26 & 13.64 & -17.02 & -15.63 & 43.74 & -24.48 \\ -34.26 & -27.79 & -33.85 & 0.23 & -24.48 & 120.15 \end{bmatrix}$$

in place of $\mathbf{S}_W$. (See 10.2.11.) The resulting best linear discriminator is

$$0.0593V_1 - 0.02679V_2 - 0.01352V_3 + 0.03142V_4 - 0.02393V_5$$

and the related distance is

$$D_\Delta^2 = 0.312 \qquad \text{or} \qquad D_\Delta = 0.559.$$

The smallness of the difference between D and D_Δ results from similarity of the two sample sizes as well as the similarity of the two sample covariance matrices.

Distances were also computed using the individual sample covariances, giving

$$D^{(1)2} = 0.286 \qquad \text{or} \qquad D^{(1)} = 0.535,$$

and

$$D^{(2)2} = 0.367 \qquad \text{or} \qquad D^{(2)} = 0.606.$$

The range from $D^{(1)}$ to $D^{(2)}$ indicates the possible range of distances resulting from different weightings of the two sample covariance matrices.

The covariance structures of the two samples are at first sight quite similar. For another look at the differences inherent in the two covariance matrices, the statistics (10.2.14) and (10.2.15) were computed. From (10.2.12)

$$D_1^2 = 0.379 \qquad \text{or} \qquad D_1 = 0.616,$$

and

$$D_2^2 = 0.286 \qquad \text{or} \qquad D_2 = 0.535.$$

Thus the ratio (10.2.13) is

$$K = (\text{ø}s_1/\text{ø}s_2)^2 = 1.325,$$

which says that the ratio of variances of the sample best linear discriminator is considerably larger than that for the original variables. Likewise the ratios (10.2.14) are

$$K_1 = (\text{ø}s_1/\text{ø}s^{(1)})^2 = 1.072,$$

and

$$K_2 = (\text{ø}s_2/\text{ø}s^{(2)})^2 = 1.048,$$

which exceed their minimum values of unity by small but significant amounts. (See Section 14.3.)

Example 10.2. The data given in this example, kindly provided by H. D. Sylwestrowicz of CIBA, are of a very common type in pharmaceutical experimentation. Many sets of such data are typically collected in routine animal experiments with drugs. The particular set considered here in isolation provides 9 measurements on each of 19 animals. The 9 variables are all measurements of renal blood pressure, but taken at intervals of half an hour over four hours. These variables will be denoted by $V_1, V_2, \ldots, V_9$ in order of time. Before the experiment, the animals had been divided randomly into two groups of sizes 12 and 7. The first group was a *control* group to which no drug was given, while the second or *treated* group received a specific drug treatment after the first of the 9 measurements had been taken. Thus differential effects between treated and control groups should not appear in V_1 but may appear in $V_2, V_3, \ldots, V_9$.

In the language of randomized experiments V_1 is usually called a *covariate*. The handling of such covariates will be discussed in some detail in Chapter 11, and the reader may wish to review this example in the light of the later discussion.

An important difference between Examples 10.1 and 10.2 is that the samples are very small in the present case. Consequently the sample means and sample covariances must be regarded with extreme caution as indicators of population means and covariances. This is especially true of the covariance structures: an unknown population covariance when $p = 9$ is determined by 36 functionally independent quantities and it is a dubious venture to attempt to estimate so many quantities from samples of sizes 7 or 12. The following analysis concentrates accordingly on mean differences.

The computations of Example 10.1, which were not described in detail, relied on straightforward subroutines for matrix addition, multiplication, and inversion. The computations of Example 10.2 rely on the SWP operator and are more efficient and informative. The original 19 × 9 data matrix was roughly centered about zero by subtracting 100.0 from each element to reduce the rounding error entering at the first sweeping operation which corrects for the grand mean. Two further columns were then added to the data matrix, representing dummy variables V_{10} taking the values zero in the control group and unity in the treated group, and V_{11} always taking the value unity. (Any three constants would serve as well.)

The resulting data matrix $\mathbf{Y}$ is reproduced below along with SWP[11, 10]$\mathbf{Y}'\mathbf{Y}$ and $-$SWP[1, 2, . . . , 9, 11, 10]$\mathbf{Y}'\mathbf{Y}$.

$$\mathbf{Y} = \begin{bmatrix}
17 & 27 & 17 & 17 & 25 & 25 & 25 & 15 & 17 & 0 & 1 \\
5 & 5 & 2 & 2 & 5 & 10 & 10 & 12 & 12 & 0 & 1 \\
20 & 20 & 20 & 20 & 18 & 17 & 17 & 17 & 15 & 0 & 1 \\
8 & 17 & 8 & 15 & 25 & 25 & 25 & 25 & 27 & 0 & 1 \\
22 & 22 & 20 & 20 & 15 & 12 & 18 & 13 & 12 & 0 & 1 \\
13 & 17 & 17 & 12 & 17 & 17 & 17 & 17 & 7 & 0 & 1 \\
35 & 23 & 25 & 23 & 28 & 27 & 42 & 42 & 30 & 0 & 1 \\
45 & 43 & 37 & 33 & 35 & 35 & 33 & 32 & 30 & 0 & 1 \\
2 & 5 & 2 & -5 & -7 & -10 & -8 & -8 & -18 & 0 & 1 \\
33 & 37 & 22 & 28 & 32 & 30 & 30 & 27 & 28 & 0 & 1 \\
25 & 35 & 22 & 28 & 28 & 30 & 28 & 25 & 22 & 0 & 1 \\
32 & 47 & 48 & 47 & 47 & 47 & 47 & 48 & 47 & 0 & 1 \\
45 & -2 & 2 & 0 & -5 & -5 & -10 & -10 & -12 & 1 & 1 \\
-3 & -27 & -30 & -33 & -35 & -35 & -33 & -33 & -33 & 1 & 1 \\
32 & 17 & 12 & 12 & 7 & 2 & 2 & 7 & 7 & 1 & 1 \\
30 & -2 & -10 & -12 & -12 & -12 & -12 & -13 & -13 & 1 & 1 \\
13 & -20 & -22 & -22 & -23 & -27 & -27 & -28 & -28 & 1 & 1 \\
20 & 18 & 2 & -13 & -18 & -18 & -22 & -22 & -23 & 1 & 1 \\
22 & 18 & 8 & -8 & -10 & -8 & -7 & -2 & 0 & 1 & 1
\end{bmatrix}$$

$$\text{SWP}[11, 10]\mathbf{Y}'\mathbf{Y} = \begin{bmatrix}
3318.3453 & 2618.4048 & 2697.1429 & 2810.2858 & 2667.9048 & 2605.1548 & 2566.5239 & 2561.7262 & 2543.4406 & 1.2976 & 21.4167 \\
2618.4048 & 4095.0954 & 3450.8572 & 3158.7144 & 3017.0954 & 3021.5954 & 2784.4764 & 2869.0239 & 3114.3096 & -24.5476 & 24.8333 \\
2697.1429 & 3450.8572 & 3409.7144 & 3126.4286 & 2893.8572 & 2858.8572 & 2757.2859 & 2925.7144 & 2984.2858 & -25.4286 & 20.0000 \\
2810.2858 & 3158.7144 & 3126.4286 & 3350.8573 & 3197.7145 & 3115.7145 & 2981.5716 & 3082.4287 & 3277.5716 & -30.8571 & 20.0000 \\
2667.9048 & 3017.0953 & 2893.8572 & 3197.7145 & 3282.0955 & 3237.0954 & 3117.4764 & 3188.5240 & 3430.8097 & -36.0476 & 22.3333 \\
2605.1548 & 3021.5954 & 2858.8572 & 3115.7144 & 3237.0953 & 3282.3454 & 3120.4763 & 3208.7739 & 3484.0596 & -36.7976 & 22.0833 \\
2566.5239 & 2784.4763 & 2757.2858 & 2981.5716 & 3117.4764 & 3120.4763 & 3242.3812 & 3350.6192 & 3502.0477 & -39.2381 & 23.6667 \\
2561.7262 & 2869.0239 & 2925.7144 & 3082.4287 & 3188.5240 & 3208.7740 & 3350.6193 & 3620.6311 & 3718.2025 & -36.5119 & 22.0833 \\
2543.4405 & 3114.3096 & 2984.2858 & 3277.5716 & 3430.8097 & 3483.0596 & 3502.0479 & 3718.2025 & 4088.6310 & -33.6548 & 19.0833 \\
1.2976 & -24.5476 & -25.4286 & -30.8571 & -36.0476 & -36.7976 & -39.2381 & -36.5119 & -33.6548 & -0.2262 & 0.0833 \\
21.4167 & 24.8333 & 20.0000 & 20.0000 & 22.3333 & 22.0833 & 23.6667 & 22.0833 & 19.0833 & 0.0833 & -0.0833
\end{bmatrix}$$

$$-\text{SWP}[1, 2, \ldots, 9, 11, 10]\mathbf{Y}'\mathbf{Y} = \begin{bmatrix}
0.0015 & 0.0001 & -0.0001 & -0.0015 & 0.0012 & -0.0011 & -0.0017 & 0.0001 & 0.0015 & -0.0552 & 0.0026 \\
0.0001 & 0.0025 & -0.0033 & 0.0016 & -0.0013 & -0.0003 & -0.0003 & 0.0024 & -0.0014 & -0.0028 & -0.0136 \\
-0.0001 & -0.0033 & 0.0072 & -0.0059 & 0.0050 & -0.0017 & 0.0006 & -0.0048 & 0.0032 & -0.0085 & 0.0162 \\
-0.0015 & 0.0016 & -0.0059 & 0.0127 & -0.0140 & 0.0054 & 0.0034 & 0.0019 & -0.0036 & 0.0560 & -0.0030 \\
0.0012 & -0.0013 & 0.0050 & -0.0140 & 0.0294 & -0.0168 & -0.0079 & 0.0004 & 0.0039 & -0.0636 & 0.0051 \\
-0.0011 & -0.0003 & -0.0017 & 0.0054 & -0.0168 & 0.0176 & 0.0009 & 0.0017 & -0.0053 & 0.0756 & -0.0112 \\
-0.0017 & -0.0003 & 0.0006 & 0.0034 & -0.0079 & 0.0009 & 0.0151 & -0.0086 & -0.0012 & 0.1040 & -0.0265 \\
0.0001 & 0.0024 & -0.0048 & 0.0019 & 0.0004 & 0.0017 & -0.0086 & 0.0127 & -0.0058 & -0.0005 & -0.0146 \\
0.0015 & -0.0014 & 0.0032 & -0.0036 & 0.0039 & -0.0053 & -0.0012 & -0.0058 & 0.0085 & -0.0937 & 0.0359 \\
-0.0552 & -0.0028 & -0.0085 & 0.0560 & -0.0636 & 0.0756 & 0.1040 & -0.0005 & -0.0937 & 3.1399 & -0.6940 \\
0.0026 & -0.0136 & 0.0162 & -0.0030 & 0.0051 & -0.0112 & -0.0265 & -0.0146 & 0.0359 & -0.6940 & 0.4998
\end{bmatrix}$$

The first 9 rows and columns provide $\mathbf{T}_W$ for the original $V_1, V_2, \ldots, V_9$ of SWP[11, 10]Y′Y. The first 9 elements of row 10 provide $\bar{\mathbf{X}}^{(2)} - \bar{\mathbf{X}}^{(1)}$ and the (10, 10) element is $-(1/n_1 + 1/n_2) = -\frac{1}{12} - \frac{1}{7} = -0.2262$. The stage is now set for distance and best linear discriminator analyses based on the inner product $\mathbf{T}_W$. The same best linear discriminators are valid for $\mathbf{S}_W$ while distances according to $\mathbf{T}_W$ should be multiplied by the degrees of freedom $n_1 + n_2 - 2 = 17$ to provide the more natural distances according to $\mathbf{S}_W$.

After applying SWP[1, 2, . . . , 9] and changing signs, the first 9 elements of row 10 provide the sample best linear discriminator while the (10, 10) element is $0.2262 + D^2$. Thus, the D^2 defined by $\mathbf{T}_W$ is 2.9137 or the more usual D^2 defined by $\mathbf{S}_W$ is 49.533.

Table 10.3.1

	Coefficient of								
s	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9
2	−10.3	12.6							
3	−18.6	−5.4	27.6						
4	−28.9	−7.0	4.7	35.6					
5	−29.5	−12.3	16.3	−1.5	33.3				
6	−29.5	−12.8	15.4	2.0	19.4	11.9			
7	−30.5	−8.3	8.0	11.9	1.9	3.4	20.4		
8	−38.9	−18.7	27.2	16.3	−20.5	17.6	90.3	−64.6	
9	−55.2	−2.8	−8.5	56.0	−63.6	75.6	104.0	−0.5	−93.7

Of course, it costs little along the way to look successively at SWP[1], SWP[1, 2], . . . , SWP[1, 2, . . . , 8] which provide the same analyses for the subsets V_1, $[V_1, V_2]$, . . . , $[V_1, V_2, \ldots, V_8]$. The weights for the best linear discriminators based on $V_1, V_2, \ldots, V_s$ are summarized in Table 10.3.1, after scaling by the arbitrary factor of 100.

In a similar fashion the successive values of D^2 for the spaces spanned by V_1, $[V_1, V_2]$, . . . , $[V_1, V_2, \ldots, V_9]$ are easily found, as shown in Table 10.3.2.

Many other decompositions of D^2 are possible. For example, each of the 9! orders of applying SWP[1], SWP[2], . . . , SWP[9] leads to a different decomposition. The given order following time is perhaps the most natural in the present context, since each successive contribution to D^2 may be regarded as a treatment effect associated with the corresponding increase of observation time. Note that the distance associated with V_1 is small, which is consistent with V_1 being a covariate. The large contributions to D^2 associated with V_8 and V_9 are rather surprising, and will be discussed later.

For comparative purposes, and to continue an empirical investigation of principal component analysis, a decomposition of D^2 based on principal variables was also computed. The principal component analysis used the total sample mean-corrected inner product matrix relative to the identity matrix as a reference inner product matrix. The computations proceeded as above, except that the operations SWP[11], SWP[10] were replaced by SWP[11], SDG[1, 2, . . . , 9], SWP[10]. The decomposition of D^2 related to the principal variables $V_1, V_2, \ldots, V_9$ is displayed in Table 10.3.3.

Table 10.3.2

s	$D(V_1, V_2, \ldots, V_s)^2$	$D(V_1, \ldots, V_s)^2 - D(V_1, \ldots, V_{s-1})^2$
1	0.009	0.009
2	5.488	5.479
3	10.097	4.609
4	18.434	8.337
5	22.198	3.764
6	22.380	0.182
7	23.893	1.513
8	31.992	8.099
9	49.534	17.542

The contributions to D^2 do appear somewhat earlier in this table than in the preceding table, but the large contribution from U_6 appears suspicious. Incidentally, the principal components of variance or, more properly, the principal components of the total sum inner product corrected for the grand mean are 65164.9, 3186.9, 1398.6, 446.0, 344.7, 190.5, 76.5, 39.5, and 21.7 which show a very rapid drop-off.

One further decomposition of D^2 was computed, based essentially on the idea of fitting polynomials of increasing order in time, i.e., linear, quadratic,

Table 10.3.3

s	$D(U_1, U_2, \ldots, U_s)^2$	$D(U_1, U_2, \ldots, U_s)^2 - D(U_1, \ldots, U_{s-1})^2$
1	5.532	5.532
2	15.511	9.979
3	17.964	2.453
4	25.305	7.341
5	29.059	3.754
6	43.388	14.329
7	43.692	0.304
8	44.216	0.524
9	49.534	5.318

cubic, etc. The variables $V_1, V_2, \ldots, V_9$ were first replaced by V_1, $V_2 - V_1$, $V_3 - V_2, \ldots, V_9 - V_8$ and the 8 differences were replaced by

$$\begin{bmatrix} W_2 \\ W_3 \\ \cdot \\ \cdot \\ \cdot \\ W_9 \end{bmatrix} = \mathbf{K} \begin{bmatrix} V_2 - V_1 \\ V_3 - V_2 \\ \cdot \\ \cdot \\ \cdot \\ V_9 - V_8 \end{bmatrix},$$

where $\mathbf{K}$ is an orthogonal matrix whose rows are successively constant, linear, quadratic, etc. The matrix $\mathbf{K}$ was computed by starting from

$$\begin{bmatrix} 1 & 1 & 1 \ldots 1 \\ 1 & 2 & 3 \ldots 8 \\ 1^2 & 2^2 & 3^2 \ldots 8^2 \\ \cdot & & \\ \cdot & & \\ \cdot & & \\ 1^7 & 2^7 & 3^7 \ldots 8^7 \end{bmatrix},$$

successively orthogonalizing its rows, and finally reducing them to unit length. These rows of $\mathbf{K}$ are sometimes called *orthogonal polynomials.* The raw sum inner product matrix for $V_1, V_2, \ldots, V_{11}$ was successively transformed into that for $V_1, V_2 - V_1, \ldots, V_9 - V_8, V_{10}, V_{11}$ and then that for $V_1, W_2, W_3, \ldots, W_9, V_{10}, V_{11}$. From there, the computations proceeded as in the original analysis of $V_1, V_2, \ldots, V_9$, i.e., SWP[11], SWP[10], SWP[1], SWP[2], . . . , SWP[9] were successively applied. The resulting decomposition of D^2 is given in Table 10.3.4.

The sampling distributions of the various sample statistics computed above are discussed in Section 14.3, on the assumption of random samples from multivariate normal populations. These distributions make possible certain postdictive inference procedures which provide uncertain information about the parameters of the normal population. In particular, significance tests are applied in Section 14.3 to the three sequences of increments of D^2 values given in Tables 10.3.2, 10.3.3, and 10.3.4.

Table 10.3.4

s	$D(V_1, W_2, \ldots, W_s)^2$	$D(V_1, W_2, \ldots, W_s)^2 - D(V_1, W_2, \ldots, W_{s-1})^2$
1	0.009	0.009
2	9.549	9.540
3	23.105	13.556
4	40.973	17.869
5	41.648	0.675
6	42.928	1.280
7	45.308	2.380
8	46.551	1.243
9	49.533	2.912

In connection with sampling theory considerations, it is sometimes argued that the stepwise procedures illustrated above should be applied in an order which takes first those variables which may be assumed to have zero population differences in their means. Thus, in the principal variable analysis one might think of reversing the order of the sweeping operations, while in the polynomial fitting analysis one might retain the initial time variable in first position while reversing the order of the rest to deal first with higher order effects. The argument for these reversals is that the variables with zero population differences act as covariates and so increase the sensitivity of tests and estimation procedures reflecting on the variables which really matter. A counter-argument is that it can never really be known that higher order population differences are not present, and their presence tends to bias the tests relating to the important variables in an unknown way. A second counter-argument is that the use of covariates consumes degrees of freedom available for covariance estimation and thereby decreases the sensitivity of the tests. The second counter-argument is quite important with these data, since the samples are very small. The reader with adequate computing facilities may wish to carry out the analyses using the alternative orders, including the significance tests based on the decomposition of D^2, for indications of a loss of sensitivity to important effects.

Polynomial fitting has a considerable literature of its own, and many variations of the above techniques have been suggested. For example, Rao (1965) has proposed applications of the orthogonal polynomials of one higher degree to the original 9 variables, instead of the application as above to the 8 time-difference variables. The latter was preferred here on the belief that the time-difference variables are more likely to be roughly uncorrelated with constant variance than are the original variables; for sampling theory considerations suggest that optimal fitting without the use of covariates is provided by orthogonal polynomials when the basic variables are uncorrelated with constant variance (in the population). Potthoff and Roy (1964) describe formally the correct set of polynomials to use in the presence of arbitrary population covariance structure. Rao (1965) also presents methods appropriate to the (artificial?) assumption that each animal corresponds to a random polynomial of given degree, while all observations are subject to independent errors of measurement with zero means and constant variance.

10.4 ANALYSIS OF MORE THAN TWO SAMPLES

The two sample theory of Section 10.2 suggests generalizations of many kinds, some of which are explained here. The emphasis will be put on the comparison of means. Techniques for comparing more than two sample means at once will be described, but the development of techniques for simultaneous comparison of more than two sample covariances will not be attempted.

A set of k sample means $m^{(1)}, m^{(2)}, \ldots, m^{(k)}$ in $\mathscr{F}$ may be considered in two ways, first without regard for any inner product and second in relation to an inner product π_d over $\mathscr{F}$. The first of these approaches will not be pursued in detail except for one comment. If $k - 1 \leq p$, then any set of k points not lying in a hyperplane of dimension $k - 2$ is affinely like any other such set, for two such sets are related by a wide sense linear transformation. On the other hand, if $k - 1 > p$, then the k points necessarily lie in a hyperplane of dimension less than $k - 1$ and so begin to have a distinctive pattern independent of any inner product. To take an extreme example, if $p = 1$ and $k \geq 3$, then the k mean points on a line do form a meaningful pattern. Such patterns are less easy to view in an affine way when $p > 1$ and are difficult to view at all when $p > 3$. No general analysis of these patterns is attempted here.

Henceforth it will be assumed that a set of k means is to be viewed in relation to an inner product π_d over $\mathscr{F}$. Later, the choice of π_d from k sample data will be discussed briefly. Having π_d one has distances defined between each of $k(k - 1)/2$ pairs of means and also a best linear discriminator for each pair of samples. The subspace of variable-space $\mathscr{E}$ spanned by this set of $k(k - 1)/2$ best linear discriminators will be called the *space of best linear discriminators among the k means.* This space is a natural generalization of the one-dimensional space of best linear discriminators defined when $k = 2$, and its properties are set out in the following generalization of Theorem 10.2.

Theorem 10.4. *The space $\mathscr{V}$ of best linear discriminators among $m^{(1)}$, $m^{(2)}, \ldots, m^{(k)}$ determined by the rank p inner product π_d over $\mathscr{F}$ is the dual space of the orthogonal complement of the subspace of $\mathscr{F}$ spanned by the differences among $m^{(1)}, m^{(2)}, \ldots, m^{(k)}$. If W is any variable in $\mathscr{E}$ and V is the orthogonal projection of W into $\mathscr{V}$, then*

$$D_{ij}(W) \leq D_{ij}(V) \qquad (10.4.1)$$

for i and $j = 1, 2, \ldots, k$, where D_{ij} denotes distance between $m^{(i)}$ and $m^{(j)}$ as defined in Section 10.2.

Theorem 10.4 gives first a characterization of $\mathscr{V}$ and second the basic property (10.4.1) which shows that only variables inside $\mathscr{V}$ need be considered for discrimination, at least in the sense that for any variable outside $\mathscr{V}$ there is another inside $\mathscr{V}$ yielding greater distances uniformly over all pairs of samples.

The proof of Theorem 10.4 requires only minor extensions of the theory of Section 10.2. The best linear discriminator V_{ij} between the pair of samples i and j defines a one-dimensional subspace $\mathscr{V}_{ij}$ whose dual is a $(p - 1)$-dimensional subspace $\mathscr{V}_{ijd}$ in $\mathscr{F}$ orthogonal to the one-dimensional subspace $\mathscr{U}_{ijd}$ spanned by $m^{(i)} - m^{(j)}$. $\mathscr{V}$ is the direct sum of the $\mathscr{V}_{ij}$, so that $\mathscr{V}_d$ is the intersection of the $\mathscr{V}_{ijd}$ and the orthogonal complement $\mathscr{U}_d$ of $\mathscr{V}_d$ is the direct sum of the $\mathscr{U}_{ijd}$, as required for the first part of Theorem 10.4. The second part of the theorem is a consequence of the fact that the sample means are identical on the subspace $\mathscr{U}$ orthogonal to $\mathscr{V}$, so that W and its orthogonal projection V into $\mathscr{V}$ have identical mean differences. In terms of the expression (10.2.2) for distance, both $D_{ij}(W)$ and $D_{ij}(V)$ have the same numerators while $(V, V) \leq (W, W)$, and (10.4.1) follows.

The dimension f of the space $\mathscr{V}$ of best linear discriminators is the same as the dimension of $\mathscr{U}_d$ or $m^{(1)} + \mathscr{U}_d$, the latter being the smallest hyperplane containing $m^{(1)}, m^{(2)}, \ldots, m^{(k)}$. Clearly $f \leq \min(k - 1, p)$. In general $f = p$ if $k - 1 \geq p$, and in this case $\mathscr{V} = \mathscr{E}$ so that no reduction occurs, i.e., the concept of a space of best linear discriminators contributes nothing. If $k - 1 < p$, however, a simplification does result from the concept of $\mathscr{V}$.

The notion of distance does not generalize so unambiguously as the notion of best linear discriminator. Consider first generalized distance for a single variable V, the aim being to define a measure reflecting separation among all k means which reduces to $D(V)$ when $k = 2$. If the k mean values of V are denoted by $\bar{X}^{(1)}, \bar{X}^{(2)}, \ldots, \bar{X}^{(k)}$, then any nonnegative quadratic form in the differences $\bar{X}^{(i)} - \bar{X}^{(j)}$ might be chosen as a measure of these differences, examples being $\sum_1^k n_i(\bar{X}^{(i)} - \bar{X})^2$ where $\bar{X} = \sum_1^k n_i\bar{X}^{(i)}/n$, or $\sum_1^k (\bar{X}^{(i)} - \bar{X}_G)^2$ where $\bar{X}_G = \sum_1^k \bar{X}^{(i)}/k$. But other choices are possible. If a particular *contrast* $\sum_1^k a_i\bar{X}^{(i)}$ with $\sum_1^k a_i = 0$ were the main source of interest, then $(\sum_1^k a_i\bar{X}^{(i)})^2$ would be a suitable distance measure. Or, more generally, a weighted linear combination of several such squared contrasts could be of interest. Any such quadratic form may be taken as a generalization of $D(V)^2$ for a given V. It would be preferable, however, to divide by (V, V) so that the generalized measure would share with $D(V)^2$ the property of having a common value over the one-dimensional subspace $\mathscr{V}$ spanned by V.

Given such a generalized $D(V) = D(\mathscr{V})$, the next step is to extend the definition to a corresponding generalization of $D(\mathscr{V})$ where $\mathscr{V}$ has dimension greater than unity. The straightforward extension of the above quadratic form concept is an inner product concept. For example, $\sum_1^k n_i(X^{(i)} - \bar{X})^2$ generalizes to the inner product defined by $\mathbf{T}_A = \sum_1^k n_i(\bar{\mathbf{X}}^{(i)} - \bar{\mathbf{X}})'(\bar{\mathbf{X}}^{(i)} - \bar{\mathbf{X}})$. "Division" of this inner product by π is generalized to the eigenvector and eigenvalue structure of the quadratic form inner product relative to π. Thus the generalized $D(V)^2$ extends to a set of eigenvalues which are themselves generalized $D(V)^2$ for a set of eigenvectors.

Sometimes a single best linear discriminator might be sought, or a single generalized distance measure. The first of these could be taken to be the eigenvector corresponding to the largest generalized distance eigenvalue. Any such eigenvector must be in the space $\mathscr{V}$ of best linear discriminators defined above, as, indeed, must any set of such eigenvectors, from (10.4.1). Similarly, a single function of the generalized distance eigenvalues could be chosen as a single generalized distance, for example, the sum of the eigenvalues. Note, however, that no single discriminator or single distance measure can tell the whole story. Indeed, there is a very strong flavor of arbitrariness about all of the analyses described in this section, since so many choices are open. Only a few guidelines can be given.

The example of Section 10.5 uses $\mathbf{T}_W$ or $\mathbf{S}_W$ to define the inner product π. The generalized distance inner product is the familiar analysis of variance choice $\mathbf{T}_A$, but several decompositions of $\mathbf{T}_A$ are considered as well.

Eigenvector and eigenvalue analyses of the kind described above are often called *multiple discriminant analyses*. The eigenvectors determine variables in the space of best linear discriminators which are often called *discriminants*.

10.5 AN EXAMPLE WITH SIX SAMPLES CROSS-CLASSIFIED INTO TWO SEXES AND THREE RACES

Example 10.3. The data of this example are based on 276 human skulls assembled and measured by Prof. W. W. Howells. The individuals were classified at the start into 6 groups as in Table 10.5.1.

Table 10.5.1

THE DISTRIBUTION OF SKULLS BY RACE AND SEX FOR EXAMPLE 10.3

	Male	Female	Total
Japanese	56	35	91
Ainu	57	55	112
Australian	38	35	73
Total	151	125	276

The Japanese and Ainus represent racial groups now living in Japan, the Ainus being descendents of one or more of the tribes which occupied the Japanese islands before the arrival of the long-dominant Japanese. The Australian racial group represents the darkskinned aboriginals who predated European immigration. It was conjectured that the present analysis might demonstrate affinities between the Australian and Ainu groups and thus support a hypothesis of a common source in prehistoric times, as has sometimes been suggested. In fact, however, the analysis indicates closer relations between the Japanese and the Ainus than between either of these and the Australians.

The analysis is based on 21 variables representing physical dimensions considered to encompass much of the important variation in human skulls, i.e., based on expert opinion of what it is that makes skulls look different. The technical definitions of these 21 variables are omitted from the present discussion, as are any interpretations of the data based on the names and meanings of the variables. A few of the values making up the 276 × 21 data matrix could not be measured directly because of incomplete skulls. These were filled in according to a professional guess as to what the complete skull would have been like. Other analyses of these data have been reported by Howells (1966).

Several analyses were carried out, only one of which is reported here. Originally, six racial groups were used, including three African groups. The latter set were later dropped since they had been measured by a different investigator and there was some evidence that investigator biases were confounded with actual physical differences between the two sets of races. Also, the analyses were done both on the original measurements and on their logarithms, but only the former is reported here because the latter is virtually indistinguishable in its outcome.

Table 10.5.2

ARTIFICIAL VARIABLES FOR EXAMPLE 10.3

	V_0	S	R_1	R_2	I_1	I_2
Japanese male	1	−1	−1	−1	1	1
Ainu male	1	−1	1	0	−1	0
Australian male	1	−1	0	1	0	−1
Japanese female	1	1	−1	−1	−1	−1
Ainu female	1	1	1	0	1	0
Australian female	1	1	0	1	0	1

Most of the following discussion is concerned with the technical details of the version of multiple discriminant analysis which was carried out. As usual, artificial variables were used to build group identifications into the data matrix. These were chosen in a special way in order that the five degrees of freedom for among group variation could be easily decomposed into a single degree of freedom for sex differences, two degrees of freedom for race differences, and two degrees of freedom for race by sex interaction. This type of decomposition is familiar to the user of a two-way "row by column" analysis of variance, except that it is carried out here for 21 variables and all of their linear combinations. Also, there is a confounding difficulty due to unequal sample sizes.

Six artificial variables were added to the original 21. These variables take the same values on all the individuals of a given group, and their values on the six groups are shown in Table 10.5.2. The symbols S, R_1, R_2, I_1, and I_2 may be regarded as abbreviations for *sex dummy*, *first race dummy*, *second race dummy*, *first interaction dummy*, and *second interaction dummy*, respectively. These variables take the place of the indicator variables of the six groups used in the original discussion of multivariate analysis of variance in Section 9.4. For the purpose of ordinary six sample multiple discriminant analysis the above set of six dummy variables is equivalent to the choice of V_0 together with any five of the six group indicator variables; the basic requirement is that both sets include V_0 and span the same six-dimensional subspace. But the present set is much more convenient for computing a decomposition into sex, race, and race by sex interaction components.

The selected dummy variables have the property that, had the sample sizes been equal, the four subspaces spanned in $\mathcal{N}$ by the representatives of V_0, S, $[R_1, R_2]$, $[I_1, I_2]$ would have been orthogonal. The difficulty caused by the lack of balanced sample sizes is not computational only but also conceptual, for it is no longer clear how to break the six-dimensional subspace of $\mathcal{N}$ spanned by the dummy variables into a direct sum of four orthogonal subspaces labelled for grand mean, sex differences, race differences, and race by sex interactions. Various decompositions are possible depending on different orders of applying successive orthogonalization to the dummy variables. For example, suppose that V_0, S', $[R_1', R_2']$, $[I_1', I_2']$ represent the set V_0, S, $[R_1, R_2]$, $[I_1, I_2]$ after successive orthogonalization in the stated order, according to the sample raw sum inner product. Then the subspaces of $\mathcal{N}$ spanned by the representatives of V_0, S', $[R_1', R_2']$ and $[I_1', I_2']$ are orthogonal and could be labelled for grand mean, sex differences, race differences, and race by sex interactions, respectively. On the other hand, if the orthogonalization is carried out in the order V_0, $[R_1, R_2]$, S, $[I_1, I_2]$, then the resulting V_0, $[R_1'', R_2'']$, S'', $[I_1', I_2']$ define alternative candidates for the orthogonal subspaces to be associated with race differences and sex differences.

Several other orders of orthogonalization each beginning with V_0 are possible, but the orders selected above are perhaps most natural on the grounds that simpler main effects should be hypothesized before more complicated interaction effects. In any case, only the selected pair of orders is carried along in this example to illustrate the fact of confounding of sex main effects and race main effects. The dilemma posed by such *confounding* may be clarified as follows.

The sex differences measured by components along S' in $\mathcal{N}$ are free of any constant addition to all of the data on a given variable, but they are not free of additions which are constant only for a given race. On the other hand, components along S'' are unaffected by systematic race differences. The difficulty with S'' is that, when sample sizes are unequal, valid sex differences necessarily contribute apparent race differences so that the use of S'' in place of S' eliminates valid as well as spurious sex differences.

The possible extent of this type of confounding may be judged by looking at the angles among the dummy variables in $\mathcal{N}$. It is assumed by convention that V_0 components should be removed first. The cosines of the angles among

the components of S, R_1, R_2, I_1, I_2 are given by the total sample correlation matrix

$$\begin{bmatrix} 1.0000 & 0.0894 & 0.0773 & 0.0970 & 0.0773 \\ 0.0894 & 1.0000 & 0.5101 & -0.1213 & -0.1213 \\ 0.0773 & 0.5101 & 1.0000 & -0.1090 & -0.1402 \\ 0.0970 & -0.1213 & -0.1090 & 1.0000 & 0.4953 \\ 0.0773 & -0.1213 & -0.1402 & 0.4953 & 1.0000 \end{bmatrix}$$

which was obtained as part of the output of the computations to be described shortly. The square of the multiple correlation coefficient between S and $[R_1, R_2]$ is 0.0083, the fraction of the squared components associated with sex difference which could conceivably be falsely eliminated when S' is replaced by S''. By such reasoning it is clear that confounding effects are necessarily rather slight between the major sources of variation associated with sex and race. Nevertheless, the computations were performed assigning the doubtful race and sex components in both ways.

Multiple discriminant analysis was carried out with the pooled within sample covariance inner product playing the role of π in Section 10.4 and with seven different choices of a numerator inner product reflecting differences among the sample means. The numerator inner products were in each case the inner products of the sample representation in a subspace of $\mathcal{N}$ of the data vectors of 21 observed variables, the seven subspaces of $\mathcal{N}$ being those spanned by S', $[R_1', R_2']$, $[R_1'', R_2'']$, S'', $[S', R_1', R_2']$, $[I_1', I_2']$, and $[S', R_1', R_2', I_1', I_2']$. These analyses produced one eigenvector or discriminant variable for each dimension, i.e., $1 + 2 + 2 + 1 + 3 + 2 + 5 = 16$ variables altogether. Of course, all of these variables lie in a single five-dimensional space of best linear discriminants, and therefore must exhibit substantial correlations which will be displayed shortly.

The computing steps were as follows. Basic linear and quadratic statistics were computed for the total sample of 276 using a program which gave a vector of means, a vector of standard deviations, and a correlation matrix for the 26 variables consisting of S, R_1, R_2, I_1, I_2 and the 21 original variables. A more convenient starting point would appear to be simply the total sum of products matrix corrected for the grand mean, but the analysis can be carried out in terms of a rescaled basis and the correlation matrix described above is in fact a total sum of products matrix corrected for the grand mean for a basis in which each of the original elements is divided by $\sqrt{275}$ times its standard deviation. To compensate for the nuisance of an additional transformation, the scaling provides an array of numbers more alike in order of magnitude and hence easier to scan by eye.

Subsequent steps were based on the correlation matrix (*cum* total sum of products matrix corrected for the grand mean) as follows. First the operation SWP[1, 2, 3, 4, 5] removed components along S, R_1, R_2, I_1, I_2 so that the lower right 21 × 21 part represented a pooled within sample sum inner product for the rescaled variables, i.e., the inner product which plays the role of π in Section 10.4 and which will be called here the *denominator inner product*. Recall that π produces the same discriminant variables as does the pooled within sample covariance which differs only by a scale factor.

Secondly, the operator MST[6, 7, . . . , 26] was applied to the output of the first step. The second input matrix of the MST operator was a diagonal matrix whose elements were the scale factors $(\sqrt{275}$ standard deviation$)^{-1}$ which express the scaled basis in terms of the original basis of variables. The MST operator effects a transformation of the first input matrix so that it refers to a basis whose last 21 elements are orthonormal with respect to the denominator inner product; achieving the orthonormal basis is always a preliminary to an eigenvector calculation. The first output matrix of the MST operator is denoted here by **Q**. The 21 × 21 part of the second MST output matrix expresses the orthonormal basis in terms of the original 21 measured variables.

The next step is to find the $\mathbf{T}_A$ matrix and its various decompositions which are called here numerator inner product matrices. These are all expressed in terms of the basis produced by the MST operator. They are found by RSW operations, which add back to **Q** various removed components, followed by subtractions to isolate these added back components. Specifically

RSW[4, 5]**Q**,
RSW[4, 5, 2, 3]**Q**,
RSW[4, 5, 1]**Q**, and
RSW[4, 5, 1, 2, 3]**Q**

were found, and thence

RSW[4, 5, 1, 2, 3]**Q** − RSW[4, 5, 2, 3]**Q**,
RSW[4, 5, 2, 3]**Q** − RSW[4, 5]**Q**,
RSW[4, 5, 1, 2, 3]**Q** − RSW[4, 5, 1]**Q**,
RSW[4, 5, 1]**Q** − RSW[4, 5]**Q**,
RSW[4, 5, 1, 2, 3]**Q** − RSW[4, 5]**Q**,
RSW[4, 5]**Q** − **Q**, and
RSW[4, 5, 1, 2, 3]**Q** − **Q**,

which provided, respectively, the numerator inner product matrices relative to the orthonormal basis, associated with

Sex, unadjusted for race,
Race, adjusted for sex,
Race, unadjusted for sex,
Sex, adjusted for race,
Sex and race together,
Race by sex interaction, adjusted for sex and race, and
Race, sex, and race by sex interaction all together.

The SDG operator was applied to the 21 × 21 parts of each of these numerator inner product matrices yielding the following nonzero eigenvalues:

2.4184,
2.6015, 0.7714,
2.6258, 0.7396,
2.4260,
2.8627, 2.1958, 0.7328,
0.1570, 0.0671, and
2.9037, 2.2021, 0.7476, 0.1017, 0.0605.

Some interpretations of these eigenvalues will be reported in Section 14.3.

The second input matrix for each of these SDG operations was the second output matrix of the MST operation. Thus the second output matrix of each SDG operation produces eigenvectors expressed as linear combinations of the original 21 variables. Since these eigenvectors have unit norms according to the pooled within sample sum inner product, they were multiplied by $\sqrt{270}$ so they finally had unit variances according to the pooled within sample covariance inner product. Altogether, 7 sets of eigenvectors were found, including 16 discriminant variables.

The eigenvectors were then applied to the 276 × 21 data matrix to produce the 276 × 16 data matrix of the 16 discriminants. The pooled within sample covariance matrix for the 16 discriminants, as directly computed from the 276 × 16 data matrix, is shown in Table 10.5.3.

Since Table 10.5.3 was computed as a sample covariance matrix, the fact that the diagonal elements are very close to unity serves as a check on the intended normalization of the discriminant variables. Likewise the zero covariances within the blocks of 1, 2, 2, 1, 3, 2, 5 variables in order are checks on these theoretically intended zero covariances.

Very high correlations may be observed in positions (6, 1), (4, 2), (5, 3), indicating that the two approaches to sex and race discriminants produced almost identical results. Other high correlations may be observed in places where they make sense. For example, discriminants 1, 2, 3, 11, 12 provide one basis of the five-dimensional space of best linear discriminators while discriminants 12, 13, 14, 15, 16 provide another. The elements of the first basis have direct relations to sex (1), race (2 and 3), and interaction (11 and 12), but the first basis is not orthonormal. The second basis is the one provided by the straightforward analysis described in Section 10.4 and this second basis is orthonormal. The correlation matrix above shows a rough pairing of these discriminators in the orders 1, 2, 3, 11, 12, and 13, 12, 14, 15, 16 but a better matching of 1 and 2 could be provided by a pair of linear combinations of 12 and 13. Thus it appears that the general discriminators 12 through 16 could reasonably be labelled for race, sex, and interaction effects, but that sharper delineations of these effects are provided by the decomposition analyses.

Table 10.5.3

1.000	−.131	.183	−.166	−.029	.998	−.675	.773	.084	.174	.017	−.657	.749	.086	−.003	−.002
−.131	1.000	.000	.999	.004	−.092	.819	.573	.014	−.519	−.122	.832	.554	.005	.026	.003
.183	.000	1.000	−.001	−.988	.233	−.143	.228	−.963	.246	−.272	−.195	.228	−.961	−.043	.026
−.166	.999	−.001	1.000	−.000	−.127	.839	.545	.006	−.521	−.123	.851	.525	−.004	.025	.003
−.029	.004	−.988	−.000	1.000	−.079	.058	−.100	.993	−.235	.276	.100	−.098	.992	.044	−.026
.998	−.092	.233	−.127	−.079	1.000	−.646	.762	.035	.166	−.002	−.628	.777	.038	−.004	−.000
−.675	.819	−.143	.839	.058	−.646	1.000	.000	−.000	.495	−.090	.999	−.023	−.009	.022	.002
.733	.573	.228	.545	−.100	.762	.000	1.000	.000	−.192	−.091	.023	1.000	−.004	.011	.003
.084	.014	−.963	.006	.993	.035	−.000	.000	1.000	−.227	.274	.008	.003	.999	.044	−.026
.174	−.519	.246	−.521	−.235	.166	−.495	−.192	−.227	1.000	−.000	−.520	−.190	−.258	.792	−.007
.017	−.122	−.272	−.123	.276	−.002	−.090	−.091	.274	−.000	1.000	−.092	−.091	.299	.024	.945
−.657	.832	−.195	.851	.100	−.628	.999	.023	.008	−.520	−.092	1.000	.000	.000	.000	.000
.749	.554	.228	.525	−.098	.777	−.023	1.000	.003	−.190	−.091	.000	1.000	.000	.000	.000
.086	.005	−.961	−.004	.992	.038	−.009	−.004	.999	−.258	.299	.000	.000	1.000	.000	.000
−.003	.026	−.043	.025	.044	−.004	.022	.011	.044	.792	.024	.000	.000	.000	1.000	.000
−.002	.003	.026	.003	−.026	−.000	.002	.003	−.026	−.007	.945	.000	.000	.000	.000	1.000

Three sets of two-dimensional plots were prepared to give pictures of the variation within and among samples for three pairs of discriminant variables 1 and 2, 2 and 3, and 7 and 8. The first pair of these sets together covers the sex and race main effect discriminators while the third set refers to the interaction discriminators. In each of the three sets, a set of 6 scatterplots show the individual skulls within each sample and also the sample concentration ellipse (Figs. 10.5.2–7, 10.5.9–14, and 10.5.16–21). A seventh plot shows all 6 ellipses of concentration (Fig. 10.5.1, 10.5.8, and 10.5.15).

Table 10.5.4

ACTUAL AND EXPECTED FREQUENCIES WITHIN CONCENTRATION ELLIPSES. THE FIRST THREE NUMBERS DENOTE ACTUAL FREQUENCIES FOR THE DISCRIMINANT PAIRS (1, 2), (2, 3), AND (7, 8). THE FOURTH NUMBER DENOTES AN APPROXIMATE EXPECTED FREQUENCY UNDER NORMALITY. THE BRACKETED NUMBER IS AN APPROPRIATE SAMPLING STANDARD DEVIATION.

	Male	Female
Japanese	22 23 20 22.03 (3.66)	15 12 19 13.77 (2.89)
Ainu	26 25 22 22.43 (3.69)	16 18 21 21.64 (3.62)
Australian	17 15 14 14.95 (3.01)	15 16 15 13.77 (2.89)

The 18 scatterplots each present to the eye much the same appearance as a sample of random drawings from a bivariate normal distribution. The distributions roughly follow the shape of the concentration ellipses without definite clusters or patterns of a nonelliptical sort. A brief analysis was done which indicates that the distributions do not have "heavy tails," i.e., do not have too much weight far from the means relative to the bivariate normal distribution. The proportion of a bivariate normal distribution contained within its concentration ellipse is the probability that a $\chi^2(2, 1)$ random variable is less than unity or $1 - e^{-1/2} = 0.39347$. The expected numbers of sample individuals within their sample concentration ellipses should under normality be roughly 0.39347 times the sample sizes, as shown in Table 10.5.4 along with the observed numbers.

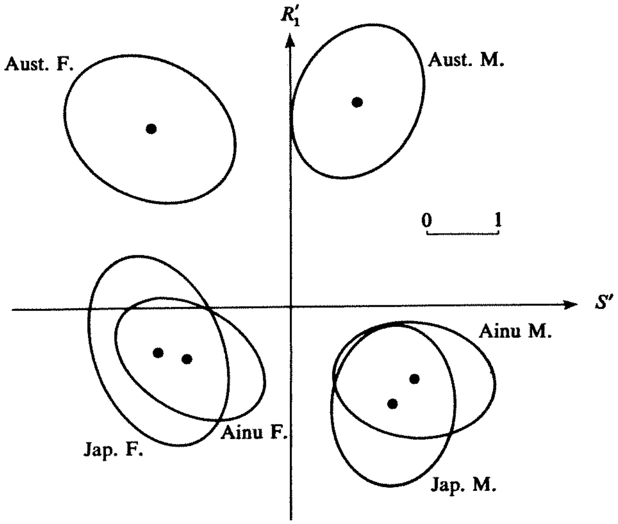

Fig. 10.5.1. Mean-centered sample concentration ellipses for a sex discriminant and first race discriminant. The pooled within sample concentration ellipse is a unit circle in the scale shown.

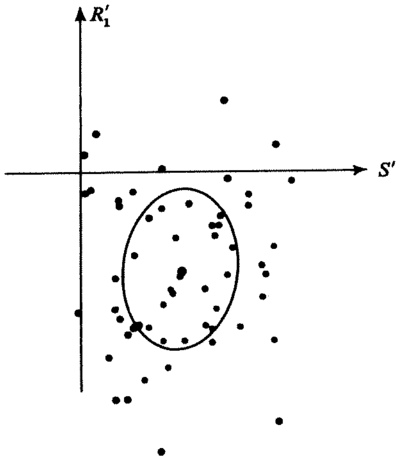

Fig. 10.5.2. The scatterplot of the Japanese male sample corresponding to the ellipse shown in Fig. 10.5.1.

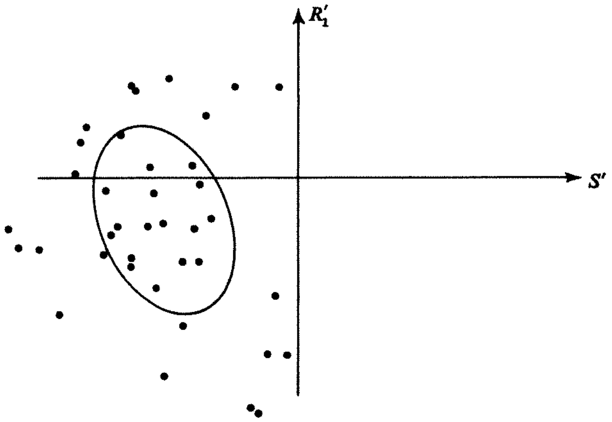

Fig. 10.5.3. The scatterplot of the Japanese female sample corresponding to the ellipse shown in Fig. 10.5.1.

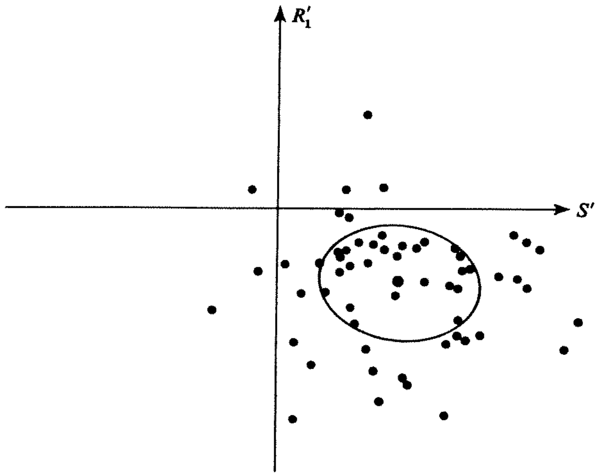

Fig. 10.5.4. The scatterplot of the Ainu male sample corresponding to the ellipse shown in Fig. 10.5.1.

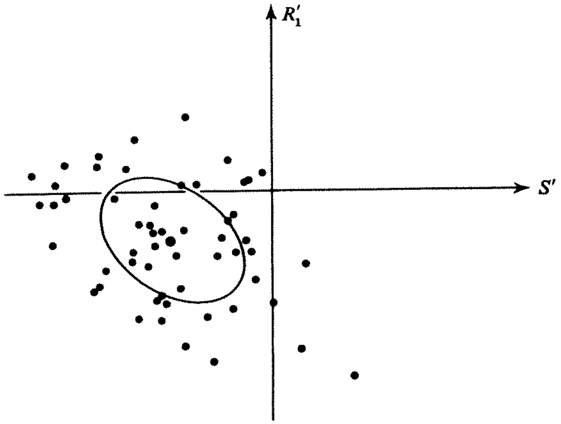

Fig. 10.5.5. The scatterplot of the Ainu female sample corresponding to the ellipse shown in Fig. 10.5.1.

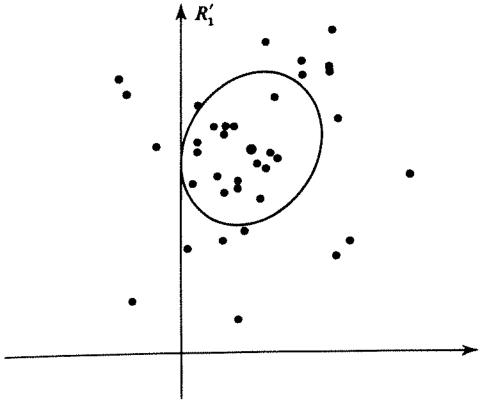

Fig. 10.5.6. The scatterplot of the Australian male sample corresponding to the ellipse shown in Fig. 10.5.1.

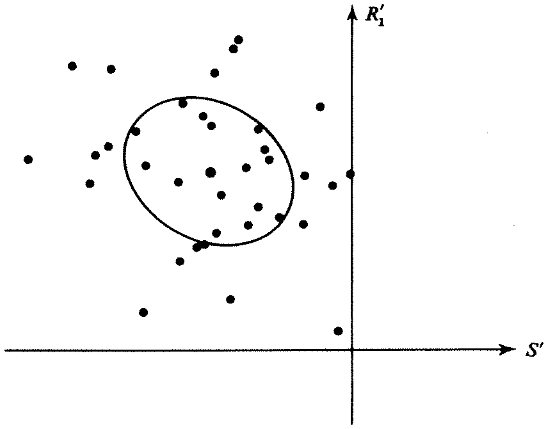

Fig. 10.5.7. The scatterplot of the Australian female sample corresponding to the ellipse shown in Fig. 10.5.1.

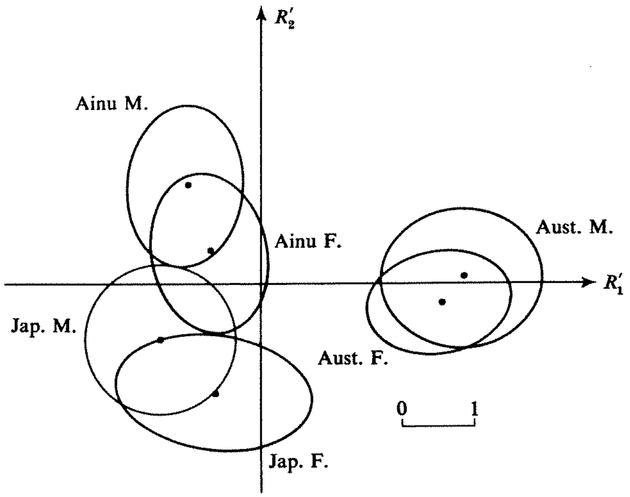

Fig. 10.5.8. Mean-centered sample concentration ellipses for a pair of race discriminants. The pooled within sample concentration ellipse is a unit circle in the scale shown.

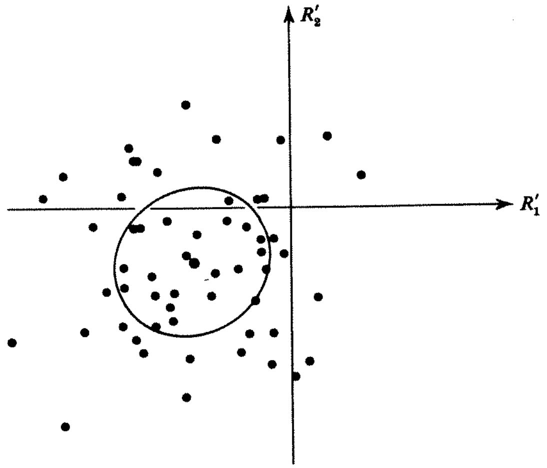

Fig. 10.5.9. The scatterplot of the Japanese male sample corresponding to the ellipse shown in Fig. 10.5.8.

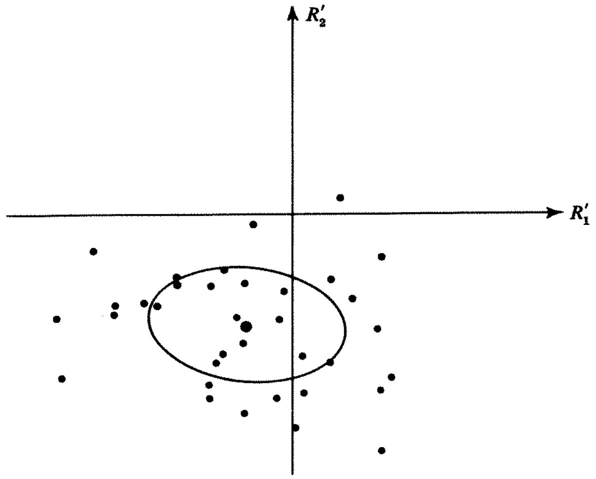

Fig. 10.5.10. The scatterplot of the Japanese female sample corresponding to the ellipse shown in Fig. 10.5.8.

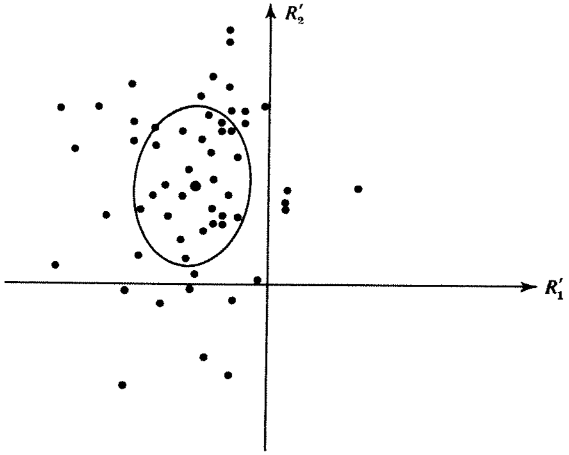

Fig. 10.5.11. The scatterplot of the Ainu male sample corresponding to the ellipse shown in Fig. 10.5.8.

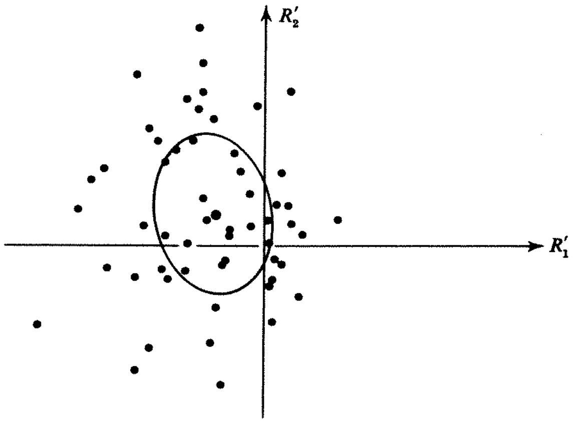

Fig. 10.5.12. The scatterplot of the Ainu female sample corresponding to the ellipse shown in Fig. 10.5.8.

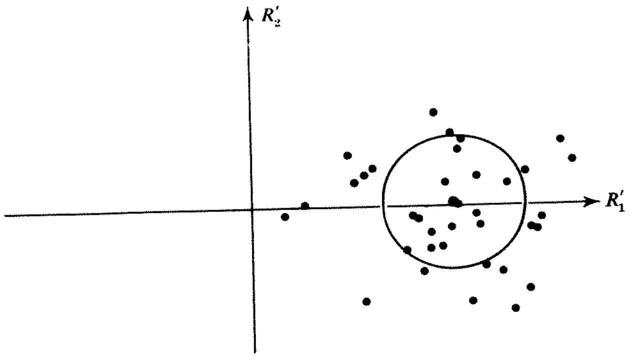

Fig. 10.5.13. The scatterplot of the Australian male sample corresponding to the ellipse shown in Fig. 10.5.8.

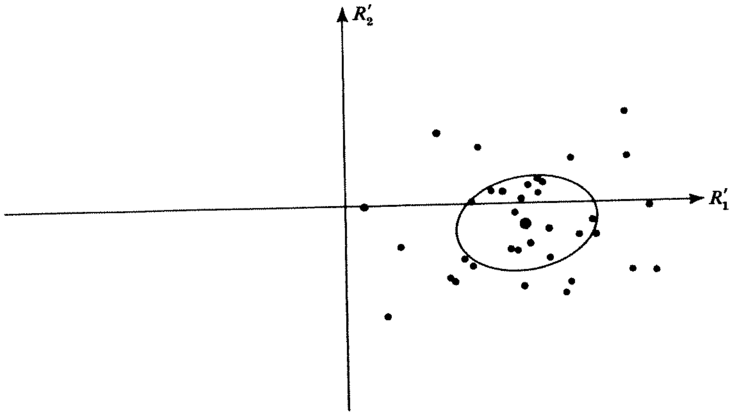

Fig. 10.5.14. The scatterplot of the Australian female sample corresponding to the ellipse shown in Fig. 10.5.8.

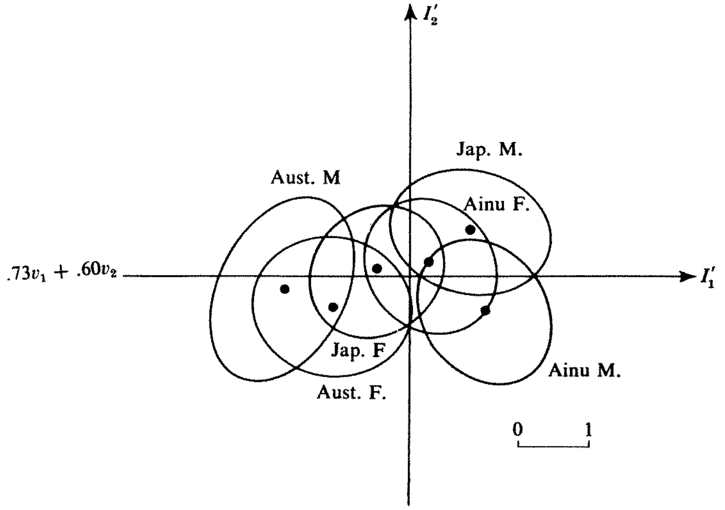

Fig. 10.5.15. Mean-centered sample concentration ellipses for a pair of sex by race interaction discriminants. The pooled within sample concentration ellipse is a unit circle in the scale shown.

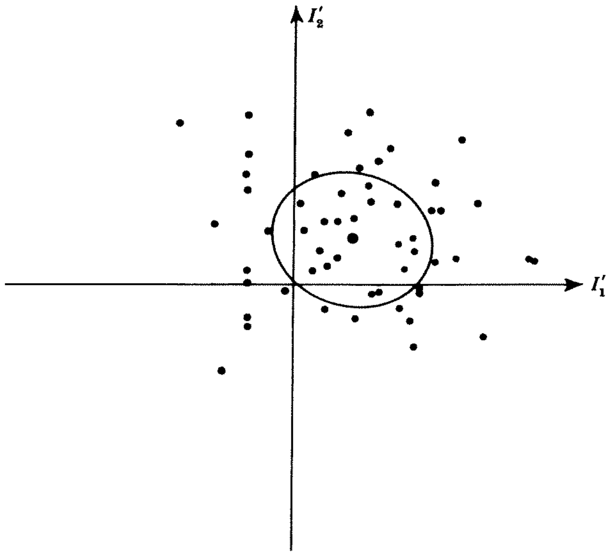

Fig. 10.5.16. The scatterplot of the Japanese male sample corresponding to the ellipse shown in Fig. 10.5.15.

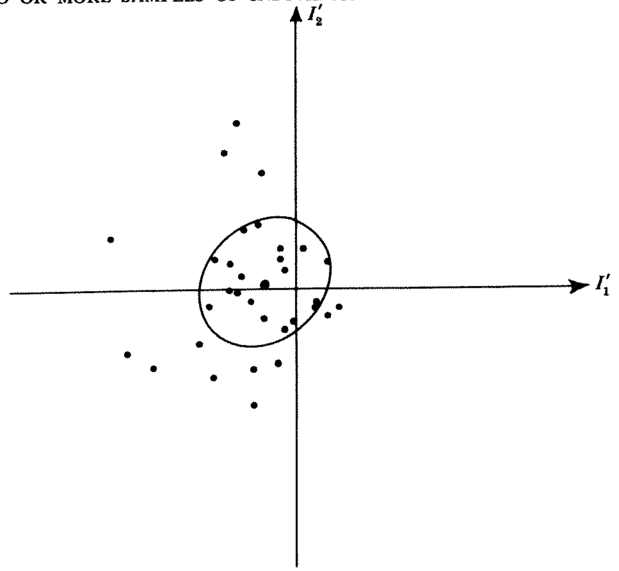

Fig. 10.5.17. The scatterplot of the Japanese female sample corresponding to the ellipse shown in Fig. 10.5.15.

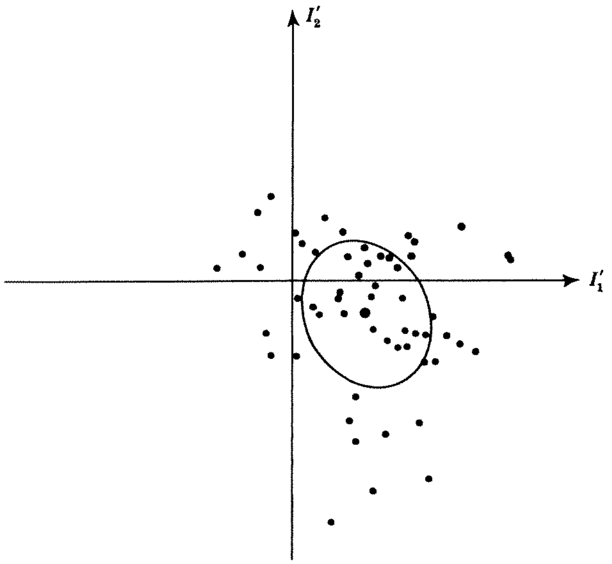

Fig. 10.5.18. The scatterplot of the Ainu male sample corresponding to the ellipse shown in Fig. 10.5.15.

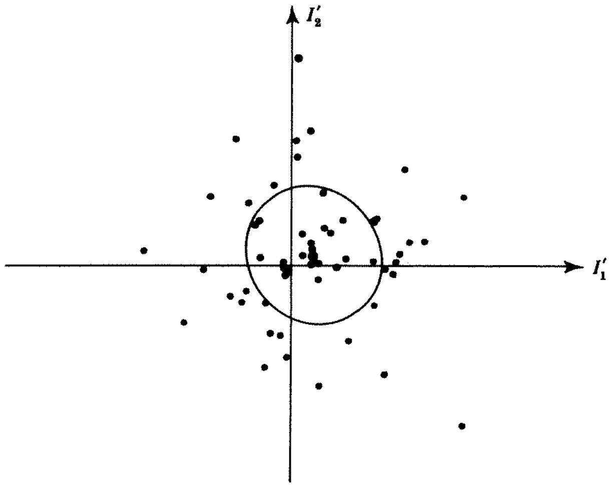

Fig. 10.5.19. The scatterplot of the Ainu female sample corresponding to the ellipse shown in Fig. 10.5.15.

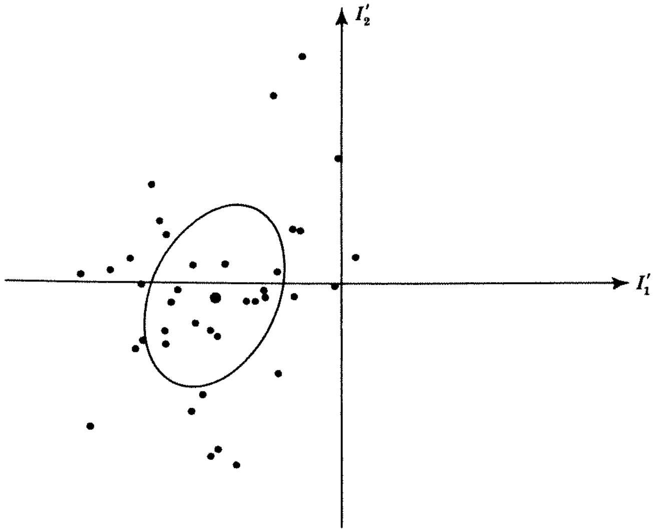

Fig. 10.5.20. The scatterplot of the Australian male sample corresponding to the ellipse shown in Fig. 10.5.15.

There are no serious discrepancies; the overall observed fraction within ellipses is $\frac{331}{828} = 0.400$ which is quite close to 0.393. Unnormally heavy tails would have shown up in expanded concentration ellipses without much effect in the center of the distribution, and one would have expected an increased observed fraction within the ellipses.

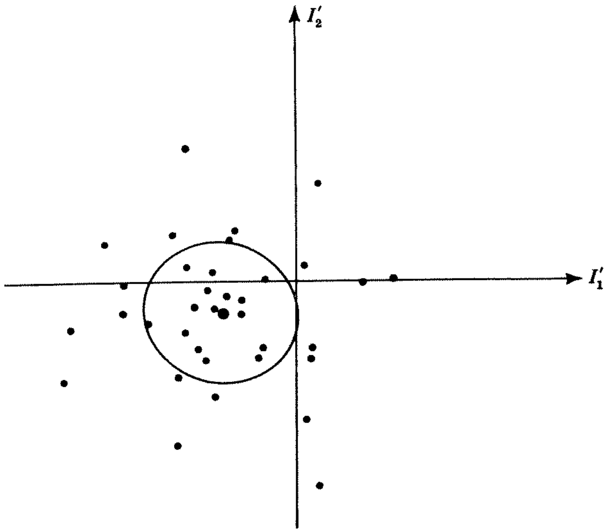

Fig. 10.5.21. The scatterplot of the Australian female sample corresponding to the ellipse shown in Fig. 10.5.15.

Attention is now directed to the sample differences visible in Figs. 10.5.1, 10.5.8, and 10.5.15. An average of the six ellipses shown on each figure is the concentration ellipse determined by the pooled within sample covariance. Thus on Figs. 10.5.8 and 10.5.15 an average is provided by a unit circle while on Fig. 10.5.1 the average corresponds to unit variances and a covariance of -0.131. The deviations of the individual sample ellipses from their averages appear minor and roughly consistent with sampling variation about a common population ellipse.

The differences among the sample means provide more obvious and interesting patterns. Significance tests are applied in Section 14.3 which leave no doubt that the apparent sex differences shown in Fig. 10.5.1 and the apparent race differences shown in Figs. 10.5.1 and 10.5.8 are inconsistent with sampling variation based on samples from normal populations with zero race and sex differences, respectively. On the other hand, interaction differences associated with the variables pictured in Fig. 10.5.15 are not significant. It is apparent also that Figs. 10.5.1 and 10.5.8 are more likely to be overestimating the population, sex, and race differences than to be underestimating them, and some rough approaches to the correction of such biases are given in Section 14.3.

It is a property of these analyses that a discriminator chosen to discriminate well for one type of difference among populations need not have zero discriminating power against other types of differences *even when the dummy variables corresponding to the two types of differences are orthogonal.* For example, in Fig. 10.5.8 it is seen that the R'_2 discriminator also shows a consistent sex difference. This property is a corollary of nonzero correlation between the sets of discriminators. For example, any variable uncorrelated with a sex discriminator must have zero mean differences on sex while any other variable does have a mean difference on sex.

Figure 10.5.8 suggests that the Ainus are somewhat closer to the Japanese than to the Australians and, moreover, are in no sense intermediate between the Japanese and Australians.

10.6 SMALL SAMPLES ON MANY VARIABLES

The distance and best linear discriminator analyses described in Sections 10.2 and 10.4 have assumed a rank p inner product π over $\mathscr{E}$. If π had been semi-definite, its dual π_d would have been a partial inner product defined over a subspace $\mathscr{V}$ of $\mathscr{F}$. Consequently, the distance D between a pair of means $m^{(1)} - m^{(2)}$ could not in general have been defined from (10.2.1). At the same time, the ratio (10.2.2) would have been either infinite or undefined for variables W in the subspace $\mathscr{U}$ of $\mathscr{E}$ over which $(W, W) = 0$, and the definition of best linear discriminator would have become meaningless.

An exception to these disasters arises when $m^{(1)} - m^{(2)}$ belongs to $\mathscr{V}$, for (10.2.1) again becomes meaningful. A meaningful definition of a best linear discriminator can then also be given, although such a best linear discriminator is unique only up to differences lying in the subspace $\mathscr{U}$ of $\mathscr{E}$. The exceptional case can arise in an artificial way, as in Example 10.1, where certain variables are redundant in the sense that their values for *every* sample individual are expressible as linear combinations of the values on the remaining variables. Barring such artificial relations among variables, with actual sample means it would almost never happen that $m^{(1)} - m^{(2)}$ would belong to $\mathscr{V}$. The purpose of this section is to draw attention to a pair of simple expedients which are available in the nonexceptional cases.

If π and π_d are sample-based covariance and concentration inner products of the various types described earlier in this chapter, the common cause of insufficient rank is insufficient sample size. The sample covariance of a p-variate sample of size n is necessarily semi-definite, in general of rank $n - 1$ when $n - 1 < p$. More generally, a linear weighting of sample covariances from k p-variate samples of sizes $n_1, n_2, \ldots, n_k$ is necessarily semi-definite, in general of rank $\sum_1^k (n_i - 1)$, when $\sum_1^k (n_i - 1) < p$. In these circumstances, the analyses of Sections 10.2 and 10.4 *using sample based inner products* become impossible. The following discussion concerns tentative ways to bypass this difficulty.

It is always possible to examine variables one at a time, or to examine them in small sets to which satisfactory multivariate methods apply. Such an approach leaves one with no easy method of forming a combined impression from the separate analyses, and sometimes a series of differences in sample means can be misleading if, due to unsuspected correlations, these differences are largely reflections and repetitions of a single phenomenon. Another approach is to use a few index variables constructed from the original variables via *a priori* weightings. In effect this asks the scientist who originally proposed the high-dimensional variable-space to give up the idea that statistical analysis *per se* can extract information from many variables and that he should condense his set of variables before coming to the statistician.

The author has proposed two methods for attempting to digest a large set of variables simultaneously (Dempster, 1960; 1963b). The aim of these methods is to be relatively free of the hard-to-assess covariation. Both methods rely on an initial reference inner product π. The first method simply suggests looking at the distance between sample means relative to π with the additional wrinkle that a scale factor for π is estimated from the data. This method would be comparable to the earlier methods of this chapter if π could be chosen proportional to a population covariance inner product. Its effectiveness in practice depends on faith in a feasible choice of π being reasonably effective. The second method is based on a faith in principal component analysis as a device for reducing the dimension of variable-space in such a way that important information is not lost. This is a considerable assumption when samples are small.

Current theories of statistical inference are not yet equipped to understand what can be learned from small samples on many variables. The author believes that such understanding, when it comes, will not be encouraging and that data collectors should not have high hopes from such data. The following example is indicative of the type of frustration that may result.

Example 10.4. The body of data considered here consists of 62 biochemical variables measured on each of 12 human subjects of whom four were alcoholics and eight were controls. The data were collected by Roger J. Williams *et al.* (1950) in a shot-gun attempt to find some biochemical differences between alcoholics and normal people. Of the 62 variables, 54 are chemical concentrations from samples of blood serum, urine, and saliva taken under controlled circumstances, five are taste threshholds, and three are phagocytic indices.

In these data, most of the individual variables have values which overlap considerably from one sample to the other, but a few show separation which would be judged significant by simple tests. The methods described above were tried out to see whether plausible methods of judging overall significance would tap hidden aspects of the data and would thereby render a clear judgment of significance. The answer on these data is negative.

An inner product π over the 62-dimensional variable-space was constructed by scaling the original variables so that they had a roughly similar scale of variation, and then choosing the scaled variables to be orthonormal according to π. It will be assumed here that the 12 × 62 data matrix $\mathbf{X}$ has been scaled to represent the orthonormal variables.

The first method of analysis is basically concerned with the squared length

$$(\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}}^{(2)})(\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}}^{(2)})'$$

which should be regarded as the squared norm according to π_d of the mean difference vector. This norm will be corrected to express it in a scale comparable to the variation in the samples.

To derive the scale factor, consider the data first as represented by 62 vectors in the 12-dimensional Euclidean space $\mathscr{N}$. As usual in analysis of variance, the orthonormal basis $\mathscr{N}$ corresponding to the original 12 sample individuals may be altered to an orthonormal basis whose first vector corresponds to the grand mean, whose second vector corresponds to the difference of sample means, and whose remaining vectors representing within sample variation are chosen in any way to complete the basis. Next consider the data plotted as 11 vectors $\mathbf{Z}^{(2)}, \mathbf{Z}^{(3)}, \ldots, \mathbf{Z}^{(12)}$ in individual-space $\mathscr{F}$, where $\mathbf{Z}^{(i)}$ for $i = 1, 2, \ldots, 12$ denote the rows of the data matrix $\mathbf{Z}$ expressed in terms of the new basis in $\mathscr{N}$.

Now

$$\mathbf{Z}^{(2)} = \pm \frac{\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}}^{(2)}}{(\frac{1}{4} + \frac{1}{8})^{1/2}} \tag{10.6.1}$$

represents the differences between sample means, while $\mathbf{Z}^{(3)}, \mathbf{Z}^{(4)}, \ldots, \mathbf{Z}^{(12)}$ are comparable vectors measuring within sample variation. The quantity

$$\frac{1}{10}\sum_{j=3}^{12} \mathbf{Z}^{(j)}\mathbf{Z}^{(j)\prime}$$

estimates a factor which makes π comparable to within sample variation. Finally, the quantity

$$\dot{D}^2 = \frac{(\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}}^{(2)})(\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}}^{(2)})'}{\frac{1}{10}\sum_{j=3}^{12} \mathbf{Z}^{(j)}\mathbf{Z}^{(j)\prime}} \tag{10.6.2}$$

is suggested as a distance measure, roughly similar in intent to D^2 of Section 10.2.

The derivation makes use of a special basis in $\mathscr{N}$, but the reader may easily check that the denominator of (10.6.2) is tr $\mathbf{S}$ where $\mathbf{S}$ is the pooled within sample covariance matrix relative to any basis orthonormal with respect to π. Thus

$$\dot{D}^2 = (\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}}^{(2)})(\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}}^{(2)})'/\text{tr}\,\mathbf{S}. \tag{10.6.3}$$

To compute $\dot{D}^2$ one needs only the vector of mean differences and the pooled within sample covariance matrix **S** both computed from the scaled data matrix **X**.

The result in the example was found to be

$$\dot{D}^2 = 3.685.$$

This value lies at roughly the 0.90 quantile of its normal theory null distribution (see Dempster, 1958; 1960) so should not be regarded as inconsistent with the hypotheses that populations of alcoholics and controls underlying the samples do not differ in their distributions on the 62 variables.

The second method of analysis begins with a reduction to a set of principal variables. The same reference inner product π was used, and the total sample sum inner product corrected for the grand mean was used to represent sample variation. An alternative choice would have been to represent the sample variation in terms of the pooled within sample covariance. The latter choice would have been more appropriate in the presence of clearly meaningful mean differences between the samples. The former was chosen here for technical reasons, to facilitate a program of stepwise significance testing (see Dempster, 1963a; 1963b).

The computations were carried out as follows. The scaled data matrix **X** was replaced by **Y** in which each element had its grand mean subtracted. The principal component analysis is defined in terms of the eigenvalues of the 62×62 matrix $\mathbf{Y}'\mathbf{Y}$ which, like **Y**, has rank 11. In view of Theorem 6.5, the 11 nonzero eigenvalues of $\mathbf{Y}'\mathbf{Y}$ may also be found as the 11 nonzero eigenvalues of $\mathbf{YY}'$ where $\mathbf{YY}'$ is a more manageable matrix of dimensions 12×12 only. Also, according to Theorem 6.5, the eigenvectors of $\mathbf{YY}'$ provide the data matrix on the resulting principal variables. The computations were done by finding $\mathbf{YY}'$ and thence its 11 nonzero eigenvalues and eigenvectors. The result is shown in Table 10.6.1.

The principal variables are shifted in this analysis so that their grand means are zero. Consequently their sample mean differences may be computed as $\frac{3}{8}$ of the sums of the four alcoholic scores. Also the principal variables are scaled to be orthonormal according to the total sum inner product (corrected or not corrected for the grand mean), so that distances D'''^2 in the sense of Section 10.2 are very easily computed as the sums of squares of mean differences. Finally, the more usual D^2 relative to the pooled within sample covariance may be computed from (10.2.8) and (10.2.10). These distances are shown in Table 10.6.2 for subsets of the principal variables.

Table 10.6.1

Principal variable		1	2	3	4	5	6	7	8	9	10	11
Eigenvalues		159.8244	113.4557	83.4787	71.8194	65.1554	53.0508	43.0775	35.0866	29.5223	23.1098	10.3909
Alcoholic scores	1	.3995	.2505	.2552	.1988	.1644	.2078	.6103	.0390	.1729	.2704	.2060
	2	−.1970	−.0382	.5664	−.4960	−.0199	.0642	.1855	−.0799	−.4117	−.3059	−.0349
	3	.3277	−.1524	−.2095	−.0069	−.0938	−.6052	.1797	−.4787	.0970	−.3113	−.0155
	4	.2105	−.7613	−.1389	−.1612	−.1188	.4261	−.0875	.0399	.1246	.0859	.1397
Control scores	5	.0537	.2816	−.4066	.0295	−.0129	.2951	−.1367	−.4082	−.5704	.2615	−.0443
	6	−.2844	−.1063	.2893	.1471	−.3960	−.3854	−.1404	−.0104	−.0023	.6220	.0840
	7	.2268	−.0516	.3032	.6281	.0101	.0701	−.3401	.1650	−.1586	−.3090	−.3277
	8	−.3296	.0392	−.4109	.0748	−.3483	.0241	.4480	.4337	−.0154	−.1607	−.3097
	9	−.3839	−.2162	−.1369	.1184	.7928	−.1972	.0284	.0856	−.0555	.0686	.0805
	10	−.4058	.2356	.0569	.0514	−.0490	.3180	−.1618	−.4105	.5946	−.1951	.0101
	11	.0891	.2831	−.1338	−.0937	−.1019	−.1070	−.3011	.3788	−.0344	−.2470	.6954
	12	.2934	.2360	−.0343	−.4902	.1731	−.1105	−.2842	.2455	.2591	.2206	−.4838

Table 10.6.2

Principal variables included	D'''^2	D^2
1	0.0772	0.97
1, 2	0.1463	2.40
1, 2, 3	0.1779	3.38
1, 2, 3, 4	0.2083	4.69
1, 2, 3, 4, 5	0.2090	4.71
1, 2, 3, . . . , 6	0.2102	4.78
1, 2, 3, . . . , 7	0.4311	22.32
1, 2, 3, . . . , 8	0.3534	61.40
1, 2, 3, . . . , 9	0.3535	61.52
1, 2, 3, . . . , 10	0.3630	113.6
1, 2, 3, . . . , 11	0.3753	∞

The values of D^2 in Table 10.6.2 are quite meaningless as estimates of population distances. Indeed the tests of the successive increments in D^2 as described in Section 14.3 show that only the jump from 22.32 to 61.40 is beyond its 0.95 quantile assuming identical normal samples. Furthermore it is only slightly beyond at 0.96, and such a result among 10 tests is not surprising. It thus appears that the principal component analysis did not succeed in isolating variables which point to a difference between alcoholics and normal people.