CHAPTER 10

Canonical Correlation Coefficients

This final chapter is concerned with the interpretation of canonical correlation coefficients and their relationship to affine dependence and independence between two random vectors. After using an invariance argument to show that population canonical correlations are a natural measure of affine dependence, these population coefficients are interpreted as cosines of the angles between subspaces (as defined in Chapter 1). Next, the sample canonical correlations are defined and interpreted as cosines of angles. The distribution theory associated with the sample coefficients is discussed briefly.

When two random vectors have a joint normal distribution, independence between the vectors is equivalent to the population canonical correlations all being zero. The problem of testing for independence is treated in the fourth section of this chapter. The relationship between the MANOVA testing problem and testing for independence is discussed in the fifth and final section of the chapter.

10.1. POPULATION CANONICAL CORRELATION COEFFICIENTS

There are a variety of ways to introduce canonical correlation coefficients and three of these are considered in this section. We begin our discussion with the notion of affine dependence between two random vectors. Let $X \in (V, (\cdot, \cdot)_1)$ and $Y \in (W, (\cdot, \cdot)_2)$ be two random vectors defined on the same probability space so the random vector $Z = \{X, Y\}$ takes values in the vector space $V \oplus W$. It is assumed that $\text{Cov}(X) = \Sigma_{11}$ and $\text{Cov}(Y) = \Sigma_{22}$ both exist and are nonsingular. Therefore, $\text{Cov}(Z)$ exists (see Proposition

2.15) and is given by

$$\Sigma = \mathrm{Cov}(Z) = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}' & \Sigma_{22} \end{pmatrix}.$$

Also, the mean vector of Z is

$$\mu = \mathcal{E}Z = \{\mathcal{E}X, \mathcal{E}Y\} = \{\mu_1, \mu_2\}.$$

Definition 10.1. Two random vectors U and $\tilde{U}$, in $(V, (\cdot, \cdot)_1)$ are *affinely equivalent* if $U = A\tilde{U} + a$ for some nonsingular linear transformation A and some vector $a \in V$.

It is clear that affine equivalence is an equivalence relation among random vectors defined on the same probability space and taking values in V.

We now consider measures of affine dependence between X and Y, which are functions of $\mu = \{\mathcal{E}X, \mathcal{E}Y\}$ and $\Sigma = \mathrm{Cov}(Z)$ where $Z = \{X, Y\}$. Let $m(\mu, \Sigma)$ be some real-valued function of μ and Σ that is supposed to measure affine dependence. If instead of X we observe $\tilde{X}$, which is affinely equivalent to X, then the affine dependence between X and Y should be the same as the affine dependence between $\tilde{X}$ and Y. Similarly, if $\tilde{Y}$ is affinely equivalent to Y, then the affine dependence between X and Y should be the same as the affine dependence between X and $\tilde{Y}$. These remarks imply that $m(\mu, \Sigma)$ should be invariant under affine transformations of both X and Y. If (A, a) is an affine transformation on V, then $(A, a)v = Av + a$ where A is nonsingular on V to V. Recall that the group of all affine transformations on V to V is denoted by $Al(V)$ and the group operation is given by

$$(A_1, a_1)(A_2, a_2) = (A_1A_2, A_1a_2 + a_1).$$

Also, let $Al(W)$ be the affine group for W. The product group $Al(V) \times Al(W)$ acts on the vector space $V \oplus W$ in the obvious way:

$$((A, a), (B, b))\{v, w\} = \{Av + b, Bw + b\}.$$

The argument given above suggests that the affine dependence between X and Y should be the same as the affine dependence between $(A, a)X$ and $(B, b)Y$ for all $(A, a) \in Al(V)$ and $(B, b) \in Al(W)$. We now need to interpret this requirement as a condition on $m(\mu, \Sigma)$. The random vector

$$((A, a), (B, b))\{X, Y\} = \{AX + a, BY + b\}$$

has a mean vector given by

$$((A,a),(B,b))\{\mu_1,\mu_2\} = \{A\mu_1 + a, A\mu_2 + b\}$$

and a covariance given by

$$\begin{pmatrix} A\Sigma_{11}A' & A\Sigma_{12}B' \\ B\Sigma'_{12}A' & B\Sigma_{22}B' \end{pmatrix}$$

Therefore, the group $Al(V) \times Al(W)$ acts on the set

$$\Theta = \{(\mu,\Sigma)|\mu \in V \oplus W, \Sigma \geqslant 0, \Sigma_{ii} > 0, i = 1,2\}.$$

For $g \equiv ((A,a),(B,b)) \in Al(V) \times Al(W)$, the group action is given by

$$(\mu,\Sigma) \to (g\mu, g(\Sigma))$$

where

$$g\mu = \{A\mu_1 + a, B\mu_2 + b\}$$

and

$$g(\Sigma) = \begin{pmatrix} A\Sigma_{11}A' & A\Sigma_{12}B' \\ B\Sigma'_{12}A' & B\Sigma_{22}B' \end{pmatrix}.$$

Requiring the affine dependence between X and Y to be equal to the affine dependence between $(A,a)X$ and $(B,b)Y$ simply means that the function m defined on Θ must be invariant under the group action given above. Therefore, m must be a function of a maximal invariant function under the action of $Al(V) \times Al(W)$ on Θ. The following proposition gives one form of a maximal invariant.

Proposition 10.1. Let $q = \dim V$, $r = \dim W$, and let $t = \min\{q,r\}$. Given

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma'_{12} & \Sigma_{22} \end{pmatrix},$$

which is positive definite on $V \oplus W$, let $\lambda_1 \geqslant \cdots \geqslant \lambda_t \geqslant 0$ be the t largest eigenvalues of

$$\Lambda(\Sigma) \equiv \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

where $\Sigma_{21} \equiv \Sigma'_{12}$. Define a function h on Θ by

$$h(\mu, \Sigma) = (\lambda_1, \lambda_2, \dots, \lambda_t),$$

where $\lambda_1 \geqslant \cdots \geqslant \lambda_t$ are defined in terms of Σ as above. Then h is a maximal invariant function under the action of $G \equiv Al(V) \times Al(W)$ on Θ.

Proof. Let $\{v_1, \dots, v_t\}$ and $\{w_1, \dots, w_t\}$ be fixed orthonormal sets in V and W. For each Σ, define $Q_{12}(\Sigma)$ by

$$Q_{12}(\Sigma) = \sum_{i=1}^{t} \lambda_i^{1/2} v_i \Box w_i$$

where $\lambda_1 \geqslant \cdots \geqslant \lambda_t$ are the t largest eigenvalues of $\Lambda(\Sigma)$. Given $(\mu, \Sigma) \in \Theta$, we first claim that there exists a $g \in G$ such that $g\mu = 0$ and

$$g(\Sigma) = \begin{pmatrix} I_V & Q_{12}(\Sigma) \\ (Q_{12}(\Sigma))' & I_W \end{pmatrix}.$$

The proof of this claim follows. For $g = ((A, a), (B, b))$, we have

$$g(\Sigma) = \begin{pmatrix} A\Sigma_{11}A' & A\Sigma_{12}B' \\ B\Sigma_{21}A' & B\Sigma_{22}B' \end{pmatrix}.$$

Choose $A = \Gamma\Sigma_{11}^{-1/2}$ and $B = \Delta\Sigma_{22}^{-1/2}$ where $\Gamma \in \mathcal{O}(V)$, $\Delta \in \mathcal{O}(W)$, and $\Sigma_{ii}^{-1/2}$ is the inverse of the positive definite square root of Σ_{ii}, $i = 1, 2$. For each Γ and Δ,

$$A\Sigma_{11}A' = \Gamma\Sigma_{11}^{-1/2}\Sigma_{11}\Sigma_{11}^{-1/2}\Gamma' = I_V$$

$$B\Sigma_{22}B' = \Delta\Sigma_{22}^{-1/2}\Sigma_{22}\Sigma_{22}^{-1/2}\Delta' = I_W$$

and

$$A\Sigma_{12}B' = \Gamma\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}\Delta'.$$

Using the singular value decomposition, write

$$\Lambda_{12} \equiv \Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2} = \sum_{i=1}^{t} \lambda_i^{1/2} x_i \Box y_i$$

where $\{x_1, \dots, x_t\}$ and $\{y_1, \dots, y_t\}$ are orthonormal sets in V and W, respectively. This representation follows by noting that the rank of Λ_{12} is at most t and

$$\Lambda_{12}\Lambda_{12}' = \Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}$$

has the same eigenvalues as $\Lambda(\Sigma)$, which are $\lambda_1 \geqslant \cdots \geqslant \lambda_t \geqslant 0$. For A and B as above, it now follows that

$$A\Sigma_{12}B' = \sum_{i=1}^{t} \lambda_i^{1/2}(\Gamma x_i)\square(\Delta y_i).$$

Choose Γ so that $\Gamma x_i = v_i$ and choose Δ so that $\Delta y_i = w_i$. Then we have

$$A\Sigma_{12}B' = Q_{12}(\Sigma)$$

so $g(\Sigma)$ has the form claimed. With these choices for A and B, now choose $a = -A\mu_1$ and $b = -B\mu_2$. Then

$$\begin{aligned} g\mu = g\{\mu_1, \mu_2\} &= ((A, a), (B, b))\{\mu_1, \mu_2\} \\ &= \{A\mu_1 + a, B\mu_2 + b\} = \{0, 0\} = 0. \end{aligned}$$

The proof of the claim is now complete. To finish the proof of Proposition 10.1, first note that Proposition 1.39 implies that h is a G-invariant function. For the maximality of h, suppose that $h(\mu, \Sigma) = h(\nu, \Psi)$. Thus

$$Q_{12}(\Sigma) = Q_{12}(\Psi),$$

which implies that there exists a g and $\tilde{g}$ such that

$$g\mu = 0, \qquad \tilde{g}\nu = 0,$$

and

$$g(\Sigma) = \begin{pmatrix} I_V & Q_{12}(\Sigma) \\ (Q_{12}(\Sigma))' & I_W \end{pmatrix} = \tilde{g}(\Psi).$$

Therefore,

$$g^{-1}\tilde{g}(\nu, \Psi) = (\mu, \Sigma)$$

so h is maximal invariant. □

The form of the singular value decomposition used in the proof of Proposition 10.1 is slightly different than that given in Theorem 1.3. For a linear transformation C of rank k defined on $(V, (\cdot, \cdot)_1)$ to $(W, (\cdot, \cdot)_2)$, Theorem 1.3 asserts that

$$C = \sum_{i}^{k} \mu_i w_i \Box x_i$$

where $\mu_i > 0$, $\{x_1, \ldots, x_k\}$, and $\{w_1, \ldots, w_k\}$ are orthonormal sets in V and W. With $q = \dim V$, $r = \dim W$, and $t = \min\{q, r\}$, obviously $k \leqslant t$. When $k < t$, it is clear that the orthonormal sets above can be extended to $\{x_1, \ldots, x_t\}$ and $\{w_1, \ldots, w_t\}$, which are still orthonormal sets in V and W. Also, setting $\mu_i = 0$ for $i = k + 1, \ldots, t$, we have

$$C = \sum_{1}^{t} \mu_i w_i \Box x_i,$$

and $\mu_1^2 \geqslant \cdots \geqslant \mu_t^2$ are the t largest eigenvalues of both CC' and $C'C$. This form of the singular value decomposition is somewhat more convenient in this chapter since the rank of C is not explicitly mentioned. However, the rank of C is just the number of μ_i, which are strictly positive. The corresponding modification of Proposition 1.48 should now be clear.

Returning to our original problem of describing measures of affine dependence, say $m(\mu, \Sigma)$, Proposition 10.1 demonstrates that m is invariant under affine relabelings of X and Y iff m is a function of the t largest eigenvalues, $\lambda_1, \ldots, \lambda_t$, of $\Lambda(\Sigma)$. Since the rank of $\Lambda(\Sigma)$ is at most t, the remaining eigenvalues of $\Lambda(\Sigma)$, if there are any, must be zero. Before suggesting some particular measures $m(\mu, \Sigma)$, the canonical correlation coefficients are discussed.

Definition 10.2. In the notation of Proposition 10.1, let $\rho_i = \lambda_i^{1/2}$, $i = 1, \ldots, t$. The numbers $\rho_1 \geqslant \rho_2 \geqslant \cdots \geqslant \rho_t \geqslant 0$ are called the *population canonical correlation coefficients*.

Since ρ_i is a one-to-one function of λ_i, it follows that the vector $(\rho_1, \ldots, \rho_t)$ also determines a maximal invariant function under the action of G on Θ. In particular, any measure of affine dependence should be a function of the canonical correlation coefficients.

The canonical correlation coefficients have a natural interpretation as cosines of angles between subspaces in a vector space. Recall that $Z = \{X, Y\}$ takes values in the vector space $V \oplus W$ where $(V, (\cdot, \cdot)_1)$ and $(W, (\cdot, \cdot)_2)$

are inner product spaces. The covariance of Z, with respect to the natural inner product, say $(\cdot, \cdot)$, on $V \oplus W$, is

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

In the discussion that follows, it is assumed that Σ is positive definite. Let $(\cdot, \cdot)_\Sigma$ denote the inner product on $V \oplus W$ defined by

$$(z_1, z_2)_\Sigma = (z_1, \Sigma z_2) = \text{cov}[(z_1, Z), (z_2, Z)],$$

for $z_1, z_2 \in V \oplus W$. The vector space V can be thought of as a subspace of $V \oplus W$—namely, just identify V with $V \oplus \{0\} \subseteq V \oplus W$. Similarly, W is a subspace of $V \oplus W$. The next result interprets the canonical correlations as the cosines of angles between the subspaces V and W when the inner product on $V \oplus W$ is $(\cdot, \cdot)_\Sigma$.

Proposition 10.2. Given Σ, the canonical correlation coefficients $\rho_1 \geqslant \cdots \geqslant \rho_t$ are the cosines of the angles between V and W as subspaces in the inner product space $(V \oplus W, (\cdot, \cdot)_\Sigma)$.

Proof. Let P_1 and P_2 be the orthogonal projections (relative to $(\cdot, \cdot)_\Sigma$) onto $V \oplus \{0\}$ and $W \oplus \{0\}$, respectively. In view of Proposition 1.48 and Definition 1.28, it suffices to show that the t largest eigenvalues of $P_1 P_2 P_1$ are $\lambda_i = \rho_i^2$, $i = 1, \ldots, t$. We claim that

$$C_1 = \begin{pmatrix} I_V & \Sigma_{11}^{-1}\Sigma_{12} \\ 0 & 0 \end{pmatrix}$$

is the orthogonal projection onto $V \oplus \{0\}$. For $\{v, w\} \in V \oplus W$,

$$\begin{pmatrix} I_V & \Sigma_{11}^{-1}\Sigma_{12} \\ 0 & 0 \end{pmatrix}\{v, w\} = \{v + \Sigma_{11}^{-1}\Sigma_{12}w, 0\}$$

so the range of C_1 is $V \oplus \{0\}$ and C_1 is the identity on $V \oplus \{0\}$. That $C_1^2 = C_1$ is easily verified. Also, since

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

the identity $C_1'\Sigma = \Sigma C_1$ holds. Here C_1' is the adjoint of C_1 relative to the inner product $(\cdot,\cdot)$—namely,

$$C_1' = \begin{pmatrix} I_V & 0 \\ \Sigma_{21}\Sigma_{11}^{-1} & 0 \end{pmatrix}.$$

This shows that C_1 is self-adjoint relative to the inner product $(\cdot,\cdot)_\Sigma$. Hence C_1 is the orthogonal projection onto $V \oplus \{0\}$ in $(V \oplus W, (\cdot,\cdot)_\Sigma)$. A similar argument yields

$$C_2 = \begin{pmatrix} 0 & 0 \\ \Sigma_{22}^{-1}\Sigma_{21} & I_W \end{pmatrix}$$

as the orthogonal projection onto $\{0\} \oplus W$ in $(V \oplus W, (\cdot,\cdot)_\Sigma)$. Therefore $P_i = C_i$, $i = 1, 2$, and a bit of algebra shows that

$$P_1P_2P_1 = \begin{pmatrix} \Lambda(\Sigma) & C \\ 0 & 0 \end{pmatrix}$$

where $\Lambda(\Sigma) = \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ and

$$C = \Lambda(\Sigma)\Sigma_{11}^{-1}\Sigma_{12}.$$

Thus the characteristic polynomial of $P_1P_2P_1$ is given by

$$p(\alpha) = \det[P_1P_2P_1 - \alpha I] = (-\alpha)^r \det[\Lambda(\Sigma) - \alpha I_V]$$

where $r = \dim W$. Since $t = \min\{q, r\}$ where $q = \dim V$, it follows that the t largest eigenvalues of $P_1P_2P_1$ are the t largest eigenvalues of $\Lambda(\Sigma)$. These are $\rho_1^2 \geq \cdots \geq \rho_t^2$, so the proof is complete. □

Another interpretation of the canonical correlation coefficients can be given using Proposition 1.49 and the discussion following Definition 1.28. Using the notation adopted in the proof of Proposition 10.2, write

$$P_2P_1 = \sum_{i=1}^{t} \rho_i \xi_i \square \eta_i$$

where $\{\eta_1, \ldots, \eta_t\}$ is an orthonormal set in $V \oplus \{0\}$ and $\{\xi_1, \ldots, \xi_t\}$ is an orthonormal set in $\{0\} \oplus W$. Here orthonormal refers to the inner product $(\cdot,\cdot)_\Sigma$ on $V \oplus W$, as does the symbol $\square$ in the expression for P_2P_1—that is, for $z_1, z_2 \in V \oplus W$,

$$(z_1 \square z_2)z = (z_2, z)_\Sigma z_1 = (z_2, \Sigma z)z_1.$$

The existence of this representation for P_2P_1 follows from Proposition 1.48, as does the relationship

$$(\eta_i, \xi_j)_\Sigma = \delta_{ij}\rho_j$$

for $i, j = 1, \ldots, t$. Define the sets D_{1i} and D_{2i}, $i = 1, \ldots, t$, as in Proposition 1.49 (with $M_1 = V \oplus \{0\}$ and $M_2 = \{0\} \oplus W$), so

$$\sup_{\eta \in D_{1i}} \sup_{\xi \in D_{2i}} (\eta, \xi)_\Sigma = (\eta_i, \xi_i)_\Sigma = \rho_i$$

for $i = 1, \ldots, t$. To interpret ρ_1, first consider the case $i = 1$. A vector η is in D_{11} iff

$$\eta = \{v, 0\}, \qquad v \in V$$

and

$$1 = (\eta, \Sigma\eta) = (v, \Sigma_{11}v)_1 = \operatorname{var}(v, X)_1.$$

Similarly, $\xi \in D_{21}$ iff

$$\xi = \{0, w\}, \qquad w \in W$$

and

$$1 = (\xi, \Sigma\xi) = (w, \Sigma_{22}w)_2 = \operatorname{var}(w, Y)_2.$$

However, for $\eta = \{v, 0\} \in D_{11}$ and $\xi = \{0, w\} \in D_{21}$,

$$(\eta, \xi)_\Sigma = (v, \Sigma_{12}w)_1 = \operatorname{cov}\{(v, X)_1, (w, Y)_2\}.$$

This is just the ordinary correlation between $(v, X)_1$ and $(w, Y)_2$ as v and w have been normalized so that $1 = \operatorname{var}(v, X)_1 = \operatorname{var}(w, Y)_2$. Since $(\eta, \xi)_\Sigma \leqslant \rho_1$ for all $\eta \in D_{11}$ and $\xi \in D_{21}$, it follows that for every $x \in V$, $x \neq 0$, and $y \in W$, $y \neq 0$, the correlation between $(x, X)_1$ and $(y, Y)_2$ is no greater than ρ_1. Further, writing $\eta_1 = \{v_1, 0\}$ and $\xi_1 = \{0, w_1\}$, we have

$$\begin{aligned}\rho_1 &= (\eta_1, \xi_1)_\Sigma = (\eta_1, \Sigma\xi_1) \\ &= (v_1, \Sigma_{12}w_1)_1 = \operatorname{cov}\{(v_1, X)_1, (w_1, Y)_2\},\end{aligned}$$

which is the correlation between $(v_1, X)_1$ and $(w_1, Y)_2$. Therefore, ρ_1 is the maximum correlation between $(x, X)_1$ and $(y, Y)_2$ for all nonzero $x \in V$ and $y \in W$. Further, this maximum correlation is achieved by choosing $x = v_1$ and $y = w_1$.

The second largest canonical correlation coefficient, ρ_2, satisfies the equality

$$\sup_{\eta \in D_{12}} \sup_{\xi \in D_{22}} (\eta, \xi)_\Sigma = (\eta_2, \xi_2)_\Sigma = \rho_2.$$

A vector η is in D_{12} iff

$$\eta = \{v, 0\}, \qquad v \in V$$

$$1 = (\eta, \eta)_\Sigma = (v, \Sigma_{11} v)_1$$

and

$$0 = (\eta, \eta_1)_\Sigma = (v, \Sigma_{11} v_1)_1.$$

Also, a vector ξ is in D_{22} iff

$$\xi = \{0, w\}, \qquad w \in W$$

$$1 = (\xi, \xi)_\Sigma = (w, \Sigma_{22} w)_2$$

and

$$0 = (\xi, \xi_1)_\Sigma = (w, \Sigma_{22} w_1)_2.$$

These relationships provide the following interpretation of ρ_2. The maximum correlation between $(x, X)_1$ and $(y, Y)_2$ is ρ_1 and is

$$\rho_1 = \operatorname{cov}\{(v_1, X)_1, (w_1, Y)_2\}$$

since $1 = \operatorname{var}(v_1, X)_1 = \operatorname{var}(w_1, Y)_2$. Suppose we now want to find the maximum correlation between $(x, X)_1$ and $(y, Y)_2$ subject to the condition

$$\text{(i)} \quad \begin{cases} \operatorname{cov}\{(x, X)_1, (v_1, X)_1\} = 0 \\ \operatorname{cov}\{(y, Y)_2, (w_1, Y)_2\} = 0. \end{cases}$$

Clearly (i) is equivalent to

$$\text{(ii)} \quad \begin{cases} (x, \Sigma_{11} v_1)_1 = 0 \\ (y, \Sigma_{22} w_1)_2 = 0. \end{cases}$$

Since correlation is invariant under multiplication of the random variables by positive constants, to find the maximum correlation between $(x, X)_1$ and $(y, Y)_2$ subject to (ii), it suffices to maximize $\operatorname{cov}\{(x, X)_1, (y, Y)_2\}$ over those x's and y's that satisfy

$$\text{(iii)}\quad \begin{cases} (x, \Sigma_{11}x)_1 = 1, (x, \Sigma_{11}v_1)_1 = 0 \\ (y, \Sigma_{22}y)_2 = 1, (y, \Sigma_{22}w_1)_2 = 0. \end{cases}$$

However, $x \in V$ satisfies (iii) iff $\eta = \{x, 0\}$ is in D_{12} and $y \in W$ satisfies (iii) iff $\xi = \{0, y\}$ is in D_{22}. Further, for such x, y, η, and ξ,

$$\operatorname{cov}\{(x, X)_1, (y, Y)_2\} = (\eta, \xi)_\Sigma.$$

Thus maximizing this covariance subject to (iii) is the same as maximizing $(\eta, \xi)_\Sigma$ for $\eta \in D_{12}$ and $\xi \in D_{22}$. Of course, this maximum is ρ_2 and is achieved at $\eta_2 \in D_{12}$ and $\xi_2 \in D_{22}$. Writing $\eta_2 = \{v_2, 0\}$ and $\xi_2 = \{0, w_2\}$, it is clear that $v_2 \in V$ and $w_2 \in W$ satisfy (iii) and

$$\operatorname{cov}\{(v_2, X)_1, (w_2, Y)_2\} = \rho_2.$$

Furthermore, Proposition 1.48 shows that

$$0 = (\eta_1, \xi_2)_\Sigma = (\eta_2, \xi_1)_\Sigma,$$

which implies that

$$0 = \operatorname{cov}\{(v_1, X)_1, (w_2, Y)_2\} = \operatorname{cov}\{(v_2, X)_1, (w_1, Y)_2\}.$$

Therefore, the problem of maximizing the correlation between $(x, X)_1$ and $(y, Y)_2$ (subject to the condition that the correlation between $(x, X)_1$ and $(v_1, X)_1$ be zero and the correlation between $(y, Y)_2$ and $(w_1, Y)_2$ be zero) has been solved.

It should now be fairly clear how to interpret the remaining canonical correlation coefficients. The easiest way to describe the coefficients is by induction. The coefficient ρ_1 is the largest possible correlation between $(x, X)_1$ and $(y, Y)_2$ for nonzero vectors $x \in V$ and $y \in W$. Further, there exist vectors $v_1 \in V$ and $w_1 \in W$ such that

$$\operatorname{cov}\{(v_1, X)_1, (w_1, Y)_2\} = \rho_1$$

and

$$1 = \operatorname{var}(v_1, X)_1 = \operatorname{var}(w_1, Y)_2.$$

These vectors came from η_1 and ξ_1 in the representation

$$P_2P_1 = \sum_{i=1}^{t} \rho_i \xi_i \square \eta_i$$

given earlier. Since $\eta_i \in V \oplus \{0\}$, we can write $\eta_i = \{v_i, 0\}$, $i = 1,\dots,t$. Similarly, $\xi_i = \{0, w_i\}$, $i = 1,\dots,t$. Using Proposition 1.48, it is easy to check that

$$\operatorname{cov}\{(v_j, X)_1, (w_k, Y)_2\} = \rho_j \delta_{jk}$$

$$\operatorname{cov}\{(v_j, X)_1, (v_k, X)_1\} = \delta_{jk}$$

$$\operatorname{cov}\{(w_j, Y)_2, (w_k, Y)_2\} = \delta_{jk}$$

for $j, k = 1,\dots,t$. Of course, these relationships are simply a restatement of the properties of $\xi_1,\dots,\xi_t$ and $\eta_1,\dots,\eta_t$. For example,

$$\operatorname{cov}\{(v_j, X)_1, (w_k, Y)_2\} = (v_j, \Sigma_{12} w_k)_1 = (\eta_j, \xi_k)_\Sigma = \rho_j \delta_{jk}.$$

However, as argued in the case of ρ_2, we can say more. Given $\rho_1,\dots,\rho_t$ and the vectors $v_1,\dots,v_{i-1}$ and $w_1,\dots,w_{i-1}$ obtained from $\eta_1,\dots,\eta_{i-1}$ and $\xi_1,\dots,\xi_{i-1}$, consider the problem of maximizing the correlation between $(x, X)_1$ and $(y, Y)_2$ subject to the conditions that

$$\begin{cases} \operatorname{cov}\{(x, X)_1, (v_j, X)_1\} = 0, & j = 1,\dots,i-1 \\ \operatorname{cov}\{(y, Y)_2, (w_j, Y)_2\} = 0, & j = 1,\dots,i-1. \end{cases}$$

By simply unravelling the notation and using Proposition 1.49, this maximum correlation is ρ_i and is achieved for $x = v_i$ and $y = w_i$. This successive maximization of correlation is often a useful interpretation of the canonical correlation coefficients.

The vectors $v_1,\dots,v_t$ and $w_1,\dots,w_t$ lead to what are called the *canonical variates*. Recall that $q = \dim V$, $r = \dim W$ and $t = \min\{q, r\}$. For definiteness, assume that $q \leqslant r$ so $t = q$. Thus $\{v_1,\dots,v_q\}$ is a basis for V and satisfies

$$(v_j, \Sigma_{11} v_k)_1 = \delta_{jk}$$

for $j, k = 1,\dots,q$ so $\{v_1,\dots,v_q\}$ is an orthonormal basis for V relative to the inner product determined by Σ_{11}. Further, the linearly independent set $\{w_1,\dots, w_q\}$ satisfies

$$\left(w_j, \Sigma_{22}w_k\right)_2 = \delta_{jk}$$

so $\{w_1,\dots, w_q\}$ is an orthonormal set relative to the inner product determined by Σ_{22}. Now, extend this set to $\{w_1,\dots, w_r\}$ so that this is an orthonormal basis for W in the Σ_{22} inner product.

Definition 10.3. The real-valued random variables defined by

$$X_i = (v_i, X)_1, \qquad i = 1,\dots, q$$

and

$$Y_i = (w_i, Y)_2, \qquad i = 1,\dots, r$$

are called the *canonical variates* of X and Y, respectively.

Proposition 10.3. The canonical variates satisfy the relationships

(i) $\operatorname{var} X_j = \operatorname{var} Y_k = 1$.

(ii) $\operatorname{cov}\{X_j, Y_k\} = \rho_j\delta_{jk}$.

These relationships hold for $j = 1,\dots, q$ and $k = 1,\dots, r$. Here, $\rho_1,\dots, \rho_q$ are the canonical correlation coefficients.

Proof. This is just a restatement of part of what we have established above. □

Let us briefly review what has been established thus far about the population canonical correlation coefficients $\rho_1,\dots, \rho_t$. These coefficients were defined in terms of a maximal invariant under a group action and this group action arose quite naturally in an attempt to define measures of affine dependence. Using Proposition 1.48 and Definition 1.28, it was then shown that $\rho_1,\dots, \rho_t$ are cosines of angles between subspaces with respect to an inner product defined by Σ. The statistical interpretation of the coefficients came from the detailed information given in Proposition 1.49 and this interpretation closely resembled the discussion following Definition 1.28. Given X in $(V, (\cdot, \cdot)_1)$ and Y in $(W, (\cdot, \cdot)_2)$ with a nonsingular covariance

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

the existence of special bases $\{v_1,\ldots,v_q\}$ and $\{w_1,\ldots,w_r\}$ for V and W was established. In terms of the canonical variates

$$X_i=(v_i,X)_1,\qquad Y_j=(w_j,Y)_2,$$

the properties of these bases can be written

$$1=\operatorname{var}X_i=\operatorname{var}Y_j$$

and

$$\operatorname{cov}\{X_i,Y_j\}=\rho_i\delta_{ij}$$

for $i=1,\ldots,q$ and $j=1,\ldots,r$. Here, the convention that $\rho_i=0$ for $i>t=\min\{q,r\}$ has been used although ρ_i is not defined for $i>t$. When $q\leqslant r$, the covariance matrix of the variates $X_1,\ldots,X_q$, $Y_1,\ldots,Y_r$ (in that order) is

$$\Sigma_0=\begin{pmatrix} I_q & (DO)\\ (DO)' & I_r\end{pmatrix}$$

where D is a $q\times q$ diagonal matrix with diagonal entries $\rho_1\geqslant\cdots\geqslant\rho_q$ and O is a $q\times(r-q)$ block of zeroes. The reader should compare this matrix representation of Σ to the assertion of Proposition 5.7.

The final point of this section is to relate a prediction problem to that of suggesting a particular measure of affine dependence. Using the ideas developed in Chapter 4, a slight generalization of Proposition 2.22 is presented below. Again, consider $X\in(V,(\cdot,\cdot)_1)$ and $Y\in(W,(\cdot,\cdot)_2)$ with $\mathcal{E}X=\mu_1$, $\mathcal{E}Y=\mu_2$, and

$$\operatorname{Cov}\{X,Y\}=\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}.$$

It is assumed that Σ_{11} and Σ_{22} are both nonsingular. Consider the problem of predicting X by an affine function of Y—say $CY+v_0$ where $C\in\mathcal{L}(W,V)$ and $v_0\in V$. Let $[\cdot,\cdot]$ be any inner product on V and let $\|\cdot\|$ be the norm defined by $[\cdot,\cdot]$. The following result shows how to choose C and v_0 to minimize

$$\mathcal{E}\|X-(CY+v_0)\|^2.$$

Of course, the inner product $[\cdot,\cdot]$ on V is related to the inner product $(\cdot,\cdot)_1$

by

$$[v_1, v_2] = (v_1, A_0 v_2)_1$$

for some positive definite A_0.

Proposition 10.4. For any $C \in \mathcal{L}(W, V)$ and $v_0 \in V$, the inequality

$$\mathcal{E}\|X - (CY + v_0)\|^2 \geqslant \langle A_0, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\rangle$$

holds. There is equality in this inequality iff

$$v_0 = \hat{v}_0 \equiv \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2$$

and

$$C = \hat{C} \equiv \Sigma_{12}\Sigma_{22}^{-1}.$$

Here, $\langle\cdot,\cdot\rangle$ is the natural inner product on $\mathcal{L}(V, V)$ inherited from $(V, (\cdot,\cdot)_1)$.

Proof. First, write

$$X - (CY + v_0) = U_1 + U_2$$

where

$$U_1 = X - \left(\hat{C}Y + \hat{v}_0\right) = X - \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(Y - \mu_2)$$

and

$$U_2 = (\hat{C} - C)Y + \hat{v}_0 - v_0.$$

Clearly, U_1 has mean zero. It follows from Proposition 2.17 that U_1 and U_2 are uncorrelated and

$$\text{Cov}(U_1) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Further, from Proposition 4.3 we have $\mathcal{E}[U_1, U_2] = 0$. Therefore,

$$\begin{aligned}\mathcal{E}\|X - (CY + v_0)\|^2 &= \mathcal{E}\|U_1 + U_2\|^2 = \mathcal{E}\|U_1\|^2 + \mathcal{E}\|U_2\|^2\\ &= \mathcal{E}(U_1, A_0U_1) + \mathcal{E}\|U_2\|^2 = \mathcal{E}\langle A_0, U_1 \square U_1\rangle + \mathcal{E}\|U_2\|^2\\ &= \langle A_0, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\rangle + \mathcal{E}\|U_2\|^2,\end{aligned}$$

where the last equality follows from the identity

$$\mathcal{E}U_1 \Box U_1 = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

established in Proposition 2.21. Thus the desired inequality holds and there is equality iff $\mathcal{E}\|U_2\|^2 = 0$. But $\mathcal{E}\|U_2\|^2$ is zero iff U_2 is zero with probability one. This holds iff $v_0 = \hat{v}_0$ and $C = \hat{C}$ since $\text{Cov}(Y) = \Sigma_{22}$ is positive definite. This completes the proof. □

Now, choose A_0 to be Σ_{11}^{-1} in Proposition 10.4. Then the mean squared error due to predicting X by $\hat{C}Y + \hat{v}_0$, measured relative to Σ_{11}^{-1}, is

$$\phi(\Sigma) \equiv \langle \Sigma_{11}^{-1}, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \rangle = \mathcal{E}\|X - (\hat{C}Y + \hat{v}_0)\|^2.$$

Here, $\|\cdot\|$ is obtained from the inner product defined by

$$[v_1, v_2] = (v_1, \Sigma_{11}^{-1}v_2).$$

We now claim that ϕ is invariant under the group of transformations discussed in Proposition 10.1, and thus ϕ is a possible measure of affine dependence between X and Y. To see this, first recall that $\langle \cdot, \cdot \rangle$ is just the trace inner product for linear transformations. Using properties of the trace, we have

$$\begin{aligned}\phi(\Sigma) &= \langle I, I - \Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2} \rangle \\ &= \text{tr}\left(I - \Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}\right) \\ &= \sum_{i=1}^{q}(1 - \lambda_i)\end{aligned}$$

where $\lambda_1 \geqslant \cdots \geqslant \lambda_q \geqslant 0$ are the eigenvalues of $\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}$. However, at most $t = \min\{q, r\}$ of these eigenvalues are nonzero and, by definition, $\rho_i = \lambda_i^{1/2}$, $i = 1, \ldots, t$, are the canonical correlation coefficients. Thus

$$\phi(\Sigma) = \sum_{1}^{t}\left(1 - \rho_i^2\right) + (q - t)$$

is a function of $\rho_1, \ldots, \rho_t$ and hence is an invariant measure of affine dependence. Since the constant $q - t$ is irrelevant, it is customary to use

$$\phi_1(\Sigma) = \sum_{i=1}^{t} \left(1 - \rho_i^2\right)$$

rather than $\phi(\Sigma)$ as a measure of affine dependence.

10.2. SAMPLE CANONICAL CORRELATIONS

To introduce the sample canonical correlation coefficients, again consider inner product spaces $(V, (\cdot, \cdot)_1)$ and $(W, (\cdot, \cdot)_2)$ and let $(V \oplus W, (\cdot, \cdot))$ be the direct sum space with the natural inner product $(\cdot, \cdot)$. The observations consist of n random vectors $Z_i = \{X_i, Y_i\} \in V \oplus W$, $i = 1, \ldots, n$. It is assumed that these random vectors are uncorrelated with each other and $\mathcal{L}(Z_i) = \mathcal{L}(Z_j)$ for all i, j. Although these assumptions are not essential in much of what follows, it is difficult to interpret canonical correlations without these assumptions. Given $Z_1, \ldots, Z_n$, define the random vector Z by specifying that Z takes on the values Z_i with probability $1/n$. Obviously, the distribution of Z is discrete in $V \oplus W$ and places mass $1/n$ at Z_i for $i = 1, \ldots, n$. Unless otherwise specified, when we speak of the distribution of Z, we mean the conditional distribution of Z given $Z_1, \ldots, Z_n$ as described above. Since the distribution of Z is nothing but the sample probability measure of $Z_1, \ldots, Z_n$, we can think of Z as a sample approximation to a random vector whose distribution is $\mathcal{L}(Z_1)$. Now, write $Z = \{X, Y\}$ with $X \in V$ and $Y \in W$ so X is X_i with probability $1/n$ and Y is Y_i with probability $1/n$. Given $Z_1, \ldots, Z_n$, the mean vector of Z is

$$\mathcal{E}Z = \bar{Z} \equiv \frac{1}{n} \sum_{i=1}^{n} Z_i = \{\bar{X}, \bar{Y}\}$$

and the covariance of Z is

$$\operatorname{Cov} Z = S \equiv \frac{1}{n} \sum_{i=1}^{n} (Z_i - \bar{Z}) \Box (Z_i - \bar{Z}).$$

This last assertion follows from Proposition 2.21 by noting that

$$\operatorname{Cov} Z = \mathcal{E}(Z - \bar{Z}) \Box (Z - \bar{Z})$$

since the mean of Z is $\bar{Z}$. When $V = R^q$ and $W = R^r$ are the standard coordinate spaces with the usual inner products, then S is just the sample covariance matrix. Since S is a linear transformation on $V \oplus W$ to $V \oplus W$, S can be written as

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}.$$

It is routine to show that

$$S_{11} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}) \square (X_i - \overline{X})$$

$$S_{12} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}) \square (Y_i - \overline{Y})$$

$$S_{22} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \overline{Y}) \square (Y_i - \overline{Y})$$

and $S_{21} = S'_{12}$. The reader should note that the symbol $\square$ appearing in the expressions for S_{11}, S_{12}, and S_{22} has a different meaning in each of the three expressions—namely, the outer product depends on the inner products on the spaces in question. Since it is clear which vectors are in which spaces, this multiple use of $\square$ should cause no confusion.

Now, to define the sample canonical correlation coefficients, the results of Section 10.1 are applied to the random vector Z. For this reason, we assume that $S = \operatorname{Cov} Z$ is nonsingular. With $q = \dim V$, $r = \dim W$, and $t = \min\{q, r\}$, the canonical correlation coefficients are the square roots of the t largest eigenvalues of

$$\Lambda(S) = S_{11}^{-1} S_{12} S_{22}^{-1} S_{21}.$$

In the sampling situation under discussion, these roots are denoted by $r_1 \geqslant \cdots \geqslant r_t \geqslant 0$ and are called the *sample canonical correlation coefficients*. The justification for such nomenclature is that $r_1^2, \ldots, r_t^2$ are the t largest eigenvalues of $\Lambda(S)$ where S is the sample covariance based on $Z_1, \ldots, Z_n$. Of course, all of the discussion of the previous section applies directly to the situation at hand. In particular, the vector $(r_1, \ldots, r_t)$ is a maximal invariant under the group action described in Proposition 10.1. Also, $r_1, \ldots, r_t$ are the cosines of the angles between the subspaces $V \oplus \{0\}$ and $\{0\} \oplus W$ in the vector space $V \oplus W$ relative to the inner product determined by S.

Now, let $\{v_1,\dots, v_q\}$ and $\{w_1,\dots, w_r\}$ be the canonical bases for V and W. Then we have

$$\operatorname{cov}\{(v_i, X)_1, (w_j, Y)_2\} = r_i\delta_{ij}$$

for $i = 1,\dots, q$ and $j = 1,\dots, r$. The convention that $r_i \equiv 0$ for $i > t$ is being used. To interpret what this means in terms of the sample $Z_1,\dots, Z_n$, consider r_1. For nonzero $x \in V$ and $y \in W$, the maximum correlation between $(x, X)_1$ and $(y, Y)_2$ is r_1 and is achieved for $x = v_1$ and $y = w_1$. However, given $Z_1,\dots, Z_n$, we have

$$\begin{aligned}\operatorname{var}(x, X)_1 &= \operatorname{var}(\{x,0\}, Z) = (\{x,0\}, S\{x,0\})\\ &= (x, S_{11}x)_1 = \frac{1}{n}\sum_{i=1}^{n}(x, X_i - \bar{X})_1^2\end{aligned}$$

and, similarly,

$$\operatorname{var}(y, Y)_2 = \frac{1}{n}\sum_{i=1}^{n}(y, Y_i - \bar{Y})_2^2.$$

An analogous calculation shows that

$$\operatorname{cov}\{(x, X)_1, (y, Y)_2\} = \frac{1}{n}\sum_{i=1}^{n}(x, X_i - \bar{X})_1(y, Y_i - \bar{Y})_2.$$

Thus $\operatorname{var}(x, X)_1$ is just the sample variance of the random variables $(x, X_i)_1$, $i = 1,\dots, n$, and $\operatorname{var}(y, Y)_2$ is the sample variance of $(y, Y_i)_2$, $i = 1,\dots, n$. Also, $\operatorname{cov}\{(x, X)_1, (y, Y)_2\}$ is the sample covariance of the random variables $(x, X_i)_1$, $(y, Y_i)_2$, $i = 1,\dots, n$. Therefore, the correlation between $(x, X)_1$ and $(y, Y)_2$ is the ordinary sample correlation coefficient for the random variables $(x, X_i)_1$, $(y, Y_i)_2$, $i = 1,\dots, n$. This observation implies that the maximum possible sample correlation coefficient for $(x, X_i)_1$, $(y, Y_i)_2$, $i = 1,\dots, n$ is the largest sample canonical correlation coefficient, r_1, and this maximum is attained by choosing $x = v_1$ and $y = w_1$. The interpretation of $r_2,\dots, r_t$ should now be fairly obvious. Given i, $2 \leqslant i \leqslant t$, and given $r_1,\dots, r_{i-1}$, $v_1,\dots, v_{i-1}$, and $w_1,\dots, w_{i-1}$, consider the problem of maximizing the correlation between $(x, X)_1$ and $(y, Y)_2$ subject to the conditions

$$\operatorname{cov}\{(x, X)_1, (v_j, X)_1\} = 0, \qquad j = 1,\dots, i-1$$

$$\operatorname{cov}\{(y, Y)_2, (w_j, Y)_2\} = 0, \qquad j = 1,\dots, i-1.$$

These conditions are easily shown to be equivalent to the conditions that the sample correlation for

$$(x, X_k)_1, \qquad (v_j, X_k)_1, \qquad k = 1, \dots, n$$

be zero for $j = 1, \dots, i-1$ with a similar statement concerning the Y's. Further, the correlation between $(x, X)_1$ and $(y, Y)_2$ is the sample correlation for $(x, X_k)_1$, $(y, Y_k)_2$, $k = 1, \dots, n$. The maximum sample correlation is r_i and is attained by choosing $x = v_i$ and $y = w_i$. Thus the sample interpretation of $r_1, \dots, r_t$ is completely analogous to the population interpretation of the population canonical correlation coefficients.

For the remainder of this section, it is assumed that $V = R^q$ and $W = R^r$ are the standard coordinate spaces with the usual inner products, so $V \oplus W$ is just R^p where $p = q + r$. Thus our sample is $Z_1, \dots, Z_n$ with $Z_i \in R^p$ and we write

$$Z_i = \begin{pmatrix} X_i \\ Y_i \end{pmatrix} \in R^p$$

with $X_i \in R^q$ and $Y_i \in R^r$, $i = 1, \dots, n$. The sample covariance matrix, assumed to be nonsingular, is

$$S = \frac{1}{n}\sum_1^n (Z_i - \bar{Z})(Z_i - \bar{Z})' = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$$

where

$$S_{11} = \frac{1}{n}\sum_1^n (X_i - \bar{X})(X_i - \bar{X})'$$

$$S_{22} = \frac{1}{n}\sum_1^n (Y_i - \bar{Y})(Y_i - \bar{Y})'$$

$$S_{12} = \frac{1}{n}\sum_1^n (X_i - \bar{X})(Y_i - \bar{Y})'$$

and $S_{21} = S_{12}'$. Now, form the random matrix $\tilde{Z}: n \times p$ whose rows are $(Z_i - \bar{Z})'$ and partition $\tilde{Z}$ into $U: n \times q$ and $V: n \times r$ so that

$$\tilde{Z} = (UV).$$

The rows of U are $(X_i - \bar{X})'$ and the rows of V are $(Y_i - \bar{Y})'$, $i = 1,\dots, n$. Obviously, we have $nS = \tilde{Z}'\tilde{Z}$, $nS_{11} = U'U$, $nS_{22} = V'V$, and $nS_{12} = U'V$. The sample canonical correlation coefficients $r_1 \geqslant \cdots \geqslant r_t$ are the square roots of the t largest eigenvalues of

$$\Lambda(S) = S_{11}^{-1}S_{12}S_{22}^{-1}S_{21} = (U'U)^{-1}U'V(V'V)^{-1}V'U.$$

However, the t largest eigenvalues of $\Lambda(S)$ are the same as the t largest eigenvalues of $P_X P_Y$ where

$$P_X = U(U'U)^{-1}U'$$

and

$$P_Y = V(V'V)^{-1}V'.$$

Now, P_X is the orthogonal projection onto the q-dimensional subspace of R^n, say M_X, spanned by the columns of U. Also, P_Y is the orthogonal projection onto the r-dimensional subspace of R^n, say M_Y, spanned by the columns of V. It follows from Proposition 1.48 and Definition 1.28 that the sample canonical correlation coefficients $r_1,\dots, r_t$ are the cosines of the angles between the two subspaces M_X and M_Y contained in R^n. Summarizing, we have the following proposition.

Proposition 10.5. Given random vectors

$$Z_i = \begin{pmatrix} X_i \\ Y_i \end{pmatrix} \in R^p, \qquad i = 1,\dots, n$$

where $X_i \in R^q$ and $Y_i \in R^r$, form the matrices $U: n \times q$ and $V: n \times r$ as above. Let $M_X \subseteq R^n$ be the subspace spanned by the columns of U and let $M_Y \subseteq R^n$ be the subspace spanned by the columns of V. Assume that the sample covariance matrix

$$S = \frac{1}{n}\sum_1^n (Z_i - \bar{Z})(Z_i - \bar{Z})'$$

is nonsingular. Then the sample canonical correlation coefficients are the cosines of the angles between M_X and M_Y.

The sample coefficients $r_1,\dots, r_t$ have been shown to be the cosines of angles between subspaces in two different vector spaces. In the first case, the interpretation followed from the material developed in Section 10.1 of this chapter: namely, $r_1, \ldots, r_t$ are the cosines of the angles between $R^q \oplus \{0\} \subseteq R^p$ and $\{0\} \oplus R^r \subseteq R^p$ when R^p has the inner product determined by the sample covariance matrix. In the second case, described in Proposition 10.5, $r_1, \ldots, r_t$ are the cosines of the angles between M_X and M_Y in R^n when R^n has the standard inner product. The subspace M_X is spanned by the columns of U where U has rows $(X_i - \overline{X})'$, $i = 1, \ldots, n$. Thus the coordinates of the jth column of U are $X_{ij} - \overline{X}_j$ for $i = 1, \ldots, n$ where X_{ij} is the jth coordinate of $X_i \in R^q$, and $\overline{X}_j$ is the jth coordinate of $\overline{X}$. This is the reason for the subscript X on the subspace M_X. Of course, similar remarks apply to M_Y.

The vector $(r_1, \ldots, r_t)$ can also be interpreted as a maximal invariant under a group action on the sample matrix. Given

$$Z_i = \begin{pmatrix} X_i \\ Y_i \end{pmatrix} \in R^p, \qquad i = 1, \ldots, n,$$

let $\tilde{X}: n \times q$ have rows X_i', $i = 1, \ldots, n$ and let $\tilde{Y}: n \times r$ have rows Y_i', $i = 1, \ldots, n$. Then the data matrix of the whole sample is

$$\tilde{Z} = (\tilde{X}\tilde{Y}) : n \times p,$$

which has rows Z_i', $i = 1, \ldots, n$. Let $e \in R^n$ be the vector of all ones. It is assumed that $\tilde{Z} \in \mathcal{Z} \subseteq \mathcal{L}_{p,n}$ where $\mathcal{Z}$ is the set of all $n \times p$ matrices such that the sample covariance mapping

$$s(\tilde{Z}) = (\tilde{Z} - e\overline{Z}')'(\tilde{Z} - e\overline{Z}')$$

has rank p. Assuming that $n \geqslant p + 1$, the complement of $\mathcal{Z}$ in $\mathcal{L}_{p,n}$ has Lebesgue measure zero. To describe the group action on $\mathcal{Z}$, let G be the set of elements $g = (\Gamma, c, C)$ where

$$\Gamma \in \mathcal{O}_n(e) = \{\Gamma | \Gamma \in \mathcal{O}_n, \Gamma e = e\}, \qquad c \in R^p$$

and

$$C = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}, \qquad A \in Gl_q, \quad B \in Gl_r.$$

For $g = (\Gamma, c, C)$, the value of g at $\tilde{Z}$ is

$$g\tilde{Z} = \Gamma\tilde{Z}C' + ec'.$$

Since

$$s(g\tilde{Z}) = Cs(\tilde{Z})C'.$$

it follows that each $g \in G$ is a one-to-one onto mapping of $\mathfrak{Z}$ to $\mathfrak{Z}$. The composition in G, defined so G acts on the left of $\mathfrak{Z}$, is

$$(\Gamma_1, c_1, C_1)(\Gamma_2, c_2, C_2) \equiv (\Gamma_1\Gamma_2, c_1 + C_1c_2, C_1C_2).$$

Proposition 10.6. Under the action of G on $\mathfrak{Z}$, a maximal invariant is the vector of canonical correlation coefficients $r_1, \ldots, r_t$ where $t = \min\{q, r\}$.

Proof. Let $\mathbb{S}_p^+$ be the space of $p \times p$ positive definite matrices so the sample covariance mapping $s: \mathfrak{Z} \to \mathbb{S}_p^+$ is onto. Given $S \in \mathbb{S}_p^+$, partition S as

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$$

where S_{11} is $q \times q$, S_{22} is $r \times r$, and S_{12} is $q \times r$. Define h on $\mathbb{S}_p^+$ by letting $h(S)$ be the vector $(\lambda_1, \ldots, \lambda_t)'$ of the t largest eigenvalues of

$$\Lambda(S) = S_{11}^{-1}S_{12}S_{22}^{-1}S_{21}.$$

Since $r_i = \sqrt{\lambda_i}$, $i = 1, \ldots, t$, the proposition will be proved if it is shown that

$$\varphi(\tilde{Z}) \equiv h(s(\tilde{Z}))$$

is a maximal invariant function. This follows since $h(s(\tilde{Z})) = (\lambda_1, \ldots, \lambda_t)'$, which is a one-to-one function of $(r_1, \ldots, r_t)$. The proof that φ is maximal invariant proceeds as follows. Consider the two subgroups G_1 and G_2 of G defined by

$$G_1 = \{g | g = (\Gamma, c, I_p) \in G\}$$

and

$$G_2 = \{g | g = (I_n, 0, C) \in G\}.$$

Note that G_2 acts on the space $\mathbb{S}_p^+$ in the obvious way— namely, if $g_2 = (I_n, 0, C)$, then

$$g_2(S) \equiv CSC', \qquad S \in \mathbb{S}_p^+.$$

Further, since

$$(\Gamma, c, C) = (\Gamma, c, I_p)(I_n, 0, C),$$

it follows that each $g \in G$ can be written as $g = g_1 g_2$ where $g_i \in G_i$, $i = 1, 2$. Now, we make two claims:

(i) $s : \mathfrak{Z} \to \mathbb{S}_p^+$ is a maximal invariant under the action of G_1 on $\mathfrak{Z}$.
(ii) $h : \mathbb{S}_p^+ \to R^t$ is a maximal invariant under the action of G_2 on $\mathbb{S}_p^+$.

Assuming (i) and (ii), we now show that $\varphi(\tilde{Z}) = h(s(\tilde{Z}))$ is maximal invariant. For $g \in G$, write $g = g_1 g_2$ with $g_i \in G_i$, $i = 1, 2$. Since

$$s(g_1 \tilde{Z}) = s(\tilde{Z}), \qquad g_1 \in G_1$$

and

$$s(g_2 \tilde{Z}) = g_2(s(\tilde{Z})), \qquad g_2 \in G_2,$$

we have

$$\varphi(g\tilde{Z}) = h(s(g_1 g_2 \tilde{Z})) = h(s(g_2 \tilde{Z})) = h(g_2 s(\tilde{Z})) = h(s(\tilde{Z})).$$

It follows that φ is invariant. To show that φ is maximal invariant, assume $\varphi(\tilde{Z}_1) = \varphi(\tilde{Z}_2)$. A $g \in G$ must be found so that $g\tilde{Z}_1 = \tilde{Z}_2$. Since h is maximal invariant under G_2 and

$$h(s(\tilde{Z}_1)) = h(s(\tilde{Z}_2)),$$

there is a $g_2 \in G_2$ such that

$$g_2(s(\tilde{Z}_1)) = s(\tilde{Z}_2).$$

However,

$$g_2(s(\tilde{Z}_1)) = s(g_2 \tilde{Z}_1) = s(\tilde{Z}_2)$$

and s is maximal invariant under G_1 so there exists a g_1 such that

$$g_1 g_2 \tilde{Z}_1 = \tilde{Z}_2.$$

This completes the proof that φ, and hence $r_1, \ldots, r_t$, is a maximal invariant

—assuming claims (i) and (ii). The proof that $s : \mathcal{Z} \to \mathbb{S}_p^+$ is a maximal invariant is an easy application of Proposition 1.20 and is left to the reader. That $h : \mathbb{S}_p^+ \to R^t$ is maximal invariant follows from an argument similar to that given in the proof of Proposition 10.1. □

The group action on $\mathcal{Z}$ treated in Proposition 10.6 is suggested by the following considerations. Assuming that the observations $Z_1, \dots, Z_n$ in R^p are uncorrelated random vectors and $\mathcal{L}(Z_i) = \mathcal{L}(Z_1)$ for $i = 1, \dots, n$, it follows that

$$\mathcal{E}\tilde{Z} = e\mu'$$

and

$$\operatorname{Cov} \tilde{Z} = I_n \otimes \Sigma$$

where $\mu = \mathcal{E}Z_1$ and $\operatorname{Cov} Z_1 = \Sigma$. When $\tilde{Z}$ is transformed by $g = (\Gamma, c, C)$, we have

$$\mathcal{E}g\tilde{Z} = e(C\mu + c)'$$

and

$$\operatorname{Cov} g\tilde{Z} = I_n \otimes (C\Sigma C').$$

Thus the induced action of g on (μ, Σ) is exactly the group action considered in Proposition 10.1. The special structure of $\mathcal{E}\tilde{Z}$ and $\operatorname{Cov}\tilde{Z}$ is reflected by the fact that, for $g = (\Gamma, 0, I_p)$, we have $\mathcal{E}g\tilde{Z} = \mathcal{E}\tilde{Z}$ and $\operatorname{Cov} g\tilde{Z} = \operatorname{Cov}\tilde{Z}$.

10.3. SOME DISTRIBUTION THEORY

The distribution theory associated with the sample canonical correlation coefficients is, to say the least, rather complicated. Most of the results in this section are derived under the assumption of normality and the assumption that the population canonical correlations are zero. However, the distribution of the sample multiple correlation coefficient is given in the general case of a nonzero population multiple correlation coefficient.

Our first result is a generalization of Example 7.12. Let $Z_1, \dots, Z_n$ be a random sample of vectors in R^p and partition Z_i as

$$Z_i = \begin{pmatrix} X_i \\ Y_i \end{pmatrix}, \qquad X_i \in R^q, \quad Y_i \in R^r.$$

Assume that Z_1 has a density on R^p given by

$$p(z|\mu,\Sigma) = |\Sigma|^{-1/2} f((z-\mu)'\Sigma^{-1}(z-\mu))$$

where f has been normalized so that

$$\int zz'f(z'z)\,dz = I_p.$$

Thus when the density of Z_1 is $p(\cdot|\mu, \Sigma)$, then

$$\mathcal{E}Z_1 = \mu, \qquad \operatorname{Cov} Z_1 = \Sigma.$$

Assuming that $n \geqslant p+1$, the sample covariance matrix

$$S = \sum_1^n (Z_i - \bar{Z})(Z_i - \bar{Z})' = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$$

is positive definite with probability one. Here S_{11} is $q \times q$, S_{22} is $r \times r$, and S_{12} is $q \times r$. Partitioning Σ as S is partitioned, we have

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Thus the squared sample coefficients, $r_1^2 \geqslant \cdots \geqslant r_t^2$, are the t largest eigenvalues of $S_{11}^{-1}S_{12}S_{22}^{-1}S_{21}$ and the squared population coefficients, $\rho_1^2 \geqslant \cdots \geqslant \rho_t^2$, are the t largest eigenvalues of $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. In the present generality, an invariance argument is given to show that the joint distribution of $(r_1,\ldots,r_t)$ depends on (μ,Σ) only through $(\rho_1,\ldots,\rho_t)$. Consider the group G whose elements are $g = (C, c)$ where $c \in R^p$ and

$$C = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}, \qquad A \in Gl_q, \quad B \in Gl_r.$$

The action of G on R^p is

$$(C,c)z = Cz + c$$

and group composition is

$$(C_1,c_1)(C_2,c_2) = (C_1C_2, C_1c_2 + c_1).$$

The group action on the sample is

$$g(Z_1,\dots,Z_n)=(gZ_1,\dots,gZ_n).$$

With the induced group action on (μ,Σ) given by

$$g(\mu,\Sigma)=(g\mu,C\Sigma C')$$

where $g=(C,c)$, it is clear that the family of distributions of $(Z_1,\dots,Z_n)$ that are indexed by elements of

$$\Theta=\{(\mu,\Sigma)|\mu\in R^p,\Sigma\in\mathbb{S}_p^+\}$$

is a G-invariant family of probability measures.

Proposition 10.7. The joint distribution of $(r_1,\dots,r_t)$ depends on (μ,Σ) only through $(\rho_1,\dots,\rho_t)$.

Proof. From Proposition 10.6, we know that $(r_1,\dots,r_t)$ is a G-invariant function of $(Z_1,\dots,Z_n)$. Thus the distribution of $(r_1,\dots,r_t)$ will depend on the parameter $\theta=(\mu,\Sigma)$ only through a maximal invariant in the parameter space. However, Proposition 10.1 shows that $(\rho_1,\dots,\rho_t)$ is a maximal invariant under the action of G on Θ. □

Before discussing the distribution of canonical correlation coefficients, even for $t=1$, it is instructive to consider the bivariate correlation coefficient. Consider pairs of random variables (X_i,Y_i), $i=1,\dots,n$, and let $X\in R^n$ and $Y\in R^n$ have coordinates X_i and Y_i, $i=1,\dots,n$. With $e\in R^n$ being the vector of ones, $P_e=ee'/n$ and $Q_e=I-P_e$, the sample correlation coefficient is defined by

$$r=\left(\frac{Q_eY}{\|Q_eY\|}\right)'\frac{Q_eX}{\|Q_eX\|}.$$

The next result describes the distribution of r when (X_i,Y_i), $i=1,\dots,n$, is a random sample from a bivariate normal distribution.

Proposition 10.8. Suppose $(X_i,Y_i)'\in R^2$, $i=1,\dots,n$, are independent random vectors with

$$\mathcal{L}\begin{pmatrix}X_i\\Y_i\end{pmatrix}=N(\mu,\Sigma),\qquad i=1,\dots,n$$

where $\mu \in R^2$ and

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$$

is positive definite. Consider random variables (U_1, U_2, U_3) with:

(i) (U_1, U_2) independent of U_3.
(ii) $\mathcal{L}(U_3) = \chi^2_{n-2}$.
(iii) $\mathcal{L}(U_2) = \chi^2_{n-1}$.
(iv) $\mathcal{L}(U_1|U_2) = N(\frac{\rho}{\sqrt{1-\rho^2}} U_2^{1/2}, 1)$.

where $\rho = \sigma_{12}/(\sigma_{11}\sigma_{22})^{1/2}$ is the correlation coefficient. Then we have

$$\mathcal{L}\left(\frac{r}{\sqrt{1-r^2}}\right) = \mathcal{L}\left(\frac{U_1}{U_3^{1/2}}\right).$$

Proof. The assumption of independence and normality implies that the matrix $(XY) \in \mathcal{L}_{2,n}$ has a distribution given by

$$\mathcal{L}(XY) = N(e\mu', I_n \otimes \Sigma).$$

It follows from Proposition 10.7 that we may assume, without loss of generality, that Σ has the form

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

When Σ has this form, the conditional distribution of X given Y is

$$\mathcal{L}(X|Y) = N\big((\mu_1 - \rho\mu_2)e + \rho Y, (1-\rho^2)I_n\big)$$

so

$$\mathcal{L}(Q_e X|Y) = N\big(\rho Q_e Y, (1-\rho^2)Q_e\big).$$

Now, let $v_1, \ldots, v_n$ be an orthonormal basis for R^n with $v_1 = e/\sqrt{n}$ and

$$v_2 = \frac{Q_e Y}{\|Q_e Y\|}.$$

Expressing Q_eX in this basis leads to

$$Q_eX=\sum_2^n(v_i'Q_eX)v_i$$

since $Q_ee=0$. Setting

$$\xi_i=\frac{v_i'Q_eX}{\sqrt{1-\rho^2}},\qquad i=2,\ldots,n,$$

it is easily seen that, conditional on Y, we have that $\xi_2,\ldots,\xi_n$ are independent with

$$\mathcal{L}(\xi_2|Y)=N\big(\rho(1-\rho^2)^{-1/2}\|Q_eY\|,1\big)$$

and

$$\mathcal{L}(\xi_i|Y)=N(0,1),\qquad i=3,\ldots,n.$$

Since

$$\|Q_eX\|^2=\sum_2^n(v_i'Q_eX)^2=(1-\rho^2)\sum_2^n\xi_i^2,$$

the identity

$$r=\frac{\xi_2}{\sqrt{\xi_2^2+\Sigma_3^n\xi_i^2}}$$

holds. This leads to

$$\frac{r}{\sqrt{1-r^2}}=\frac{\xi_2}{\sqrt{\Sigma_3^n\xi_i^2}}.$$

Setting $U_1=\xi_2$, $U_2=\|Q_eY\|^2$, and $U_3=\Sigma_3^n\xi_i^2$ yields the assertion of the proposition. □

The result of this proposition has a couple of interesting consequences. When $\rho=0$, then the statistic

$$W=\sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}}=\sqrt{n-2}\,\frac{U_1}{U_3^{1/2}}$$

has a Students t distribution with $n - 2$ degrees of freedom. In the general case, the distribution of W can be described by saying that: conditional on U_2, W has a noncentral t distribution with $n - 2$ degrees of freedom and noncentrality parameter

$$\delta = \frac{\rho}{\sqrt{1-\rho^2}} U_2^{1/2}$$

where $\mathcal{L}(U_2) = \chi^2_{n-1}$. Let $p_m(\cdot|\delta)$ denote the density function of a noncentral t distribution with m degrees of freedom and noncentrality parameter δ. The results in the Appendix show that $p_m(\cdot|\delta)$ has a monotone likelihood ratio. It is clear that the density of W is

$$h(w|\rho) = \int_0^\infty p_m\left(w|\left(\rho(1-\rho^2)^{-1/2}\right)u^{1/2}\right)f(u)\,du$$

where f is the density of U_2 and $m = n - 2$. From this representation and the results in the Appendix, it is not difficult to show that $h(\cdot|\rho)$ has a monotone likelihood ratio. The details of this are left to the reader.

In the case that the two random vectors X and Y in R^n are independent, the conditions under which W has a t_{n-2} distribution can be considerably weakened.

Proposition 10.9. Suppose X and Y in R^n are independent and both $\|Q_e X\|$ and $\|Q_e Y\|$ are positive with probability one. Also assume that, for some number $\mu_1 \in R$, the distribution of $X - \mu_1 e$ is orthogonally invariant. Under these assumptions, the distribution of

$$W = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$$

where

$$r = \left(\frac{Q_e Y}{\|Q_e Y\|}\right)' \frac{Q_e X}{\|Q_e X\|}$$

is a t_{n-2} distribution.

Proof. The two random vectors $Q_e X$ and $Q_e Y$ take values in the $(n-1)$-dimensional subspace

$$M = \{x | x \in R^n, x'e = 0\}.$$

Fix Y so the vector

$$y \equiv \frac{Q_e Y}{\|Q_e Y\|} \in M$$

has length one. Since the distribution of $X - \mu_1 e$ is $\mathcal{O}_n$ invariant, it follows that the distribution of $Q_e X$ is invariant under the group

$$G = \{\Gamma | \Gamma \in \mathcal{O}_n, \Gamma e = e\},$$

which acts on M. Therefore, the distribution of $Q_e X / \|Q_e X\|$ is G-invariant on the set

$$\mathfrak{X} = \{x | x \in M, \|x\| = 1\}.$$

But G is compact and acts transitively on $\mathfrak{X}$ so there is a unique G-invariant distribution for $Q_e X / \|Q_e X\|$ in $\mathfrak{X}$. From this uniqueness it follows that

$$\mathcal{L}\left(\frac{Q_e X}{\|Q_e X\|}\right) = \mathcal{L}\left(\frac{Q_e Z}{\|Q_e Z\|}\right)$$

where $\mathcal{L}(Z) = N(0, I_n)$ on R^n. Therefore, we have

$$\mathcal{L}(r) = \mathcal{L}\left(y' \frac{Q_e Z}{\|Q_e Z\|}\right),$$

and for each y, the claimed result follows from the argument given to prove Proposition 10.8. □

We now turn to the canonical correlation coefficients in the special case that $t = 1$. Consider random vectors X_i and Y_i with $X_i \in R^1$ and $Y_i \in R^r$, $i = 1, \ldots, n$. Let $X \in R^n$ have coordinates $X_1, \ldots, X_n$ and let $Y \in \mathcal{L}_{r,n}$ have rows $Y_1', \ldots, Y_n'$. Assume that $Q_e Y$ has rank r so

$$P \equiv Q_e Y[(Q_e Y)' Q_e Y]^{-1} (Q_e Y)'$$

is the orthogonal projection onto the subspace spanned by the columns of $Q_e Y$. Since $t = 1$, the canonical correlation coefficient is the square root of the largest, and only nonzero, eigenvalue of

$$\frac{(Q_e X)(Q_e X)'}{\|Q_e X\|^2} P,$$

which is

$$r_1^2 \equiv \frac{(Q_eX)'P(Q_eX)}{\|Q_eX\|^2} = \frac{\|PQ_eX\|^2}{\|Q_eX\|^2}.$$

For the case at hand, r_1 is commonly called the *multiple correlation coefficient*. The distribution of r_1^2 is described next under the assumption of normality.

Proposition 10.10. Assume that the distribution of $(XY) \in \mathcal{L}_{r+1,n}$ is given by

$$\mathcal{L}(XY) = N(e\mu', I_n \otimes \Sigma)$$

and partition Σ as

$$\Sigma = \begin{pmatrix} \sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} : (r+1) \times (r+1)$$

where $\sigma_{11} > 0$, Σ_{12} is $1 \times r$, and Σ_{22} is $r \times r$. Consider random variables U_1, U_2, and U_3 whose joint distribution is specified by:

(i) (U_1, U_2) and U_3 are independent.
(ii) $\mathcal{L}(U_3) = \chi^2_{n-r-1}$.
(iii) $\mathcal{L}(U_2) = \chi^2_{n-1}$.
(iv) $\mathcal{L}(U_1|U_2) = \chi^2_r(\Delta)$, where $\Delta = \rho^2(1-\rho^2)^{-1}U_2$.

Here $\rho = (\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}/\sigma_{11})^{1/2}$ is the population multiple correlation coefficient. Then we have

$$\mathcal{L}\left(\frac{r_1^2}{1-r_1^2}\right) = \mathcal{L}\left(\frac{U_1}{U_3}\right).$$

Proof. Combining the results of Proposition 10.1 and Proposition 5.7, without loss of generality, Σ can be assumed to have the form

$$\Sigma = \begin{pmatrix} 1 & \rho\varepsilon_1' \\ \rho\varepsilon_1 & I_r \end{pmatrix}$$

where $\varepsilon_1 \in R^r$ and $\varepsilon_1' = (1, 0, \ldots, 0)$. When Σ has this form, the conditional distribution of X given Y is

$$\mathcal{L}(X|Y) = N\big((\mu_1 - \rho\mu_2'\varepsilon_1)e + \rho Y\varepsilon_1, (1-\rho^2)I_n\big)$$

where $\mathcal{E}X = \mu_1 e$ and $\mathcal{E}Y = e\mu_2'$. Since $Q_e e = 0$, we have

$$\mathcal{L}(Q_e X|Y) = N\big(\rho Q_e Y\varepsilon_1, (1-\rho^2)Q_e\big).$$

The subspace spanned by the columns of $Q_e Y$ is contained in the range of Q_e and this implies that $Q_e P = PQ_e = P$ so

$$\|Q_e X\|^2 = \|(Q_e - P)X\|^2 + \|PX\|^2 = \|(Q_e - P)Q_e X\|^2 + \|PQ_e X\|^2.$$

Since

$$r_1^2 = \frac{\|PQ_e X\|^2}{\|Q_e X\|^2},$$

it follows that

$$\frac{r_1^2}{1-r_1^2} = \frac{\|PQ_e X\|^2}{\|(Q_e - P)Q_e X\|^2}.$$

Given Y, the conditional covariance of $Q_e X$ is $(1-\rho^2)Q_e$ and, therefore, the identity $PQ_e(Q_e - P) = 0$ implies that $PQ_e X$ and $(Q_e - P)Q_e X$ are conditionally independent. It is clear that

$$\mathcal{L}((Q_e - P)Q_e X|Y) = N\big(0, (1-\rho^2)(Q_e - P)\big),$$

so we have

$$\mathcal{L}\big(\|(Q_e - P)Q_e X\|^2|Y\big) = (1-\rho^2)\chi^2_{n-r-1}$$

since $Q_e - P$ is an orthogonal projection of rank $n - r - 1$. Again, conditioning on Y,

$$\mathcal{L}(PQ_e X|Y) = N\big(\rho Q_e Y\varepsilon_1, (1-\rho^2)P\big)$$

since $PQ_e = P$ and $Q_e Y\varepsilon_1$ is in the range of P. It follows from Proposition 3.8 that

$$\mathcal{L}\big(\|PQ_e X\|^2|Y\big) = (1-\rho^2)\chi_r^2(\Delta)$$

where the noncentrality parameter Δ is given by

$$\Delta = \frac{\rho^2}{1-\rho^2}\varepsilon_1'Y'Q_eY\varepsilon_1.$$

That $U_2 \equiv \varepsilon_1'YQ_eY\varepsilon_1$ has a χ^2_{n-1} distribution is clear. Defining U_1 and U_3 by

$$U_1 = (1-\rho^2)^{-1}\|PQ_eX\|^2$$

and

$$U_3 = (1-\rho^2)^{-1}\|(Q_e - P)Q_eX\|^2,$$

the identity

$$\frac{r_1^2}{1-r_1^2} = \frac{U_1}{U_3}$$

holds. That U_3 is independent of (U_1, U_2) follows by conditioning on Y. Since

$$\mathcal{L}(U_1|Y) = \chi_n(\Delta)$$

where

$$\Delta = \frac{\rho^2}{1-\rho^2}\varepsilon_1'Y'Q_eY\varepsilon_1 = \frac{\rho^2}{1-\rho^2}U_2,$$

the conditional distribution of U_1 given Y is the same as the conditional distribution of U_1 given U_2. This completes the proof. □

When $\rho = 0$, Proposition 10.10 shows that

$$\mathcal{L}\left(\frac{r_1^2}{1-r_1^2}\right) = \mathcal{L}\left(\frac{\chi_r^2}{\chi_{n-r-1}^2}\right) = F_{r,n-r-1},$$

which is the unnormalized F-distribution on $(0, \infty)$. More generally,

$$\mathcal{L}\left(\frac{r_1^2}{1-r_1^2}\right) = F(r, n-r-1; \Delta)$$

where

$$\Delta = \frac{\rho^2}{1-\rho^2}\chi^2_{n-1}$$

is random. This means that, conditioning on $\Delta = \delta$,

$$\mathcal{L}\left(\frac{r_1^2}{1-r_1^2}\middle|\delta\right) = F(r, n-r-1; \delta).$$

Let $f(\cdot|\delta)$ denote the density function of an $F(r, n-r-1; \delta)$ distribution, and let $h(\cdot)$ be the density of a χ^2_{n-1} distribution. Then the density of $r_1^2/(1-r_1^2)$ is

$$k(w|\rho) = \int_0^\infty f\left(w|\rho^2(1-\rho^2)^{-1}u\right)h(u)\,du.$$

From this representation, it can be shown, using the results in the Appendix, that $k(w|\rho)$ has a monotone likelihood ratio.

The final exact distributional result of this section concerns the function of the sample canonical correlations given by

$$W = \prod_{i=1}^{t}(1-r_i^2)$$

when the random sample $(X_i, Y_i)'$, $i = 1,\ldots, n$, is from a normal distribution and the population coefficients are all zero. This statistic arises in testing for independence, which is discussed in detail in the next section. To be precise, it is assumed that the random sample

$$Z_i = \begin{pmatrix} X_i \\ Y_i \end{pmatrix} \in R^p, \qquad i = 1,\ldots, n$$

satisfies

$$\mathcal{L}\begin{pmatrix} X_i \\ Y_i \end{pmatrix} = N(\mu, \Sigma).$$

As usual, $X_i \in R^q$, $Y_i \in R^r$, and the sample covariance matrix

$$S = \sum_1^n (Z_i - \bar{Z})(Z_i - \bar{Z})'$$

is partitioned as

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$$

where S_{11} is $q \times q$ and S_{22} is $r \times r$. Under the assumptions made this far, S has a Wishart distribution—namely,

$$\mathcal{L}(S) = W(\Sigma, p, n-1).$$

Partitioning Σ, we have

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

and the population canonical correlation coefficients, say $\rho_1 \geqslant \cdots \geqslant \rho_t$, are all zero iff $\Sigma_{12} = 0$.

Proposition 10.11. Assume $n - 1 \geqslant p$ and let $r_1 \geqslant \cdots \geqslant r_t$ be the sample canonical correlations. When $\Sigma_{12} = 0$, then

$$\mathcal{L}\left(\prod_1^t (1 - r_i^2)\right) = U(n - r - 1, r, q)$$

where the distribution $U(n - r - 1, r, q)$ is described in Proposition 8.14.

Proof. Since $r_1^2, \ldots, r_t^2$ are the t largest eigenvalues of

$$\Lambda(S) = S_{11}^{-1} S_{12} S_{22}^{-1} S_{21}$$

and the remaining $q - t$ eigenvalues of $\Lambda(S)$ are zero, it follows that

$$W \equiv \prod_{i-1}^{t} (1 - r_i^2) = |I_q - S_{11}^{-1} S_{12} S_{22}^{-1} S_{21}|.$$

Since W is a function of the sample canonical correlations and $\Sigma_{12} = 0$, Proposition 10.1 implies that we can take

$$\Sigma = \begin{pmatrix} I_q & 0 \\ 0 & I_r \end{pmatrix}$$

without loss of generality to find the distribution of W. Using properties of determinants, we have

$$W = |S_{11}^{-1}||S_{11} - S_{12}S_{22}^{-1}S_{21}| = \frac{|S_{11\cdot 2}|}{|S_{11\cdot 2} + S_{12}S_{22}^{-1}S_{21}|}.$$

Proposition 8.7 implies that

$$\mathcal{L}(S_{11\cdot 2}) = W(I_q, q, n - r - 1)$$

$$\mathcal{L}\left(S_{22}^{-1/2}S_{21}|S_{22}\right) = N(0, I_r \otimes I_q)$$

and $S_{11\cdot 2}$ and $S_{12}S_{22}^{-1}S_{21}$ are independent. Therefore,

$$\mathcal{L}\left(S_{12}S_{22}^{-1}S_{21}\right) = W(I_q, q, r)$$

and by definition, it follows that

$$\mathcal{L}(W) = U(n - r - 1, r, q). \qquad \square$$

Since

$$W = \frac{|S_{22\cdot 1}|}{|S_{22\cdot 1} + S_{21}S_{11}^{-1}S_{12}|},$$

the proof of Proposition 10.11 shows that $\mathcal{L}(W) = U(n - q - 1, q, r)$ so $U(n - q - 1, q, r) = U(n - r - 1, r, q)$ as long as $n - 1 \geqslant q + r$. Using the ideas in the proof of Proposition 8.15, the distribution of W can be derived when Σ_{12} has rank one—that is, when one population canonical correlation is positive and the rest are zero. The details of this are left to the reader.

We close this section with a discussion that provides some qualitative information about the distribution of $r_1 \geqslant \cdots \geqslant r_t$ when the data matrices $X \in \mathcal{L}_{q,n}$ and $Y \in \mathcal{L}_{r,n}$ are independent. As usual, let P_X and P_Y denote the orthogonal projections onto the column spaces of Q_eX and Q_eY. Then the sample canonical correlations are the t largest eigenvalues of P_YP_X—say

$$\varphi(P_YP_x) \equiv \begin{pmatrix} r_1 \\ \vdots \\ r_t \end{pmatrix} \in R^t.$$

It is assumed that Q_eX has rank q and Q_eY has rank r. Since the distribution of $\varphi(P_YP_X)$ is of interest, it is reasonable to investigate the distributional properties of the two random projections P_X and P_Y. Since X and Y are assumed to be independent, it suffices to focus our attention on P_X. First note that P_X is an orthogonal projection onto a q-dimensional subspace contained in

$$M = \{x | x \in R^n, x'e = 0\}.$$

Therefore, P_X is an element of

$$\mathcal{P}_{q,n}(e) = \left\{ P \middle| \begin{array}{c} P \text{ is an } n \times n \text{ rank } q \text{ orthogonal} \\ \text{projection, } Pe = 0 \end{array} \right\}$$

Furthermore, the space $\mathcal{P}_{q,n}(e)$ is a compact subset of R^{n^2} and is acted on by the compact group

$$\mathcal{O}_n(e) = \{\Gamma | \Gamma \in \mathcal{O}_n, \Gamma e = e\},$$

with the group action given by $P \to \Gamma P \Gamma'$. Since $\mathcal{O}_n(e)$ acts transitively on $\mathcal{P}_{q,n}(e)$, there is a unique $\mathcal{O}_n(e)$-invariant probability distribution on $\mathcal{P}_{q,n}(e)$. This is called the *uniform distribution* on $\mathcal{P}_{q,n}(e)$.

Proposition 10.12. If $\mathcal{L}(X) = \mathcal{L}(\Gamma X)$ for $\Gamma \in \mathcal{O}_n(e)$, then P_X has a uniform distribution on $\mathcal{P}_{q,n}(e)$.

Proof. It is readily verified that

$$P_{\Gamma X} = \Gamma P_X \Gamma', \qquad \Gamma \in \mathcal{O}_n(e).$$

Therefore, if $\mathcal{L}(\Gamma X) = \mathcal{L}(X)$, then

$$\mathcal{L}(P_X) = \mathcal{L}(\Gamma P_X \Gamma'),$$

which implies that the distribution $\mathcal{L}(P_X)$ on $\mathcal{P}_{q,n}(e)$ is $\mathcal{O}_n(e)$-invariant. The uniqueness of the uniform distribution on $\mathcal{P}_{q,n}(e)$ yields the result. □

When $\mathcal{L}(X) = N(e\mu_1', I_n \otimes \Sigma_{11})$, then $\mathcal{L}(X) = \mathcal{L}(\Gamma X)$ for $\Gamma \in \mathcal{O}_n(e)$, so Proposition 10.12 applies to this case. For any two $n \times n$ positive semidefinite matrices B_1 and B_2, define the function $\varphi(B_1 B_2)$ to be the vector of the t largest eigenvalues of $B_1 B_2$. In particular, $\varphi(P_Y P_X)$ is the vector of sample canonical correlations.

Proposition 10.13. Assume X and Y are independent, $\mathcal{L}(\Gamma X) = \mathcal{L}(X)$ for $\Gamma \in \mathcal{O}_n(e)$, Q_eX has rank q, and Q_eY has rank r. Then

$$\mathcal{L}(\varphi(P_YP_X)) = \mathcal{L}(\varphi(P_0P_X))$$

where P_0 is any fixed rank r projection in $\mathcal{P}_{r,n}(e)$.

Proof. First note that

$$\varphi(P_Y\Gamma P_X\Gamma') = \varphi(\Gamma' P_Y \Gamma P_X)$$

since the eigenvalues of $P_Y\Gamma P_X\Gamma'$ are the same as the eigenvalues of $\Gamma' P_Y\Gamma P_X$. From Proposition 10.12, we have

$$\mathcal{L}(P_X) = \mathcal{L}(\Gamma P_X\Gamma'), \qquad \Gamma \in \mathcal{O}_n(e).$$

Conditioning on Y, the independence of X and Y implies that

$$\mathcal{L}(\varphi(P_YP_X)|Y) = \mathcal{L}(\varphi(P_Y\Gamma P_X\Gamma')|Y) = \mathcal{L}(\varphi(\Gamma' P_Y\Gamma P_X)|Y)$$

for all $\Gamma \in \mathcal{O}_n(e)$. The group $\mathcal{O}_n(e)$ acts transitively on $\mathcal{P}_{r,n}(e)$, so for Y fixed, there exists a $\Gamma \in \mathcal{O}_n(e)$ such that $\Gamma' P_Y \Gamma = P_0$. Therefore, the equation

$$\mathcal{L}(\varphi(P_YP_X)|Y) = \mathcal{L}(\varphi(P_0P_X)|Y) = \mathcal{L}(\varphi(P_0P_X))$$

holds for each Y since X and Y are independent. Averaging $\mathcal{L}(\varphi(P_YP_X)|Y)$ over Y yields $\mathcal{L}(\varphi(P_YP_X))$, which must then equal $\mathcal{L}(\varphi(P_0P_X))$. This completes the proof. □

The preceeding result shows that $\mathcal{L}(\varphi(P_YP_X))$ does not depend on the distribution of Y as long as X and Y are independent and $\mathcal{L}(X) = \mathcal{L}(\Gamma X)$ for $\Gamma \in \mathcal{O}_n(e)$. In this case, the distribution of $\varphi(P_YP_X)$ can be derived under the assumption that $\mathcal{L}(X) = N(0, I_n \otimes I_q)$ and $\mathcal{L}(Y) = N(0, I_n \otimes I_r)$. Suppose that $q \leqslant r$ so $t = q$. Then $\mathcal{L}(\varphi(P_YP_X))$ is the distribution of $r_1 \geqslant \cdots \geqslant r_q$ where $\lambda_i = r_i^2$, $i = 1, \ldots, q$, are the eigenvalues of $S_{11}^{-1}S_{12}S_{22}^{-1}S_{21}$ and

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$$

is the sample covariance matrix. To find the distribution of $r_1, \ldots, r_q$, it would obviously suffice to find the distribution of $\gamma_i = 1 - \lambda_i$, $i = 1, \ldots, q$, which are the eigenvalues of

$$I_q - S_{11}^{-1}S_{12}S_{22}^{-1}S_{21} = (T_1 + T_2)^{-1}T_1$$

where

$$T_1 = S_{11} - S_{12}S_{22}^{-1}S_{21}; \qquad T_2 = S_{12}S_{22}^{-1}S_{21}.$$

It was shown in the proof of Proposition 10.11 that T_1 and T_2 are independent and

$$\mathcal{L}(T_1) = W(I_q, q, n - r - 1)$$

and

$$\mathcal{L}(T_2) = W(I_q, q, r).$$

Since the matrix

$$B = (T_1 + T_2)^{-1/2}T_1(T_1 + T_2)^{-1/2}$$

has the same eigenvalues as $(T_1 + T_2)^{-1}T_1$, it suffices to find the distribution of the eigenvalues of B. It is not too difficult to show (see the Problems at end of this chapter) that the density of B is

$$p(B) = \frac{\omega(n - r - 1, q)\omega(r, q)}{\omega(n - 1, q)}|B|^{(n-r-q-2)/2}|I_q - B|^{(r-q-1)/2}$$

with respect to Lebesgue measure dB restricted to the set

$$\mathcal{X} = \{B | B \in \mathbb{S}_q^+, I_q - B \in \mathbb{S}_q^+\}.$$

Here, $\omega(\cdot, \cdot)$ is the Wishart constant defined in Example 5.1. Now, the ordered eigenvalues of B are a maximal invariant under the action of the group $\mathcal{O}_q$ on $\mathcal{X}$ given by $B \to \Gamma B \Gamma'$, $\Gamma \in \mathcal{O}_q$. Let λ be the vector of ordered eigenvalues of B so $\lambda \in R^q$, $1 \geqslant \lambda_1 \geqslant \cdots \geqslant \lambda_q \geqslant 0$. Since $p(\Gamma B \Gamma') = p(B)$, $\Gamma \in \mathcal{O}_q$, it follows from Proposition 7.15 that the density of λ is $q(\lambda) = p(D_\lambda)$ where D_λ is a $q \times q$ diagonal matrix with diagonal entries $\lambda_1, \ldots, \lambda_q$. Of course, $q(\cdot)$ is the density of λ with respect to the measure $\nu(d\lambda)$ induced by the maximal invariant mapping. More precisely, let

$$\mathcal{Z} = \{\lambda | \lambda \in R^q, 1 \geqslant \lambda_1 \geqslant \cdots \geqslant \lambda_q \geqslant 0\}$$

and consider the mapping φ on $\mathcal{X}$ to $\mathcal{Z}$ defined by $\varphi(B) = \lambda$ where λ is the vector of eigenvalues of B. For any Borel set $C \subseteq \mathcal{Z}$, $\nu(C)$ is defined by

$$\nu(C) = \int_{\varphi^{-1}(C)} dB.$$

Since $q(\lambda)$ has been calculated, the only step left to determine the distribution of λ is to find the measure ν. However, it is rather nontrivial to find ν and the details are not given here. We have included the above argument to show that the only step in obtaining $\mathcal{L}(\lambda)$ that we have not solved is the calculation of ν. This completes our discussion of distributional problems associated with canonical correlations.

The measure ν above is just the restriction to $\mathcal{Z}$ of the measure ν_2 discussed in Example 6.1. For one derivation of ν_2, see Muirhead (1982, p. 104).

10.4. TESTING FOR INDEPENDENCE

In this section, we consider the problem of testing for independence based on a random sample from a normal distribution. Again, let $Z_1, \dots, Z_n$ be independent random vectors in R^p and partition Z_i as

$$Z_i = \begin{pmatrix} X_i \\ Y_i \end{pmatrix}, \qquad X_i \in R^q, \qquad Y_i \in R^r.$$

It is assumed that $\mathcal{L}(Z_i) = N(\mu, \Sigma)$, so

$$\operatorname{Cov}(Z_i) = \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \operatorname{Cov}\begin{pmatrix} X_i \\ Y_i \end{pmatrix}$$

for $i = 1, \dots, n$. The problem is to test the null hypothesis $H_0 : \Sigma_{12} = 0$ against the alternative $H_1 : \Sigma_{12} \neq 0$. As usual, let Z have rows Z_i', $i = 1, \dots, n$ so $\mathcal{L}(Z) = N(e\mu', I_n \otimes \Sigma)$. Assuming $n \geqslant p + 1$, the set $\mathcal{Z} \subseteq \mathcal{L}_{p,n}$ where

$$S \equiv (Z - e\bar{Z}')'(Z - e\bar{Z}') = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$$

has rank p is a set of probability one and $\mathcal{Z}$ is taken as the sample space for Z. The group G considered in Proposition 10.6 acts on $\mathcal{Z}$ and a maximal invariant is the vector of canonical correlation coefficients $r_1, \dots, r_t$ where $t = \min\{q, r\}$.

Proposition 10.14. The problem of testing $H_0 : \Sigma_{12} = 0$ versus $H_1 : \Sigma_{12} \neq 0$ is invariant under G. Every G-invariant test is a function of the sample canonical correlation coefficients $r_1, \dots, r_t$. When $t = 1$, the test that rejects for large values of r_1 is a uniformly most powerful invariant test.

Proof. That the testing problem is G-invariant is easily checked. From Proposition 10.6, the function mapping Z into $r_1, \dots, r_t$ is a maximal invariant so every invariant test is a function of $r_1, \dots, r_t$. When $t = 1$, the test that rejects for large values of r_1 is equivalent to the test that rejects for large values of $U \equiv r_1^2/(1 - r_1^2)$. It was argued in the last section (see Proposition 10.10) that the density of U, say $k(u|\rho)$, has a monotone likelihood ratio where ρ is the only nonzero population canonical correlation coefficient. Since the null hypothesis is that $\rho = 0$ and since every invariant test is a function of U, it follows that the test that rejects for large values of U is a uniformly most powerful invariant test.

When $t = 1$, the distribution of U is specified in Proposition 10.10, and this can be used to construct a test of level α for H_0. For example, if $q = 1$, then $\mathcal{L}(U) = F_{r, n-r-1}$ and a constant $c(\alpha)$ can be found from standard tables of the normalized $\mathcal{F}$-distribution such that, under H_0, $P\{U > c(\alpha)\} = \alpha$.

In the case that $t > 1$, there is no obvious function of $r_1, \dots, r_t$ that provides an optimum test of H_0 versus H_1. Intuitively, if some of the r_i's are "too big," there is reason to suspect that H_0 is not true. The likelihood ratio test provides one possible criterion for testing $\Sigma_{12} = 0$.

Proposition 10.15. The likelihood ratio test of H_0 versus H_1 rejects if the statistic

$$W = \prod_{i=1}^{t} \left(1 - r_i^2\right) = \frac{|S|}{|S_{11}||S_{22}|}$$

is too small. Under H_0, $\mathcal{L}(W) = U(n - r - 1, r, q)$, which is the distribution described in Proposition 8.14.

Proof. The density function of Z is

$$p(Z|\mu, \Sigma) = \left(\sqrt{2\pi}\right)^{-np} |\Sigma|^{-n/2} \exp\left[-\tfrac{1}{2}\operatorname{tr}(Z - e\mu')'(Z - e\mu')\Sigma^{-1}\right].$$

Under both H_0 and H_1, the maximum likelihood estimate of μ is $\hat{\mu} = \bar{Z}$. Under H_1, the maximum likelihood estimate of Σ is $\hat{\Sigma} = (1/n)S$. Partitioning S as Σ is partitioned, we have

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$$

where S_{11} is $q \times q$, S_{12} is $q \times r$, and S_{22} is $r \times r$. Under H_0, Σ has the form

$$\Sigma = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}$$

so

$$\Sigma^{-1} = \begin{pmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{pmatrix}.$$

When Σ has this form,

$$\begin{aligned} p(Z|\hat{\mu}, \Sigma) &= (\sqrt{2\pi})^{-np}|\Sigma_{11}|^{-n/2}|\Sigma_{22}|^{-n/2}\exp[-\tfrac{1}{2}\operatorname{tr} S\Sigma^{-1}] \\ &= (\sqrt{2\pi})^{-np}|\Sigma_{11}|^{-n/2}\exp[-\tfrac{1}{2}\operatorname{tr} S_{11}\Sigma_{11}^{-1}]|\Sigma_{22}|^{-n/2} \\ &\quad \times \exp[-\tfrac{1}{2}\operatorname{tr} S_{22}\Sigma_{22}^{-1}]. \end{aligned}$$

From this it is clear that, under H_0, $\hat{\Sigma}_{11} = (1/n)S_{11}$ and $\hat{\Sigma}_{22} = (1/n)S_{22}$. Substituting these estimates into the densities under H_0 and H_1 leads to a likelihood ratio of

$$\Lambda(Z) = \left(\frac{|S_{11}||S_{22}|}{|S|}\right)^{-n/2}.$$

Rejecting H_0 for small values of $\Lambda(Z)$ is equivalent to rejecting for small values of

$$W = (\Lambda(Z))^{2/n} = \frac{|S|}{|S_{11}||S_{22}|}.$$

The identity $|S| = |S_{11}||S_{22} - S_{21}S_{11}^{-1}S_{12}|$ shows that

$$W = \frac{|S_{22} - S_{21}S_{11}^{-1}S_{21}|}{|S_{22}|} = |I_r - S_{22}^{-1}S_{21}S_{11}^{-1}S_{12}| = \prod_{i=1}^{t}(1 - r_i^2)$$

where $r_1^2, \ldots, r_t^2$ are the t largest eigenvalues of $S_{22}^{-1}S_{21}S_{11}^{-1}S_{12}$. Thus the likelihood ratio test is equivalent to the test that rejects for small values of W. That $\mathcal{L}(W) = U(n-r-1, r, q)$ under H_0 follows from Proposition 10.11. □

The distribution of W under H_1 is quite complicated to describe except in the case that Σ_{12} has rank one. As mentioned in the last section, when Σ_{12} has rank one, the methods used in Proposition 8.15 yields a description of the distribution of W.

Rather than discuss possible alternatives to the likelihood test, in the next section we show that the testing problem above is a special case of the MANOVA testing problem considered in Chapter 9. Thus the alternatives to the likelihood ratio test for the MANOVA problem are also alternatives to the likelihood ratio test for independence.

We now turn to a slight generalization of the problem of testing that $\Sigma_{12} = 0$. Again suppose that $Z \in \mathfrak{X}$ satisfies $\mathcal{L}(Z) = N(e\mu', I_n \otimes \Sigma)$ where $\mu \in R^p$ and Σ are both unknown parameters and $n \geqslant p+1$. Given an integer $k \geqslant 2$, let $p_1, \dots, p_k$ be positive integers such that $\Sigma_1^k p_i = p$. Partition Σ into blocks Σ_{ij} of dimension $p_i \times p_j$ for $i, j = 1, \dots, k$. We now discuss the likelihood ratio test for testing $H_0: \Sigma_{ij} = 0$ for all i, j with $i \neq j$. For example, when $k = p$ and each $p_i = 1$, then the null hypothesis is that Σ is diagonal with unknown diagonal elements. By mimicking the proof of Proposition 10.15, it is not difficult to show that the likelihood ratio test for testing H_0 versus the alternative that Σ is completely unknown rejects for small values of

$$\Lambda = \frac{|S|}{\prod_{i=1}^{k} |S_{ii}|}.$$

Here, $S = (Z - e\bar{Z}')'(Z - e\bar{Z}')$ is partitioned into $S_{ij}: p_i \times p_j$ as Σ was partitioned. Further, for $i = 1, \dots, k$, define $S_{(ii)}$ by

$$S_{(ii)} = \begin{pmatrix} S_{ii} & S_{i(i+1)} & \cdots & S_{ik} \\ & & & \vdots \\ & & & S_{kk} \end{pmatrix}$$

so $S_{(ii)}$ is $(p_i + \cdots + p_k) \times (p_i + \cdots + p_k)$. Noting that $S_{(11)} = S$, we can write

$$\Lambda = \frac{|S|}{\prod_{i=1}^{k} |S_{ii}|} = \prod_{i=1}^{k-1} \frac{|S_{(ii)}|}{|S_{ii}||S_{(i+1,i+1)}|}.$$

Define W_i, $i = 1, \dots, k - 1$, by

$$W_i = \frac{|S_{(ii)}|}{|S_{ii}||S_{(i+1,i+1)}|}.$$

Under the null hypothesis, it follows from Proposition 10.11 that

$$\mathcal{L}(W_i) = U\left(n - 1 - \sum_{j=i+1}^{k} p_j, \sum_{j=i+1}^{k} p_j, p_i\right).$$

To derive the distribution of Λ under H_0, we now show that $W_1, \dots, W_{k-1}$ are independent random variables under H_0. From this it follows that, under H_0,

$$\mathcal{L}(\Lambda) = \prod_{i=1}^{k-1} U\left(n - 1 - \sum_{j=i+1}^{k} p_j, \sum_{j=i+1}^{k} p_j, p_i\right)$$

so Λ is distributed as a product of independent beta random variables. The independence of $W_1, \dots, W_{k-1}$ for a general k follows easily by induction once independence has been verified for $k = 3$.

For $k = 3$, we have

$$\Lambda = W_1 W_2 = \frac{|S|}{|S_{11}||S_{(22)}|} \frac{|S_{(22)}|}{|S_{22}||S_{33}|}$$

and, under H_0,

$$\mathcal{L}(S) = W(\Sigma, p, n - 1)$$

where Σ has the form

$$\Sigma = \begin{pmatrix} \Sigma_{11} & 0 & 0 \\ 0 & \Sigma_{22} & 0 \\ 0 & 0 & \Sigma_{33} \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{(22)} \end{pmatrix}.$$

To show W_1 and W_2 are independent, Proposition 7.19 is applied. The sample space for S is $\mathbb{S}_p^+$—the space of $p \times p$ positive definite matrices. Fix Σ of the above form and let P_0 denote the probability measure of S so P_0 is the probability measure of a $W(\Sigma, p, n - 1)$ distribution on $\mathbb{S}_p^+$. Consider the group G whose elements are (A, B) where $A \in Gl_{p_1}$ and $B \in Gl_{(p_2+p_3)}$

and the group composition is

$$(A_1, B_1)(A_2, B_2) = (A_1A_2, B_1B_2).$$

It is easy to show that the action

$$(A, B)[S] \equiv \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} S \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}'$$

defines a left action of G on $\mathcal{S}_p^+$. If $\mathcal{L}(S) = W(\Sigma, p, n - 1)$, then

$$\mathcal{L}((A, B)[S]) = W((A, B)[\Sigma], p, n - 1)$$

where

$$(A, B)[\Sigma] = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} \Sigma \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}' = \begin{pmatrix} A\Sigma_{11}A_1 & 0 \\ 0 & B\Sigma_{(22)}B' \end{pmatrix}.$$

This last equality follows from the special form of Σ. The first thing to notice is that

$$W_1 = W_1(S) = \frac{|S|}{|S_{11}||S_{(22)}|}$$

is invariant under the action of G on $\mathcal{S}_p^+$. Also, because of the special form of Σ, the statistic

$$\tau(S) \equiv (S_{11}, S_{(22)}) \in \mathcal{S}_{p_1}^+ \times \mathcal{S}_{(p_2+p_3)}^+$$

is a sufficient statistic for the family of distributions $\{gP_0 | g \in G\}$. This follows from the factorization criterion applied to the family $\{gP_0 | g \in G\}$, which is the Wishart family

$$\left\{ W(\Sigma, p, n - 1) | \Sigma = \begin{pmatrix} \gamma_{11} & 0 \\ 0 & \gamma_{22} \end{pmatrix}, \quad \gamma_{11} \in \mathcal{S}_{p_1}^+, \gamma_{22} \in \mathcal{S}_{(p_2+p_3)}^+ \right\}.$$

However, G acts transitively on $\mathcal{S}_{p_1}^+ \times \mathcal{S}_{(p_2+p_3)}^+$ in the obvious way:

$$(A, B)[S_1, S_2] \equiv (AS_1A', BS_2B')$$

for $[S_1, S_2] \in \mathcal{S}_{p_1}^+ \times \mathcal{S}_{(p_2+p_3)}^+$. Further, the sufficient statistic $\tau(S) \in \mathcal{S}_{p_1}^+ \times \mathcal{S}_{(p_2+p_3)}^+$ satisfies

$$\tau((A, B)[S]) = (A, B)[\tau(S)]$$

so $\tau(\cdot)$ is an equivariant function. It now follows from Proposition 7.19 that the invariant statistic $W_1(S)$ is independent of the sufficient statistic $\tau(S)$. But

$$W_2(S) = \frac{|S_{(22)}|}{|S_{22}||S_{33}|}$$

is a function of $S_{(22)}$ and so is a function of $\tau(S) = [S_{11}, S_{(22)}]$. Thus W_1 and W_2 are independent for each value of Σ in the null hypothesis. Summarizing, we have the following result.

Proposition 10.16. Assume $k = 3$ and Σ has the form specified under H_0. Then, under the action of the group G on both $\mathbb{S}_p^+$ and $\mathbb{S}_{p_1}^+ \times \mathbb{S}_{(p_2+p_3)}^+$, the invariant statistic

$$W_1(S) = \frac{|S|}{|S_{11}||S_{(22)}|}$$

and the equivariant statistic

$$\tau(S) = \left[S_{11}, S_{(22)}\right]$$

are independent. In particular, the statistic

$$W_2(S) = \frac{|S_{(22)}|}{|S_{22}||S_{33}|},$$

being a function of $\tau(S)$, is independent of W_1.

The application and interpretation of the previous paragraph for general k should be fairly clear. The details are briefly outlined. Under the null hypothesis that $\Sigma_{ij} = 0$ for $i, j = 1, \ldots, k$ and $i \neq j$, we want to describe the distribution of

$$\Lambda = \prod_{i=1}^{k-1} \frac{|S_{(ii)}|}{|S_{ii}||S_{(i+1,i+1)}|} = \prod_{i=1}^{k-1} W_i.$$

It was remarked earlier that each W_i is distributed as a product of independent beta random variables. To see that $W_1, \ldots, W_{k-1}$ are independent, Proposition 10.16 shows that

$$W_1 = \frac{|S|}{|S_{11}||S_{(22)}|}$$

and $S_{(22)}$ are independent. Since $(W_2, \ldots, W_{k-1})$ is a function of $S_{(22)}$, W_1 and $(W_2, \ldots, W_{k-1})$ are independent. Next, apply Proposition 10.16 to $S_{(22)}$ to conclude that

$$W_2 = \frac{|S_{(22)}|}{|S_{22}||S_{(33)}|}$$

and $S_{(33)}$ are independent. Since $(W_3, \ldots, W_{k-1})$ is a function of $S_{(33)}$, W_2 and $(W_3, \ldots, W_{k-1})$ are independent. The conclusion that $W_1, \ldots, W_{k-1}$ are independent now follows easily. Thus the distribution of Λ under H_0 has been described.

To interpret the decomposition of Λ into the product $\prod_1^{k-1} W_i$, first consider the null hypothesis

$$H_0^{(1)}: \Sigma_{1j} = 0 \quad \text{for } j = 2, \ldots, k.$$

An application of Proposition 10.15 shows that the likelihood ratio test of $H_0^{(1)}$ versus the alternative that Σ is unknown rejects for small values of

$$W_1 = \frac{|S|}{|S_{11}||S_{(22)}|}.$$

Assuming $H_0^{(1)}$ to be true, consider testing

$$H_0^{(2)}: \Sigma_{2j} = 0 \quad \text{for } j = 3, \ldots, k$$

versus

$$H_1^{(2)}: \Sigma_{2j} \neq 0 \quad \text{for some } j = 3, \ldots, k.$$

A minor variation of the proof of Proposition 10.15 yields a likelihood ratio test of $H_0^{(2)}$ versus $H_1^{(2)}$ (given $H_0^{(1)}$) that rejects for small values of

$$W_2 = \frac{|S_{(22)}|}{|S_{22}||S_{(33)}|}.$$

Proceeding by induction, assume null hypotheses $H_0^{(i)}$, $i = 1, \ldots, m-1$, to be true and consider testing

$$H_0^{(m)}: \Sigma_{mj} = 0, \qquad j = m+1, \ldots, k$$

versus

$$H_1^{(m)}: \Sigma_{mj} \neq 0 \quad \text{for some } j = m+1, \ldots, k.$$

Given the null hypotheses $H_0^{(i)}$, $i = 1, \ldots, m - 1$, the likelihood ratio test of $H_0^{(m)}$ versus $H_1^{(m)}$ rejects for small values of

$$W_m = \frac{|S_{(mm)}|}{|S_{mm}||S_{(m+1,\, m+1)}|}.$$

The overall likelihood ratio test is one possible way of combining the likelihood ratio tests of $H_0^{(m)}$ versus $H_1^{(m)}$, given that $H_0^{(i)}$, $i = 1, \ldots, m - 1$, is true.

10.5. MULTIVARIATE REGRESSION

The purpose of this section is to show that testing for independence can be viewed as a special case of the general MANOVA testing problem treated in Chapter 9. In fact, the results below extend those of the previous section by allowing a more general mean structure for the observations. In the notation of the previous section, consider a data matrix $Z: n \times p$ that is partitioned as $Z = (XY)$ where X is $n \times q$ and Y is $n \times r$ so $p = q + r$. It is assumed that

$$\mathcal{L}(Z) = N(TB, I_n \otimes \Sigma)$$

where T is an $n \times k$ known matrix of rank k and B is a $k \times p$ matrix of unknown parameters. As usual, Σ is a $p \times p$ positive definite matrix. This is precisely the linear model discussed in Section 9.1 and clearly includes the model of previous sections of this chapter as a special case.

To test that X and Y are independent, it is illuminating to first calculate the conditional distribution of Y given X. Partition the matrix B as $B = (B_1 B_2)$ where B_1 is $k \times q$ and B_2 is $k \times r$. In describing the conditional distribution of Y given X, say $\mathcal{L}(Y|X)$, the notation

$$\Sigma_{22\cdot 1} \equiv \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

is used. Following Example 3.1, we have

$$\begin{aligned}\mathcal{L}(Y|X) &= N\big(TB_2 + (I_n \otimes \Sigma_{21}\Sigma_{11}^{-1})(X - TB_1), I_n \otimes \Sigma_{22\cdot 1}\big)\\ &= N\big(T(B_2 - B_1\Sigma_{11}^{-1}\Sigma_{12}) + X\Sigma_{11}^{-1}\Sigma_{12}, I_n \otimes \Sigma_{22\cdot 1}\big)\end{aligned}$$

and the marginal distribution of X is

$$\mathcal{L}(X) = N(TB_1, I_n \otimes \Sigma_{11}).$$

Let W be the $n \times (q + k)$ matrix (XT) and let C be the $(q + k) \times r$ matrix of parameters

$$C = \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} \equiv \begin{pmatrix} \Sigma_{11}^{-1}\Sigma_{12} \\ B_2 - B_1\Sigma_{11}^{-1}\Sigma_{12} \end{pmatrix}$$

so

$$X\Sigma_{11}^{-1}\Sigma_{12} + T(B_2 - B_1\Sigma_{11}^{-1}\Sigma_{12}) = (XT)\begin{pmatrix} C_1 \\ C_2 \end{pmatrix} = WC.$$

In this notation, we have

$$\mathcal{L}(Y|X) = N(WC, I_n \otimes \Sigma_{22\cdot 1})$$

and

$$\mathcal{L}(X) = N(TB_1, I_n \otimes \Sigma_{11}).$$

Assuming $n \geqslant p + k$, the matrix W has rank $q + k$ with probability one so the conditional model for Y is of the MANOVA type. Further, testing $H_0 : \Sigma_{12} = 0$ versus $H_1 : \Sigma_{12} \neq 0$ is equivalent to testing $\tilde{H}_0 : C_1 = 0$ versus $\tilde{H}_1 : C_1 \neq 0$. In other words, based on the model for Z,

$$\mathcal{L}(Z) = N(TB, I_n \otimes \Sigma),$$

the null hypothesis concerns the covariance matrix. But in terms of the conditional model, the null hypothesis concerns the matrix of regression parameters.

With the above discussion and models in mind, we now want to discuss various approaches to testing H_0 and $\tilde{H}_0$. In terms of the model

$$\mathcal{L}(Z) = N(TB, I_n \otimes \Sigma)$$

and assuming H_1, the maximum likelihood estimators of B and Σ are

$$\hat{B} = (T'T)^{-1}T'Z, \qquad \hat{\Sigma} = \frac{1}{n}S$$

where

$$S = (Z - T\hat{B})'(Z - T\hat{B}),$$

so

$$\mathcal{L}(S) = W(\Sigma, p, n - k).$$

Under H_0, the maximum likelihood estimator of B is still $\hat{B}$ as above and since

$$\Sigma = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix},$$

it is readily verified that

$$\hat{\Sigma}_{ii} = \frac{1}{n} S_{ii}, \qquad i = 1, 2$$

where

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}.$$

Substituting these estimators into the density of Z under H_0 and H_1 demonstrates that the likelihood ratio test rejects for small values of

$$\Lambda(Z) = \frac{|S|}{|S_{11}||S_{22}|}.$$

Under H_0, the proof of Proposition 10.11 shows that the distribution of $\Lambda(Z)$ is $U(n - k - r, r, q)$ as described in Proposition 8.14. Of course, symmetry in r and q implies that $U(n - k - r, r, q) = U(n - k - q, q, r)$. An alternative derivation of this likelihood ratio test can be given using the conditional distribution of Y given X and the marginal distribution of X. This follows from two facts: (i) the density of Z is proportional to the conditional density of Y given X multiplied by the marginal density of X, and (ii) the relabeling of the parameters is one-to-one—namely, the mapping from (B, Σ) to $(C, B_1, \Sigma_{11}, \Sigma_{22\cdot 1})$ is a one-to-one onto mapping of $\mathcal{L}_{p,k} \times \mathcal{S}_p^+$ to $\mathcal{L}_{r,(q+k)} \times \mathcal{L}_{q,k} \times \mathcal{S}_q^+ \times \mathcal{S}_r^+$. We now turn to the likelihood ratio test of $\tilde{H}_0$ versus $\tilde{H}_1$ based on the conditional model

$$\mathcal{L}(Y|X) = N(WC, I_n \otimes \Sigma_{22\cdot 1})$$

where X is treated as fixed. With X fixed, testing $\tilde{H}_0$ versus $\tilde{H}_1$ is a special case of the MANOVA testing problem and the results in Chapter 9 are directly applicable. To express $\tilde{H}_0$ in the MANOVA testing problem form, let K be the $q \times (q + k)$ matrix $K = (I_q \; 0)$, so the null hypothesis $\tilde{H}_0$ is

$$\tilde{H}_0 : KC = 0.$$

Recall that

$$\hat{C} \equiv (W'W)^{-1}W'Y$$

is the maximum likelihood estimator of C under $\tilde{H}_1$. Let $P_W = W(W'W)^{-1}W'$ denote the orthogonal projection onto the column space of W, let $Q_W = I_n - P_W$, and define $V \in \mathcal{S}_r^+$ by

$$V \equiv Y'Q_WY = (Y - W\hat{C})'(Y - W\hat{C}).$$

As shown in Section 9.1, based on the model

$$\mathcal{L}(Y|X) = N(WC, I_n \otimes \Sigma_{22\cdot 1}),$$

the likelihood ratio test of $\tilde{H}_0 : KC = 0$ versus $\tilde{H}_1 : KC \neq 0$ rejects H_0 for small values of

$$\Lambda_1(Y) \equiv \frac{|V|}{|V + (K\hat{C})'\left(K(W'W)^{-1}K'\right)^{-1}(K\hat{C})|}.$$

For each fixed X, Proposition 9.1 shows that under H_0, the distribution of $\Lambda_1(Y)$ is $U(n - q - k, q, r)$, which is the distribution (unconditional) of $\Lambda(Z)$ under H_0. In fact, much more is true.

Proposition 10.17. In the notation above:

(i) $V = S_{22\cdot 1}$.
(ii) $(K\hat{C})'(K(W'W)^{-1}K')^{-1}(K\hat{C}) = S_{21}S_{11}^{-1}S_{12}$.
(iii) $\Lambda_1(Y) = \Lambda(Z)$.

Further, under H_0, the conditional (given X) and unconditional distribution of $\Lambda_1(Y)$ and $\Lambda(Z)$ are the same.

Proof. To establish (i), first write S as

$$S = (Z - T\hat{B})'(Z - T\hat{B}) = Z'(I - P_T)Z$$

where $P_T = T(T'T)^{-1}T'$ is the orthogonal projection onto the column space of T. Setting $Q_T = I - P_T$ and writing $Z = (XY)$, we have

$$S = Z'Q_TZ = \begin{pmatrix} X' \\ Y' \end{pmatrix} Q_T(XY) = \begin{pmatrix} X'Q_TX & X'Q_TY \\ Y'Q_TX & Y'Q_TY \end{pmatrix} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}.$$

This yields the identity

$$S_{22\cdot 1} = Y'Q_TY - Y'Q_TX(X'Q_TX)^{-1}X'Q_TY = Y'(I - P_T)Y - Y'P_0Y$$

where $P_0 = Q_TX(X'Q_TX)^{-1}X'Q_T$ is the orthogonal projection onto the column space of Q_TX. However, a bit of reflection reveals that $P_0 = P_W - P_T$ so

$$S_{22\cdot 1} = Y'(I - P_T)Y - Y'(P_W - P_T)Y = Y'(I - P_W)Y = Y'Q_WY = V.$$

This establishes assertion (i). For (ii), we have

$$S_{21}S_{11}^{-1}S_{12} = Y'P_0Y$$

and

$$\begin{aligned}(K\hat{C})'\big(K(W'W)^{-1}K'\big)^{-1}K\hat{C} \\ &= Y'W(W'W)^{-1}K'\big(K(W'W)^{-1}W'W(W'W)^{-1}K'\big)^{-1} \\ &\quad \times K(W'W)^{-1}W'Y \\ &= Y'U(U'U)^{-1}U'Y \equiv Y'P_UY\end{aligned}$$

where $U \equiv W(W'W)^{-1}K'$ and P_U is the orthogonal projection onto the column space of U. Thus it must be shown that $P_U = P_0$ or, equivalently, that the column space of U is the same as the column space of Q_TX. Since $W = (XT)$, the relationship

$$W'U = W'W(W'W)^{-1}K' = K' = \begin{pmatrix} I_q \\ 0 \end{pmatrix}$$

proves that the q columns of U are orthogonal to the k columns of T. Thus the columns of U span a q-dimensional subspace contained in the column space of W and orthogonal to the column space of T. But there is only one subspace with these properties. Since the column space of $Q_T X$ also has these properties, it follows that $P_U = P_0$ so (ii) holds. Relationship (iii) is a consequence of (i), (ii), and

$$\Lambda(Z) = \frac{|S|}{|S_{11}||S_{22}|} = \frac{|S_{22\cdot 1}|}{|S_{22\cdot 1} + S_{21}S_{11}^{-1}S_{12}|}.$$

The validity of the final assertion concerning the distribution of $\Lambda_1(Y)$ and $\Lambda(Z)$ was established earlier. □

The results of Proposition 10.17 establish the connection between testing for independence and the MANOVA testing problem. Further, under H_0, the conditional distribution of $\Lambda_1(Y)$ is $U(n - q - k, q, r)$ for each value of X, so the marginal distribution of X is irrelevant. In other words, as long as the conditional model for Y given X is valid, we can test $\tilde{H}_0$ using the likelihood ratio test and under H_0, the distribution of the test statistic does not depend on the value of X. Of course, this implies that the conditional (given X) distribution of $\Lambda(Z)$ is the same as the unconditional distribution of $\Lambda(Z)$ under H_0. However, under H_1, the conditional and unconditional distributions of $\Lambda(Z)$ are not the same.

PROBLEMS

1. Given positive integers t, q, and r with $t \leqslant q, r$, consider random vectors $U \in R^t$, $V \in R^q$, and $W \in R^r$ where $\text{Cov}(U) = I_t$ and U, V, and W are uncorrelated. For $A : q \times t$ and $B : r \times t$, construct $X = AU + V$ and $Y = BU + W$.

   (i) With $\Lambda_{11} = \text{Cov}(V)$ and $\Lambda_{22} = \text{Cov}(W)$, show that

   $$\text{Cov}(X) = AA' + \Lambda_{11}$$

   $$\text{Cov}(Y) = BB' + \Lambda_{22}$$

   and the cross covariance between X and Y is AB'. Conclude that the number of nonzero population canonical correlations between X and Y is at most t.

   (ii) Conversely, given $\tilde{X} \in R^q$ and $\tilde{Y} \in R^r$ with t nonzero population canonical correlations, construct U, V, W, A, and B as above so that $X = AU + V$ and $Y = BU + W$ have the same joint covariance as $\tilde{X}$ and $\tilde{Y}$.

2. Consider $X \in R^q$ and $Y \in R^r$ and assume that $\text{Cov}(X) = \Sigma_{11}$ and $\text{Cov}(Y) = \Sigma_{22}$ exist. Let Σ_{12} be the cross covariance of X with Y. Recall that $\mathcal{P}_n$ denotes the group of $n \times n$ permutation matrices.
   (i) If $g\Sigma_{12}h = \Sigma_{12}$ for all $g \in \mathcal{P}_q$ and $h \in \mathcal{P}_r$, show that $\Sigma_{12} = \delta e_1 e_2'$ for some $\delta \in R^1$ where e_1 (e_2) is the vector of ones in R^q (R^r).
   (ii) Under the assumptions in (i), show that there is at most one nonzero canonical correlation and it is $|\delta|(e_1'\Sigma_{11}^{-1}e_1)^{1/2}(e_2'\Sigma_{22}^{-1}e_2)^{1/2}$. What is a set of canonical coordinates?

3. Consider $X \in R^p$ with $\text{Cov}(X) = \Sigma > 0$ (for simplicity, assume $\mathcal{E}X = 0$). This problem has to do with the approximation of X by a lower dimensional random vector—say $Y = BX$ where B is a $t \times p$ matrix of rank t.
   (i) In the notation of Proposition 10.4, suppose $A_0 : p \times p$ is used to define the inner product $[\cdot, \cdot]$ on R^n and prediction error is measured by $\mathcal{E}\|X - CY\|^2$ where $\|\cdot\|$ is defined by $[\cdot, \cdot]$ and C is $p \times t$. Show that the minimum prediction error (B fixed) is

$$\delta(B) = \text{tr}\, A_0\big(\Sigma - \Sigma B'(B\Sigma B')^{-1}B\Sigma\big)$$

   and the minimum is achieved for $C = \hat{C} = \Sigma B(B\Sigma B')^{-1}$.
   (ii) Let $A = \Sigma^{1/2}A_0\Sigma^{1/2}$ and write A in spectral form as $A = \Sigma_1^p \lambda_i a_i a_i'$ where $\lambda_1 \geqslant \cdots \geqslant \lambda_p > 0$ and $a_1, \dots, a_p$ is an orthonormal basis for R^p. Show that $\delta(B) = \text{tr}\, A(I - Q(B))$ where $Q(B) = \Sigma^{1/2}B'(B\Sigma B')^{-1}B\Sigma^{1/2}$ is a rank t orthogonal projection. Using this, show that $\delta(B)$ is minimized by choosing $Q = \hat{Q} = \Sigma_1^t a_i a_i'$, and the minimum is $\Sigma_{t+1}^p \lambda_i$. What is a corresponding $\hat{B}$ and $\hat{X} = \hat{C}\hat{B}X$ that gives the minimum? Show that $\hat{X} = \hat{C}\hat{B}X = \Sigma^{1/2}\hat{Q}\Sigma^{-1/2}X$.
   (iii) In the special case that $A_0 = I_p$, show that

$$\hat{X} = \sum_{i=1}^{t} (a_i'X)a_i$$

   where $a_1, \dots, a_p$ are the eigenvectors of Σ and $\Sigma a_i = \lambda_i a_i$ with $\lambda_1 \geqslant \cdots \geqslant \lambda_p$. (The random variables $a_i'X$ are often called the *principal components* of X, $i = 1, \dots, p$. It is easily verified that $\text{cov}(a_i'X, a_j'X) = \delta_{ij}\lambda_i$.)

4. In R^p, consider a translated subspace $M + a_0$ where $a_0 \in R^p$—such a set in R^p is called a *flat* and the dimension of the flat is the dimension of M.

(i) Given any flat $M + a_0$, show that $M + a_0 = M + b_0$ for some *unique* $b_0 \in M^{\perp}$.

Consider a flat $M + a_0$, and define the orthogonal projection onto $M + a_0$ by $x \to P(x - a_0) + a_0$ where P is the orthogonal projection onto M. Given n points $x_1, \ldots, x_n$ in R^p, consider the problem of finding the "closest" k-dimensional flat $M + a_0$ to the n points. As a measure of distance of the n points from $M + a_0$, we use $\Delta(M, a_0) = \Sigma_1^n \|x_i - \hat{x}_i\|^2$ where $\|\cdot\|$ is the usual Euclidean norm and $\hat{x}_i = P(x_i - a_0) + a_0$ is the projection of x_i onto $M + a_0$. The problem is to find M and a_0 to minimize $\Delta(M, a_0)$ over all k-dimensional subspaces M and all a_0.

(ii) First, regard a_0 as fixed, and set $S(b) = \Sigma_1^n(x_i - b)(x_i - b)'$ for any $b \in R^p$. With $Q = I - P$, show that $\Delta(M, a_0) = \operatorname{tr} S(a_0)Q = \operatorname{tr} S(\bar{x})Q + n(a_0 - \bar{x})'Q(a_0 - \bar{x})$ where $\bar{x} = n^{-1}\Sigma_1^n x_i$.

(iii) Write $S(\bar{x}) = \Sigma_1^p \lambda_i v_i v_i'$ in spectral form where $\lambda_1 \geqslant \cdots \geqslant \lambda_p \geqslant 0$ and $v_1, \ldots, v_p$ is an orthonormal basis for R^p. Using (ii), show that $\Delta(M, a_0) \geqslant \Sigma_{k+1}^p \lambda_i$ with equality for $z_0 = \bar{x}$ and for $M = \operatorname{span}\{v_1, \ldots, v_k\}$.

5. Consider a sample covariance matrix

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$$

with $S_{ii} > 0$ for $i = 1, 2$. With $t = \min\{\dim S_{ii}, i = 1, 2\}$, show that the t sample canonical correlations are the t largest solutions (in λ) to the equation $|S_{12}S_{22}^{-1}S_{21} - \lambda^2 S_{11}| = 0$, $\lambda \in [0, \infty)$.

6. (The Eckhart–Young Theorem, 1936.) Given a matrix $A : n \times p$ (say $n \geqslant p$), let $k \leqslant p$. The problem is to find a matrix $B : n \times p$ of rank no greater than k that is "closest" to A in the usual trace inner product on $\mathcal{L}_{p,n}$. Let $\mathfrak{B}_k$ be all the $n \times p$ matrices of rank no larger than k, so the problem is to find

$$\inf_{B \in \mathfrak{B}_k} \|A - B\|^2$$

where $\|M\|^2 = \operatorname{tr} MM'$ for $M \in \mathcal{L}_{p,n}$.

(i) Show that every $B \in \mathfrak{B}_k$ can be written ψC where ψ is $n \times k$, $\psi'\psi = I_k$, and C is $k \times p$. Conversely, $\psi C \in \mathfrak{B}_k$, for each such ψ and C.

(ii) Using the results of Example 4.4, show that, for A and ψ fixed,

$$\inf_{C \in \mathcal{L}_{p,k}} \|A - \psi C\|^2 = \|A - \psi\psi' A\|^2.$$

(iii) With $Q = I - \psi\psi'$, Q is a rank $n - k$ orthogonal projection. Show that, for each $B \in \mathcal{B}_k$,

$$\|A - B\|^2 \geqslant \inf_Q \|AQ\|^2 = \inf_Q \operatorname{tr} QAA' = \Sigma_{k+1}^p \lambda_i^2$$

where $\lambda_1 \geqslant \cdots \geqslant \lambda_p$ are the singular values of A. Here Q ranges over all rank $n - k$ orthogonal projections.

(iv) Write $A = \Sigma_1^p \lambda_i u_i v_i'$ as the singular value decomposition for A. Show that $\hat{B} = \Sigma_1^k \lambda_i u_i v_i'$ achieves the infimum of part (iii).

7. In the case of a random sample from a bivariate normal distribution $N(\mu, \Sigma)$, use Proposition 10.8 and Karlin's Lemma in the Appendix to show that the density of $W = \sqrt{n-2}\, r(1 - r^2)^{-1/2}$ (r is the sample correlation coefficient) has a monotone likelihood ratio in $\theta = \rho(1 - \rho^2)^{-1/2}$. Conclude that the density of r has a monotone likelihood ratio in ρ.

8. Let $f_{p,q}$ denote the density function on $(0, \infty)$ of an unnormalized $F_{p,q}$ random variable. Under the assumptions of Proposition 10.10, show that the distribution of $W = r_1^2(1 - r_1^2)^{-1}$ has a density given by

$$f(w|\rho) = \sum_{k=1}^{\infty} f_{r+2k,\, n-r-1}(w) h(k|\rho)$$

where

$$h(k|\rho) = \frac{(1 - \rho^2)^{(n-1)/2}\Gamma((n-1)/2 + k)}{k!\Gamma((n-1)/2)}(\rho^2)^k,$$

$$k = 0, 1, \ldots .$$

Note that $h(\cdot|\rho)$ is the probability mass function of a negative binomial distribution, so $f(w|\rho)$ is a mixture of F distributions. Show that $f(\cdot|\rho)$ has a monotone likelihood ratio.

9. (A generalization of Proposition 10.12.) Consider the space R^n and an integer k with $1 \leqslant k < n$. Fix an orthogonal projection P of rank k, and for $s \leqslant n - k$, let $\mathcal{P}_s$ be the set of all $n \times n$ orthogonal projections R of rank s that satisfy $RP = 0$. Also, consider the group $\mathcal{O}(P) = \{\Gamma | \Gamma \in \mathcal{O}_n, \Gamma P = P\Gamma\}$.

(i) Show that the group $\mathcal{O}(P)$ acts transitively on $\mathcal{P}_s$ under the action $R \to \Gamma R \Gamma'$.

(ii) Argue that there is a unique $\mathcal{O}(P)$ invariant probability distribution on $\mathcal{P}_s$.

(iii) Let Δ have a uniform distribution on $\mathcal{O}(P)$ and fix $R_0 \in \mathcal{P}_s$. Show that $\Delta R_0 \Delta'$ has the unique $\mathcal{O}(P)$ invariant distribution on $\mathcal{P}_s$.

10. Suppose $Z \in \mathcal{L}_{p,n}$ has an $\mathcal{O}_n$-left invariant distribution and has rank p with probability one. Let Q be a rank $n-k$ orthogonal projection with $p+k \leqslant n$ and form $W = QZ$.

(i) Show that W has rank p with probability one.

(ii) Show that $R = W(W'W)^{-1}W$ has the uniform distribution on $\mathcal{P}_p$ (in the notation of Problem 9 above with $P = I - Q$ and $s = p$).

11. After the proof of Proposition 10.13, it was argued that, when $q \leqslant r$, to find the distribution of $r_1 \geqslant \cdots \geqslant r_q$, it suffices to find the distribution of the eigenvalues of the matrix $B = (T_1 + T_2)^{-1/2}T_1(T_1 + T_2)^{-1/2}$ where T_1 and T_2 are independent with $\mathcal{L}(T_1) = W(I_q, q, n-r-1)$ and $\mathcal{L}(T_2) = W(I_q, q, r)$. It is assumed that $q \leqslant n-r-1$. Let $f(\cdot|m)$ denote the density function of the $W(I_q, q, m)$ distribution $(m \geqslant q)$ with respect to Lebesgue measure dS on $\mathcal{S}_q$. Thus $f(S|m) = \omega(m,q)|S|^{(m-q-1)/2}\exp[-\frac{1}{2}\operatorname{tr} S]I(S)$ where

$$I(S) = \begin{cases} 1 & \text{if } S > 0 \\ 0 & \text{otherwise} \end{cases}.$$

(i) With $W_1 = T_1$ and $W_2 = T_1 + T_2$, show that the joint density of W_1 and W_2 with respect to $dW_1\,dW_2$ is $f(W_1|n-r-1)f(W_2 - W_1|r)$.

(ii) On the set where $W_1 > 0$ and $W_2 > 0$, define $B = W_2^{-1/2}W_1W_2^{-1/2}$ and $W_2 = V$. Using Proposition 5.11, show that the Jacobian of this transformation is $|\det V|^{(q+1)/2}$. Show that the joint density of B and V on the set where $B > 0$ and $V > 0$ is given by

$$f\left(V^{1/2}BV^{1/2}|n-r-1\right)f\left(V^{1/2}(I-B)V^{1/2}|r\right)|\det V|^{(q+1)/2}.$$

(iii) Now, integrate out V to show that the density of B on the set $0 < B < I_q$ is

$$p(B) = \frac{\omega(n-r-1,q)\,\omega(r,q)}{\omega(n-1,q)}$$
$$\times |B|^{(n-r-q-2)/2}|I_q - B|^{(r-q-1)/2}.$$

12. Suppose the random orthogonal transformation Γ has a uniform distribution on $\mathcal{O}_n$. Let Δ be the upper left-hand $k \times p$ block of Γ and assume $p \leqslant k$. Under the additional assumption that $p \leqslant n - k$, the following argument shows that Δ has a density with respect to Lebesgue measure on $\mathcal{L}_{p,k}$.

(i) Let $\psi : n \times p$ consist of the first p columns of Γ so $\Delta : k \times p$ has rows that are the first k rows of ψ. Show that ψ has a uniform distribution on $\mathcal{F}_{p,n}$. Conclude that ψ has the same distribution as $Z(Z'Z)^{-1/2}$ where $Z : n \times p$ is $N(0, I_n \otimes I_p)$.

(ii) Now partition Z as $Z = \begin{pmatrix} X \\ Y \end{pmatrix}$ where X is $k \times p$ and Y is $(n-k) \times p$. Show that $Z'Z = X'X + Y'Y$ and that Δ has the same distribution as $X(X'X + Y'Y)^{-1/2}$.

(iii) Using (ii) and Problem 11, show that $B = \Delta'\Delta$ has the density

$$p(B) = \frac{\omega(k, p)\omega(n-k, p)}{\omega(n, p)} |B|^{(k-p-1)/2} |I_p - B|^{(n-k-p-1)/2}$$

with respect to Lebesgue measure on the set $0 < B < I_p$.

(iv) Consider a random matrix $L : k \times p$ with a density with respect to Lebesgue measure given by

$$h(L) = c|I_p - L'L|^{(n-k-p-1)/2} \phi(L'L)$$

where for $B \in \mathcal{S}_p$,

$$\phi(B) = \begin{cases} 1 & \text{if } 0 < B < I_p \\ 0 & \text{otherwise} \end{cases}$$

and

$$c = \frac{\omega(n-k, p)}{(\sqrt{2\pi})^{kp} \omega(n, p)}.$$

Show that $B = L'L$ has the density $p(B)$ given in part (iii) (use Proposition 7.6).

(v) Now, to conclude that Δ has h as its density, first prove the following proposition: Suppose $\mathcal{X}$ is acted on measurably by a compact group G and $\tau : \mathcal{X} \to \mathcal{Y}$ is a maximal invariant. If P_1 and P_2 are both G-invariant measures on $\mathcal{X}$ such that $P_1(\tau^{-1}(C)) = P_2(\tau^{-1}(C))$ for all measurable $C \subseteq \mathcal{Y}$, then $P_1 = P_2$.

(vi) Now, apply the proposition above with $\mathcal{X} = \mathcal{L}_{p,k}$, $G = \mathcal{O}_k$, $\tau(x) = x'x$, P_1 the distribution of Δ, and P_2 the distribution of L as given in (iv). This shows that Δ has density h.

13. Consider a random matrix $Z: n \times p$ with a density given by $f(Z|B, \Sigma) = |\Sigma|^{-n/2} h(\operatorname{tr}(Z - TB)\Sigma^{-1}(Z - TB)')$ where $T: n \times k$ of rank k is known, $B: k \times p$ is a matrix of unknown parameters, and $\Sigma: p \times p$ is positive definite and unknown. Assume that $n \geqslant p + k$, that

$$\sup_{C \in \mathcal{S}_p^+} |C|^{n/2} h(\operatorname{tr}(C)) < +\infty,$$

and that h is a nonincreasing function defined on $[0, \infty)$. Partition Z into $X: n \times q$ and $Y: n \times r$, $q + r = p$, so $Z = (XY)$. Also, partition Σ into Σ_{ij}, $i, j = 1,2$, where Σ_{11} is $q \times q$, Σ_{22} is $r \times r$, and Σ_{12} is $q \times r$.

(i) Show that the maximum likelihood estimator of B is $\hat{B} = (T'T)^{-1}TZ$ and $f(Z|\hat{B}, \Sigma) = |\Sigma|^{-n/2}h(\operatorname{tr} S\Sigma^{-1})$ where $S = Z'QZ$ with $Q = I - P$ and $P = T(T'T)^{-1}T'$.

(ii) Derive the likelihood ratio test of $H_0: \Sigma_{12} = 0$ versus $H_1: \Sigma_{12} \neq 0$. Show that the test rejects for small values of

$$\Lambda(Z) = \frac{|S|}{|S_{11}||S_{22}|}.$$

(iii) For $U: n \times q$ and $V: n \times r$, establish the identity $\operatorname{tr}(UV)\Sigma^{-1}(UV)' = \operatorname{tr}(V - U\Sigma_{11}^{-1}\Sigma_{12})\Sigma_{22\cdot1}^{-1}(V - U\Sigma_{11}^{-1}\Sigma_{12}) + \operatorname{tr} U\Sigma_{11}^{-1}U'$. Use this identity to derive the conditional distribution of Y given X in the above model. Using the notation of Section 10.5, show that the conditional density of Y given X is

$$f_1(Y|C, B_1, \Sigma_{11}, \Sigma_{22\cdot1}, X)$$

$$= |\Sigma_{22\cdot1}|^{-n/2}h\big(\operatorname{tr}(Y - WC)\Sigma_{22\cdot1}^{-1}(Y - WC)' + \eta\big)\phi(\eta)$$

where $\eta = \operatorname{tr}(X - TB_1)\Sigma_{11}^{-1}(X - TB_1)$ and $(\phi(\eta))^{-1} = \int_{\mathcal{L}_{r,n}} h(\operatorname{tr} uu' + \eta)\, du$.

(iv) The null hypothesis is now that $C_1 = 0$. Show that, for each fixed η, the likelihood ratio test (with C and $\Sigma_{22\cdot1}$ as parameters) based on the conditional density rejects for large values of $\Lambda(Z)$. Verify (i), (ii), and (iii) of Proposition 10.17.

(v) Now, assume that

$$\sup_{\eta > 0} \sup_{C \in \mathbb{S}_r^+} |C|^{n/2} h(\operatorname{tr} C + \eta)\phi(\eta) = k_2 < +\infty.$$

Show that the likelihood ratio test for $C_1 = 0$ (with C, $\Sigma_{22\cdot 1}$, B_1, and Σ_{11} as parameters) rejects for large values of $\Lambda(Z)$.

(vi) Show that, under H_0, the sample canonical correlations based on S_{11}, S_{12}, S_{22} (here $S = Z'QZ$) have the same distribution as when Z is $N(TB, I_n \otimes \Sigma)$. Conclude that under H_0, $\Lambda(Z)$ has the same distribution as when Z is $N(TB, I_n \otimes \Sigma)$.

NOTES AND REFERENCES

1. Canonical correlation analysis was first proposed in Hotelling (1935, 1936). There are as many approaches to canonical correlation analysis as there are books covering the subject. For a sample of these, see Anderson (1958), Dempster (1969), Kshirsagar (1972), Rao (1973), Mardia, Kent, and Bibby (1979), Srivastava and Khatri (1979), and Muirhead (1982).

2. See Eaton and Kariya (1981) for some material related to Proposition 10.13.