6-1. Canonical Correlation Analysis

- Canonical Correlatin analysis focuses on the correlation between a *linear combination* of the variable in one set and a *linear combination* of the variables in another set.

- Canonical variables, Canonical correlation.

- Examples: Arithmetic speed and arithmetic power to reading speed and reading power, Governmental policy variables with economic goal variables, College "performance" variables with precollege "achievement" variables.

Canonical Variates and Canonical Correlation

Let $\boldsymbol{X}^{(1)}$ be a $p \times 1$ random vector, $\boldsymbol{X}^{(2)}$ be a $q \times 1$ random vector with $p \leq q$, and

$$\mathrm{E}(\boldsymbol{X}^{(1)}) = \boldsymbol{\mu}^{(1)}; \quad \mathrm{Cov}(\boldsymbol{X}^{(1)}) = \Sigma_{11};$$
$$\mathrm{E}(\boldsymbol{X}^{(2)}) = \boldsymbol{\mu}^{(2)}; \quad \mathrm{Cov}(\boldsymbol{X}^{(2)}) = \Sigma_{22};$$
$$\mathrm{Cov}(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}) = \Sigma_{12} = \Sigma_{21}'.$$

Set

$$U = \mathbf{a}'\boldsymbol{X}^{(1)}, \quad V = \mathbf{b}'\boldsymbol{X}^{(2)},$$

Then we shall seek coefficient vectors **a** and **b** such that

$$\mathrm{Corr}(U, V) = \frac{\mathbf{a}'\Sigma_{12}\mathbf{b}}{\sqrt{\mathbf{a}'\Sigma_{11}\mathbf{a}}\sqrt{\mathbf{b}'\Sigma_{22}\mathbf{b}}}$$

is as large as possible.

- *The first pair of canonical variables, or first canonical variate pair* the pair linear combination U_1 and V_1 having unit variances, which maximize the correlation Corr(U, V);

- *The second pair of canonical variables, or second canonical variate pair* the pair of linear combinations U_2 and V_2 having unit variances , which maximize the correlation Corr(U, V) among all choices that are uncorrelated with the first pair of canonical variables.

- *The kth pair of canonical variables, or kth canonical variate pair* the pair of linear combinations U_k, V_k having unit variances, which maximize the correlation Corr(U, V) among all choices uncorrelated with the previous $k - 1$ canonical variable pares

The correlation between the kth pair of canonical variable s is called the *kth canonical correlation*

Result 6-1.1: Suppose $p \leq q$ and let the p-dimensional random vectors $\boldsymbol{X}^{(}1)$ and q dimensional $\boldsymbol{X}^{(2)}$ have $\text{Cov}(\boldsymbol{X}^{(1)}) = \Sigma_{11}$, $\text{Cov}(\boldsymbol{X}^{(1)}) = \Sigma_{22}$, and $\text{Cov}(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}) = \Sigma_{12}$, where Σ has full rank. For coefficients $p \times 1$ vector $\mathbf{a}$ and $q \times 1$ vector $\mathbf{b}$, form the linear combination $U = \mathbf{a}'\boldsymbol{X}^{(1)}$ and $V = \mathbf{b}'\boldsymbol{X}^{(2)}$. Then

$$\max_{\mathbf{a},\mathbf{b}} \text{Corr}(U, V) = \rho_1^*$$

attained by the linear combinations (first canonical variate pair)

$$U_1 = \mathbf{e}_1'\Sigma_{11}^{-1/2}\boldsymbol{X}^{(1)} \quad \text{and} \quad V_1 = \mathbf{f}_1'\Sigma_{22}^{-1/2}\boldsymbol{X}^{(2)},$$

The kth pair of canonical variates, $k = 2, 3, \ldots, p$

$$U_k = \mathbf{e}_k'\Sigma_{11}^{-1/2}\boldsymbol{X}^{(1)} \quad \text{and} \quad V_k = \mathbf{f}_k'\Sigma_{22}^{-1/2}\boldsymbol{X}^{(2)},$$

maximize

$$\max_{\mathbf{a},\mathbf{b}} \text{Corr}(U, V) = \rho_k^*$$

among those linear combinations uncorrelated with the preceding $1, 2, \ldots, k-1$ canonical variables.

- Here $\rho_1^{*2} \geq \rho_2^{*2} \cdots \geq \rho_p^{*2}$ are the eigenvalues of $\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}$, and $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p$ are associated $p \times 1$ eigenvectors.

- The quantities $\rho_1^{*2} \geq \rho_2^{*2} \cdots \geq \rho_p^{*2}$ are also the p largest eigenvalues of the matrix $\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}$ with corresponding $q \times 1$ eigenvectors $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_p$.

- Each $\mathbf{f}_i$ is proportional to $\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1/2}\mathbf{e}_i$.

- The canonical variates have the properties

$$\mathsf{Var}(U_k) = \mathsf{Var}(V_k) = 1, k = 1, \ldots, p,$$

$$\mathsf{Corr}(U_k, U_\ell) = \mathsf{Corr}(V_k, V_\ell) = \mathsf{Corr}(U_k, V_\ell) = 0$$

  for $k, \ell = 1, 2, \ldots, p$ and $k \neq \ell$.

Example 6-1.1 Suppose $\mathbf{Z}^{(1)} = [Z_1^{(1)}, Z_2^{(1)}]'$ are standardize variables and $\mathbf{Z}^{(2)} = [Z_1^{(2)}, Z_2^{(2)}]'$ are also standardized variables. Let $\mathbf{Z} = [\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}]'$ and

$$\text{Cov}(\mathbf{Z}) = \begin{bmatrix} 1.0 & 4 & .5 & .6 \\ .4 & 1.0 & .3 & .4 \\ .5 & .3 & 1.0 & .2 \\ .6 & .4 & .2 & 1.0 \end{bmatrix}$$

Calculate canonical variates and canonical correlations for standardized variables $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$.

Example 6-1.2 Compute the computing correlations between the first pair canonical variates and their component variables for the situation considered in Example 6-1.1.

Example 6-1.3 Consider the covariance matrix

$$\text{Cov}\begin{pmatrix} X_1^{(1)} \\ X_2^{(1)} \\ X_1^{(2)} \\ X_2^{(2)} \end{pmatrix} = \begin{bmatrix} 100 & 0 & 0 & 0 \\ 0 & 1 & 0.95 & 0 \\ 0 & 0.95 & 1 & 0 \\ 0 & 0 & 0 & 100 \end{bmatrix}$$

Calculate the canonical correlation between $[X_1^{(1)}, X_2^{(1)}]'$ and $[X_1^{(2)}, X_2^{(2)}]'$

The Sample Canonical Variates and Sample Canonical Correlation

Results 6-1.2. Let $\hat{\rho}_1^{*2} \geq \hat{\rho}_2^{*2} \geq \ldots \geq \hat{\rho}_p^{*2}$ be the p ordered eigenvalues of $S_{11}^{-1/2} S_{12} S_{22}^{-1} S_{21} S_{11}^{-1/2}$ with corresponding eigenvectors $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \ldots, \hat{\mathbf{e}}_p$, where $p \leq q$. Let $\hat{\mathbf{f}}_1, \ldots, \hat{\mathbf{f}}_p$ be the eigenvectors of $S_{22}^{-1/2} S_{21} S_{11}^{-1} S_{12} S_{22}^{-1/2}$. Then the kth sample canonical variate pair is

$$\hat{U}_k = \hat{\mathbf{e}}_k S_{11}^{-1/2} \mathbf{x}^{(1)}, \quad \hat{V}_k = \hat{\mathbf{f}}_k S_{22}^{-1/2} \mathbf{x}^{(2)}$$

where $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ are the values of the variables $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ for a particular experimental unit. Also for the kth pair, $k = 1, \ldots, p$

$$r_{\hat{U}_k, \hat{V}_k} = \hat{\rho}_k^*.$$

The quantities $\hat{\rho}_1^*, \ldots, \hat{\rho}_p^*$ are the sample canonical correlations.

Large Sample Inference

Results 6-1.3 Let

$$\boldsymbol{X}_j = \begin{bmatrix} \boldsymbol{X}_j^{(1)} \\ \boldsymbol{X}_j^{(2)} \end{bmatrix}, j = 1, 2, \ldots, n$$

bee a random sample from an $N_{p+q}(\boldsymbol{\mu}, \Sigma)$ population with

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

Then the likelihood ratio test $H_0 : \Sigma_{12} = 0$ vs $H_1 : \Sigma_{12} \neq 0$ reject H_0 for large value of

$$-2 \ln \Gamma = n \ln \left(\frac{|S_{11}||S_{22}|}{|S|} \right) = -n \ln \prod_{i=1}^{p} (1 - \hat{\rho}_i^{*2})$$

where

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

is the unbiased estimator of Σ. For large n the test statistic $-2 \ln \Gamma$ is approximately distributed as a chi-square random variable with pq degree freedom.

and we can define

$$\mathbf{R}_{\hat{\mathbf{U}},\mathbf{x}^{(1)}} = \text{matrix of sample correlations of } \hat{\mathbf{U}} \text{ with } \mathbf{x}^{(1)}$$
$$\mathbf{R}_{\hat{\mathbf{V}},\mathbf{x}^{(2)}} = \text{matrix of sample correlations of } \hat{\mathbf{V}} \text{ with } \mathbf{x}^{(2)}$$
$$\mathbf{R}_{\hat{\mathbf{U}},\mathbf{x}^{(2)}} = \text{matrix of sample correlations of } \hat{\mathbf{U}} \text{ with } \mathbf{x}^{(2)}$$
$$\mathbf{R}_{\hat{\mathbf{V}},\mathbf{x}^{(1)}} = \text{matrix of sample correlations of } \hat{\mathbf{V}} \text{ with } \mathbf{x}^{(1)}$$

Corresponding to (10-19), we have

$$\begin{aligned}\mathbf{R}_{\hat{\mathbf{U}},\mathbf{x}^{(1)}} &= \hat{\mathbf{A}}\mathbf{S}_{11}\mathbf{D}_{11}^{-1/2}\\ \mathbf{R}_{\hat{\mathbf{V}},\mathbf{x}^{(2)}} &= \hat{\mathbf{B}}\mathbf{S}_{22}\mathbf{D}_{22}^{-1/2}\\ \mathbf{R}_{\hat{\mathbf{U}},\mathbf{x}^{(2)}} &= \hat{\mathbf{A}}\mathbf{S}_{12}\mathbf{D}_{22}^{-1/2}\\ \mathbf{R}_{\hat{\mathbf{V}},\mathbf{x}^{(1)}} &= \hat{\mathbf{B}}\mathbf{S}_{21}\mathbf{D}_{11}^{-1/2}\end{aligned} \tag{10-29}$$

where $\mathbf{D}_{11}^{-1/2}$ is the $(p \times p)$ diagonal matrix with ith diagonal element (sample $\text{var}(x_i^{(1)}))^{-1/2}$ and $\mathbf{D}_{22}^{-1/2}$ is the $(q \times q)$ diagonal matrix with ith diagonal element (sample $\text{var}(x_i^{(2)}))^{-1/2}$.

Comment. If the observations are standardized [see (8-25)], the data matrix becomes

$$\mathbf{Z} = \left[\mathbf{Z}^{(1)} \vdots \mathbf{Z}^{(2)}\right] = \begin{bmatrix} \mathbf{z}_1^{(1)\prime} & \vdots & \mathbf{z}_1^{(2)\prime} \\ \vdots & & \vdots \\ \mathbf{z}_n^{(1)\prime} & \vdots & \mathbf{z}_n^{(2)\prime} \end{bmatrix}$$

and the sample canonical variates become

$$\underset{(p\times 1)}{\hat{\mathbf{U}}} = \hat{\mathbf{A}}_{\mathbf{z}}\mathbf{z}^{(1)} \qquad \underset{(q\times 1)}{\hat{\mathbf{V}}} = \hat{\mathbf{B}}_{\mathbf{z}}\mathbf{z}^{(2)} \tag{10-30}$$

where $\hat{\mathbf{A}}_{\mathbf{z}} = \hat{\mathbf{A}}\mathbf{D}_{11}^{1/2}$ and $\hat{\mathbf{B}}_{\mathbf{z}} = \hat{\mathbf{B}}\mathbf{D}_{22}^{1/2}$. The sample canonical correlations are unaffected by the standardization. The correlations displayed in (10–29) remain unchanged and may be calculated, for standardized observations, by substituting $\hat{\mathbf{A}}_{\mathbf{z}}$ for $\hat{\mathbf{A}}$, $\hat{\mathbf{B}}_{\mathbf{z}}$ for $\hat{\mathbf{B}}$, and $\mathbf{R}$ for $\mathbf{S}$. Note that $\mathbf{D}_{11}^{-1/2} = \underset{(p\times p)}{\mathbf{I}}$ and $\mathbf{D}_{22}^{-1/2} = \underset{(q\times q)}{\mathbf{I}}$ for standardized observations.

Example 10.4 (Canonical correlation analysis of the chicken-bone data) In Example 9.14, data consisting of bone and skull measurements of white leghorn fowl were described. From this example, the chicken-bone measurements for

$$\text{Head}\,(\mathbf{X}^{(1)}): \begin{cases} X_1^{(1)} = \text{skull length} \\ X_2^{(1)} = \text{skull breadth} \end{cases}$$

$$\text{Leg}\,(\mathbf{X}^{(2)}): \begin{cases} X_1^{(2)} = \text{femur length} \\ X_2^{(2)} = \text{tibia length} \end{cases}$$

have the sample correlation matrix

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{bmatrix} = \left[\begin{array}{cc:cc} 1.0 & .505 & .569 & .602 \\ .505 & 1.0 & .422 & .467 \\ \hdashline .569 & .422 & 1.0 & .926 \\ .602 & .467 & .926 & 1.0 \end{array}\right]$$

A canonical correlation analysis of the head and leg sets of variables using $\mathbf{R}$ produces the two canonical correlations and corresponding pairs of variables

$$\widehat{\rho_1^*} = .631 \qquad \begin{aligned} \hat{U}_1 &= .781z_1^{(1)} + .345z_2^{(1)} \\ \hat{V}_1 &= .060z_1^{(2)} + .944z_2^{(2)} \end{aligned}$$

and

$$\widehat{\rho_2^*} = .057 \qquad \begin{aligned} \hat{U}_2 &= -.856z_1^{(1)} + 1.106z_2^{(1)} \\ \hat{V}_2 &= -2.648z_1^{(2)} + 2.475z_2^{(2)} \end{aligned}$$

Here $z_i^{(1)}$, $i = 1, 2$ and $z_i^{(2)}$, $i = 1, 2$ are the standardized data values for sets 1 and 2, respectively. The preceding results were taken from the SAS statistical software output shown in Panel 10.1. In addition, the correlations of the original variables with the canonical variables are highlighted in that panel. ■

Example 10.5 (Canonical correlation analysis of job satisfaction) As part of a larger study of the effects of organizational structure on "job satisfaction," Dunham [4] investigated the extent to which measures of job satisfaction are related to job characteristics. Using a survey instrument, Dunham obtained measurements of $p = 5$ job characteristics and $q = 7$ job satisfaction variables for $n = 784$ executives from the corporate branch of a large retail merchandising corporation. Are measures of job satisfaction associated with job characteristics? The answer may have implications for job design.

PANEL 10.1 SAS ANALYSIS FOR EXAMPLE 10.4 USING PROC CANCORR.

```
title 'Canonical Correlation Analysis';
data skull (type = corr);
_type_ = 'CORR';
input _name_$ x1 x2 x3 x4;
cards;
x1   1.0      .      .      .
x2    .505   1.0     .      .
x3    .569    .422  1.0     .
x4    .602    .467   .926  1.0
;
proc cancorr data = skull vprefix = head wprefix = leg;
   var x1 x2; with x3 x4;
```

PROGRAM COMMANDS

(continues on next page)

PANEL 10.1 *(continued)*

Canonical Correlation Analysis

	Canonical Correlation	Adjusted Canonical Correlation	Approx Standard Error	Squared Canonical Correlation
1	0.631085	0.628291	0.036286	0.398268
2	0.056794		0.060108	0.003226

Raw Canonical Coefficient for the 'VAR' Variables

	HEAD1	HEAD2
X1	0.7807924389	−0.855973184
X2	0.3445068301	1.1061835145

Raw Canonical Coefficient for the 'WITH' Variables

	LEG1	LEG2
X3	0.0602508775	−2.648156338
X4	0.943948961	2.4749388913

OUTPUT

Canonical Structure

Correlations Between the 'VAR' Variables and Their Canonical Variables

	HEAD1	HEAD2	
X1	0.9548	−0.2974	(see 10-29)
X2	0.7388	0.6739	

Correlations Between the 'WITH' Variables and Their Canonical Variables

	LEG1	LEG2	
X3	0.9343	−0.3564	(see 10-29)
X4	0.9997	0.0227	

Correlations Between the 'VAR' Variables and the Canonical Variables of the 'WITH' Variables

	LEG1	LEG2	
X1	0.6025	−0.0169	(see 10-29)
X2	0.4663	0.0383	

Correlations Between the 'WITH' Variables and the Canonical Variables of the 'VAR' Variables

	HEAD1	HEAD2	
X3	0.5897	−0.0202	(see 10-29)
X4	0.6309	0.0013	

The original job characteristic variables, $\mathbf{X}^{(1)}$, and job satisfaction variables, $\mathbf{X}^{(2)}$, were respectively defined as

$$\mathbf{X}^{(1)} = \begin{bmatrix} X_1^{(1)} \\ X_2^{(1)} \\ X_3^{(1)} \\ X_4^{(1)} \\ X_5^{(1)} \end{bmatrix} = \begin{bmatrix} \text{feedback} \\ \text{task significance} \\ \text{task variety} \\ \text{task identity} \\ \text{autonomy} \end{bmatrix}$$

$$\mathbf{X}^{(2)} = \begin{bmatrix} X_1^{(2)} \\ X_2^{(2)} \\ X_3^{(2)} \\ X_4^{(2)} \\ X_5^{(2)} \\ X_6^{(2)} \\ X_7^{(2)} \end{bmatrix} = \begin{bmatrix} \text{supervisor satisfaction} \\ \text{career-future satisfaction} \\ \text{financial satisfaction} \\ \text{workload satisfaction} \\ \text{company identification} \\ \text{kind-of-work-satisfaction} \\ \text{general satisfaction} \end{bmatrix}$$

Responses for variables $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ were recorded on a scale and then standardized. The sample correlation matrix based on 784 responses is

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{bmatrix}$$

$$= \left[\begin{array}{ccccc:ccccccc} 1.0 & & & & & .33 & .32 & .20 & .19 & .30 & .37 & .21 \\ .49 & 1.0 & & & & .30 & .21 & .16 & .08 & .27 & .35 & .20 \\ .53 & .57 & 1.0 & & & .31 & .23 & .14 & .07 & .24 & .37 & .18 \\ .49 & .46 & .48 & 1.0 & & .24 & .22 & .12 & .19 & .21 & .29 & .16 \\ .51 & .53 & .57 & .57 & 1.0 & .38 & .32 & .17 & .23 & .32 & .36 & .27 \\ \hdashline .33 & .30 & .31 & .24 & .38 & 1.0 & & & & & & \\ .32 & .21 & .23 & .22 & .32 & .43 & 1.0 & & & & & \\ .20 & .16 & .14 & .12 & .17 & .27 & .33 & 1.0 & & & & \\ .19 & .08 & .07 & .19 & .23 & .24 & .26 & .25 & 1.0 & & & \\ .30 & .27 & .24 & .21 & .32 & .34 & .54 & .46 & .28 & 1.0 & & \\ .37 & .35 & .37 & .29 & .36 & .37 & .32 & .29 & .30 & .35 & 1.0 & \\ .21 & .20 & .18 & .16 & .27 & .40 & .58 & .45 & .27 & .59 & .31 & 1.0 \end{array}\right]$$

The $\min(p, q) = \min(5, 7) = 5$ sample canonical correlations and the sample canonical variate coefficient vectors (from Dunham [4]) are displayed in the following table:

Canonical Variate Coefficients and Canonical Correlations

	Standardized variables							Standardized variables						
	$z_1^{(1)}$	$z_2^{(1)}$	$z_3^{(1)}$	$z_4^{(1)}$	$z_5^{(1)}$	$\widehat{\rho_1^*}$		$z_1^{(2)}$	$z_2^{(2)}$	$z_3^{(2)}$	$z_4^{(2)}$	$z_5^{(2)}$	$z_6^{(2)}$	$z_7^{(2)}$
$\hat{\mathbf{a}}_1'$:	.42	.21	.17	−.02	.44	.55	$\hat{\mathbf{b}}_1'$:	.42	.22	−.03	.01	.29	.52	−.12
$\hat{\mathbf{a}}_2'$:	−.30	.65	.85	−.29	−.81	.23	$\hat{\mathbf{b}}_2'$:	.03	−.42	.08	−.91	.14	.59	−.02
$\hat{\mathbf{a}}_3'$:	−.86	.47	−.19	−.49	.95	.12	$\hat{\mathbf{b}}_3'$:	.58	−.76	−.41	−.07	.19	−.43	.92
$\hat{\mathbf{a}}_4'$:	.76	−.06	−.12	−1.14	−.25	.08	$\hat{\mathbf{b}}_4'$:	.23	.49	.52	−.47	.34	−.69	−.37
$\hat{\mathbf{a}}_5'$:	.27	1.01	−1.04	.16	.32	.05	$\hat{\mathbf{b}}_5'$:	−.52	−.63	.41	.21	.76	.02	.10

For example, the first sample canonical variate pair is

$$\hat{U}_1 = .42z_1^{(1)} + .21z_2^{(1)} + .17z_3^{(1)} - .02z_4^{(1)} + .44z_5^{(1)}$$

$$\hat{V}_1 = .42z_1^{(2)} + .22z_2^{(2)} - .03z_3^{(2)} + .01z_4^{(2)} + .29z_5^{(2)} + .52z_6^{(2)} - .12z_7^{(2)}$$

with sample canonical correlation $\widehat{\rho_1^*} = .55$.

According to the coefficients, $\hat{U}_1$ is primarily a feedback and autonomy variable, while $\hat{V}_1$ represents supervisor, career-future, and kind-of-work satisfaction, along with company identification.

To provide interpretations for $\hat{U}_1$ and $\hat{V}_1$, the sample correlations between $\hat{U}_1$ and its component variables and between $\hat{V}_1$ and its component variables were computed. Also, the following table shows the sample correlations between variables in one set and the first sample canonical variate of the other set. These correlations can be calculated using (10-29).

Sample Correlations Between Original Variables and Canonical Variables

	Sample canonical variates			Sample canonical variates	
$\mathbf{X}^{(1)}$ variables	$\hat{U}_1$	$\hat{V}_1$	$\mathbf{X}^{(2)}$ variables	$\hat{U}_1$	$\hat{V}_1$
1. Feedback	.83	.46	1. Supervisor satisfaction	.42	.75
2. Task significance	.74	.41	2. Career-future satisfaction	.35	.65
3. Task variety	.75	.42	3. Financial satisfaction	.21	.39
4. Task identity	.62	.34	4. Workload satisfaction	.21	.37
5. Autonomy	.85	.48	5. Company identification	.36	.65
			6. Kind-of-work satisfaction	.44	.80
			7. General satisfaction	.28	.50

All five job characteristic variables have roughly the same correlations with the first canonical variate $\hat{U}_1$. From this standpoint, $\hat{U}_1$ might be interpreted as a job characteristic "index." This differs from the preferred interpretation, based on coefficients, where the task variables are not important.

The other member of the first canonical variate pair, $\hat{V}_1$, seems to be representing, primarily, supervisor satisfaction, career-future satisfaction, company identification, and kind-of-work satisfaction. As the variables suggest, $\hat{V}_1$ might be regarded as a job satisfaction–company identification index. This agrees with the preceding interpretation based on the canonical coefficients of the $z_i^{(2)}$'s. The sample correlation between the two indices $\hat{U}_1$ and $\hat{V}_1$ is $\widehat{\rho_1^*} = .55$. There appears to be some overlap between job characteristics and job satisfaction. We explore this issue further in Example 10.7. ■

Scatter plots of the first $(\hat{U}_1, \hat{V}_1)$ pair may reveal atypical observations $\mathbf{x}_j$ requiring further study. If the canonical correlations $\widehat{\rho_2^*}, \widehat{\rho_3^*}, \ldots$ are also moderately large,

scatter plots of the pairs $(\hat{U}_2, \hat{V}_2)$, $(\hat{U}_3, \hat{V}_3), \ldots$ may also be helpful in this respect. Many analysts suggest plotting "significant" canonical variates against their component variables as an aid in subject-matter interpretation. These plots reinforce the correlation coefficients in (10-29).

If the sample size is large, it is often desirable to split the sample in half. The first half of the sample can be used to construct and evaluate the sample canonical variates and canonical correlations. The results can then be "validated" with the remaining observations. The change (if any) in the nature of the canonical analysis will provide an indication of the sampling variability and the stability of the conclusions.

10.5 Additional Sample Descriptive Measures

If the canonical variates are "good" summaries of their respective sets of variables, then the associations between variables can be described in terms of the canonical variates and their correlations. It is useful to have summary measures of the extent to which the canonical variates account for the variation in their respective sets. It is also useful, on occasion, to calculate the proportion of variance in one set of variables explained by the canonical variates of the other set.

Matrices of Errors of Approximations

Given the matrices $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ defined in (10-27), let $\hat{\mathbf{a}}^{(i)}$ and $\hat{\mathbf{b}}^{(i)}$ denote the ith columns of $\hat{\mathbf{A}}^{-1}$ and $\hat{\mathbf{B}}^{-1}$, respectively. Since $\hat{\mathbf{U}} = \hat{\mathbf{A}}\mathbf{x}^{(1)}$ and $\hat{\mathbf{V}} = \hat{\mathbf{B}}\mathbf{x}^{(2)}$ we can write

$$\underset{(p\times 1)}{\mathbf{x}^{(1)}} = \underset{(p\times p)}{\hat{\mathbf{A}}^{-1}}\ \underset{(p\times 1)}{\hat{\mathbf{U}}} \qquad \underset{(q\times 1)}{\mathbf{x}^{(2)}} = \underset{(q\times q)}{\hat{\mathbf{B}}^{-1}}\ \underset{(q\times 1)}{\hat{\mathbf{V}}} \tag{10-31}$$

Because sample $\mathrm{Cov}(\hat{\mathbf{U}}, \hat{\mathbf{V}}) = \hat{\mathbf{A}}\mathbf{S}_{12}\hat{\mathbf{B}}'$, sample $\mathrm{Cov}(\hat{\mathbf{U}}) = \hat{\mathbf{A}}\mathbf{S}_{11}\hat{\mathbf{A}}' = \underset{(p\times p)}{\mathbf{I}}$, and sample $\mathrm{Cov}(\hat{\mathbf{V}}) = \hat{\mathbf{B}}\mathbf{S}_{22}\hat{\mathbf{B}}' = \underset{(q\times q)}{\mathbf{I}}$,

$$\mathbf{S}_{12} = \hat{\mathbf{A}}^{-1}\left[\begin{array}{cccc|c} \widehat{\rho_1^*} & 0 & \cdots & 0 & \\ 0 & \widehat{\rho_2^*} & \cdots & 0 & \\ \vdots & \vdots & \ddots & \vdots & \mathbf{0} \\ 0 & 0 & \cdots & \widehat{\rho_p^*} & \end{array}\right] (\hat{\mathbf{B}}^{-1})' = \widehat{\rho_1^*}\hat{\mathbf{a}}^{(1)}\hat{\mathbf{b}}^{(1)\prime} + \widehat{\rho_2^*}\hat{\mathbf{a}}^{(2)}\hat{\mathbf{b}}^{(2)\prime} + \cdots + \widehat{\rho_p^*}\hat{\mathbf{a}}^{(p)}\hat{\mathbf{b}}^{(p)\prime} \tag{10-32}$$

$$\mathbf{S}_{11} = (\hat{\mathbf{A}}^{-1})(\hat{\mathbf{A}}^{-1})' = \hat{\mathbf{a}}^{(1)}\hat{\mathbf{a}}^{(1)\prime} + \hat{\mathbf{a}}^{(2)}\hat{\mathbf{a}}^{(2)\prime} + \cdots + \hat{\mathbf{a}}^{(p)}\hat{\mathbf{a}}^{(p)\prime}$$

$$\mathbf{S}_{22} = (\hat{\mathbf{B}}^{-1})(\hat{\mathbf{B}}^{-1})' = \hat{\mathbf{b}}^{(1)}\hat{\mathbf{b}}^{(1)\prime} + \hat{\mathbf{b}}^{(2)}\hat{\mathbf{b}}^{(2)\prime} + \cdots + \hat{\mathbf{b}}^{(q)}\hat{\mathbf{b}}^{(q)\prime}$$

Since $\mathbf{x}^{(1)} = \hat{\mathbf{A}}^{-1}\hat{\mathbf{U}}$ and $\hat{\mathbf{U}}$ has sample covariance $\mathbf{I}$, the first r columns of $\mathbf{S}_{11}$ contain the sample covariances of the first r canonical variates $\hat{U}_1, \hat{U}_2, \ldots, \hat{U}_r$ with their component variables $X_1^{(1)}, X_2^{(1)}, \ldots, X_p^{(1)}$. Similarly, the first r columns of $\mathbf{S}_{12}$ contain the sample covariances of $\hat{V}_1, \hat{V}_2, \ldots, \hat{V}_r$ with their component variables

If only the first r canonical pairs are used, so that for instance,

$$\tilde{\mathbf{x}}^{(1)} = [\hat{\mathbf{a}}^{(1)} \mid \hat{\mathbf{a}}^{(2)} \mid \cdots \mid \hat{\mathbf{a}}^{(r)}] \begin{bmatrix} \hat{U}_1 \\ \hat{U}_2 \\ \vdots \\ \hat{U}_r \end{bmatrix}$$

and (10-33)

$$\tilde{\mathbf{x}}^{(2)} = [\hat{\mathbf{b}}^{(1)} \mid \hat{\mathbf{b}}^{(2)} \mid \cdots \mid \hat{\mathbf{b}}^{(r)}] \begin{bmatrix} \hat{V}_1 \\ \hat{V}_2 \\ \vdots \\ \hat{V}_r \end{bmatrix}$$

then $\mathbf{S}_{12}$ is approximated by sample $\text{Cov}(\tilde{\mathbf{x}}^{(1)}, \tilde{\mathbf{x}}^{(2)})$.

Continuing, we see that the *matrices of errors of approximation* are

$$\begin{aligned}
\mathbf{S}_{11} - (\hat{\mathbf{a}}^{(1)}\hat{\mathbf{a}}^{(1)\prime} + \hat{\mathbf{a}}^{(2)}\hat{\mathbf{a}}^{(2)\prime} + \cdots + \hat{\mathbf{a}}^{(r)}\hat{\mathbf{a}}^{(r)\prime}) &= \hat{\mathbf{a}}^{(r+1)}\hat{\mathbf{a}}^{(r+1)\prime} + \cdots + \hat{\mathbf{a}}^{(p)}\hat{\mathbf{a}}^{(p)\prime} \\
\mathbf{S}_{22} - (\hat{\mathbf{b}}^{(1)}\hat{\mathbf{b}}^{(1)\prime} + \hat{\mathbf{b}}^{(2)}\hat{\mathbf{b}}^{(2)\prime} + \cdots + \hat{\mathbf{b}}^{(r)}\hat{\mathbf{b}}^{(r)\prime}) &= \hat{\mathbf{b}}^{(r+1)}\hat{\mathbf{b}}^{(r+1)\prime} + \cdots + \hat{\mathbf{b}}^{(q)}\hat{\mathbf{b}}^{(q)\prime} \\
\mathbf{S}_{12} - (\widehat{\rho_1^*}\hat{\mathbf{a}}^{(1)}\hat{\mathbf{b}}^{(1)\prime} + \widehat{\rho_2^*}\hat{\mathbf{a}}^{(2)}\hat{\mathbf{b}}^{(2)\prime} + \cdots + \widehat{\rho_r^*}\hat{\mathbf{a}}^{(r)}\hat{\mathbf{b}}^{(r)\prime}) & \\
= \widehat{\rho_{r+1}^*}\hat{\mathbf{a}}^{(r+1)}\hat{\mathbf{b}}^{(r+1)\prime} + \cdots + \widehat{\rho_p^*}\hat{\mathbf{a}}^{(p)}\hat{\mathbf{b}}^{(p)\prime} &
\end{aligned} \tag{10-34}$$

The approximation error matrices (10-34) may be interpreted as descriptive summaries of how well the first r sample canonical variates reproduce the sample covariance matrices. Patterns of large entries in the rows and/or columns of the approximation error matrices indicate a poor "fit" to the corresponding variable(s).

Ordinarily, the first r variates do a better job of reproducing the elements of $\mathbf{S}_{12} = \mathbf{S}_{21}'$ than the elements of $\mathbf{S}_{11}$ or $\mathbf{S}_{22}$. Mathematically, this occurs because the residual matrix in the former case is directly related to the smallest $p - r$ sample canonical correlations. These correlations are usually all close to zero. On the other hand, the residual matrices associated with the approximations to the matrices $\mathbf{S}_{11}$ and $\mathbf{S}_{22}$ depend only on the last $p - r$ and $q - r$ coefficient vectors. The elements in these vectors may be relatively large, and hence, the residual matrices can have "large" entries.

For standardized observations, $\mathbf{R}_{kl}$ replaces $\mathbf{S}_{kl}$ and $\hat{\mathbf{a}}_z^{(k)}$, $\hat{\mathbf{b}}_z^{(l)}$ replace $\hat{\mathbf{a}}^{(k)}$, $\hat{\mathbf{b}}^{(l)}$ in (10-34).

Example 10.6 (Calculating matrices of errors of approximation) In Example 10.4, we obtained the canonical correlations between the two head and the two leg variables for white leghorn fowl. Starting with the sample correlation matrix

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{bmatrix} = \left[\begin{array}{cc:cc} 1.0 & .505 & .569 & .602 \\ .505 & 1.0 & .422 & .467 \\ \hdashline .569 & .422 & 1.0 & .926 \\ .602 & .467 & .926 & 1.0 \end{array}\right]$$

we obtained the two sets of canonical correlations and variables

$$\widehat{\rho_1^*} = .631 \qquad \begin{aligned} \hat{U}_1 &= .781z_1^{(1)} + .345z_2^{(1)} \\ \hat{V}_1 &= .060z_1^{(2)} + .944z_2^{(2)} \end{aligned}$$

and

$$\widehat{\rho_2^*} = .057 \qquad \begin{aligned} \hat{U}_2 &= -.856z_1^{(1)} + 1.106z_2^{(1)} \\ \hat{V}_2 &= -2.648z_1^{(2)} + 2.475z_2^{(2)} \end{aligned}$$

where $z_i^{(1)}, i = 1, 2$ and $z_i^{(2)}, i = 1, 2$ are the standardized data values for sets 1 and 2, respectively.

We first calculate (see Panel 10.1)

$$\hat{\mathbf{A}}_z^{-1} = \begin{bmatrix} .781 & .345 \\ -.856 & 1.106 \end{bmatrix}^{-1} = \begin{bmatrix} .9548 & -.2974 \\ .7388 & .6739 \end{bmatrix}$$

$$\hat{\mathbf{B}}_z^{-1} = \begin{bmatrix} .9343 & -.3564 \\ .9997 & .0227 \end{bmatrix}$$

Consequently, the matrices of errors of approximation created by using only the first canonical pair are

$$\mathbf{R}_{12} - \text{sample Cov}(\tilde{\mathbf{z}}^{(1)}, \tilde{\mathbf{z}}^{(2)}) = (.057) \begin{bmatrix} -.2974 \\ .6739 \end{bmatrix} [-.3564 \quad .0227]$$

$$= \begin{bmatrix} .006 & -.000 \\ -.014 & .001 \end{bmatrix}$$

$$\mathbf{R}_{11} - \text{sample Cov}(\tilde{\mathbf{z}}^{(1)}) = \begin{bmatrix} -.2974 \\ .6739 \end{bmatrix} [-.2974 \quad .6739]$$

$$= \begin{bmatrix} .088 & -.200 \\ -.200 & .454 \end{bmatrix}$$

$$\mathbf{R}_{22} - \text{sample Cov}(\tilde{\mathbf{z}}^{(2)}) = \begin{bmatrix} -.3564 \\ .0227 \end{bmatrix} [-.3564 \quad .0227]$$

$$= \begin{bmatrix} .127 & -.008 \\ -.008 & .001 \end{bmatrix}$$

where $\tilde{\mathbf{z}}^{(1)}, \tilde{\mathbf{z}}^{(2)}$ are given by (10-33) with $r = 1$ and $\hat{\mathbf{a}}_z^{(1)}, \hat{\mathbf{b}}_z^{(1)}$ replace $\hat{\mathbf{a}}^{(1)}, \hat{\mathbf{b}}^{(1)}$, respectively.

We see that the first pair of canonical variables effectively summarizes (reproduces) the intraset correlations in $\mathbf{R}_{12}$. However, the individual variates are not particularly effective summaries of the sampling variability in the original $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(2)}$ sets, respectively. This is especially true for $\hat{U}_1$. ■

Proportions of Explained Sample Variance

When the observations are standardized, the sample covariance matrices $\mathbf{S}_{kl}$ are correlation matrices $\mathbf{R}_{kl}$. The canonical coefficient vectors are the *rows* of the matrices $\hat{\mathbf{A}}_{\mathbf{z}}$ and $\hat{\mathbf{B}}_{\mathbf{z}}$ and the *columns* of $\hat{\mathbf{A}}_{\mathbf{z}}^{-1}$ and $\hat{\mathbf{B}}_{\mathbf{z}}^{-1}$ are the sample correlations between the canonical variates and their component variables.

Specifically,

$$\text{sample Cov}(\mathbf{z}^{(1)}, \hat{\mathbf{U}}) = \text{sample Cov}(\hat{\mathbf{A}}_{\mathbf{z}}^{-1}\hat{\mathbf{U}}, \hat{\mathbf{U}}) = \hat{\mathbf{A}}_{\mathbf{z}}^{-1}$$

and

$$\text{sample Cov}(\mathbf{z}^{(2)}, \hat{\mathbf{V}}) = \text{sample Cov}(\hat{\mathbf{B}}_{\mathbf{z}}^{-1}\hat{\mathbf{V}}, \hat{\mathbf{V}}) = \hat{\mathbf{B}}_{\mathbf{z}}^{-1}$$

so

$$\hat{\mathbf{A}}_{\mathbf{z}}^{-1} = [\hat{\mathbf{a}}_{\mathbf{z}}^{(1)}, \hat{\mathbf{a}}_{\mathbf{z}}^{(2)}, \ldots, \hat{\mathbf{a}}_{\mathbf{z}}^{(p)}] = \begin{bmatrix} r_{\hat{U}_1, z_1^{(1)}} & r_{\hat{U}_2, z_1^{(1)}} & \cdots & r_{\hat{U}_p, z_1^{(1)}} \\ r_{\hat{U}_1, z_2^{(1)}} & r_{\hat{U}_2, z_2^{(1)}} & \cdots & r_{\hat{U}_p, z_2^{(1)}} \\ \vdots & \vdots & \ddots & \vdots \\ r_{\hat{U}_1, z_p^{(1)}} & r_{\hat{U}_2, z_p^{(1)}} & \cdots & r_{\hat{U}_p, z_p^{(1)}} \end{bmatrix}$$

$$\hat{\mathbf{B}}_{\mathbf{z}}^{-1} = [\hat{\mathbf{b}}_{\mathbf{z}}^{(1)}, \hat{\mathbf{b}}_{\mathbf{z}}^{(2)}, \ldots, \hat{\mathbf{b}}_{\mathbf{z}}^{(q)}] = \begin{bmatrix} r_{\hat{V}_1, z_1^{(2)}} & r_{\hat{V}_2, z_1^{(2)}} & \cdots & r_{\hat{V}_q, z_1^{(2)}} \\ r_{\hat{V}_1, z_2^{(2)}} & r_{\hat{V}_2, z_2^{(2)}} & \cdots & r_{\hat{V}_q, z_2^{(2)}} \\ \vdots & \vdots & \ddots & \vdots \\ r_{\hat{V}_1, z_q^{(2)}} & r_{\hat{V}_2, z_q^{(2)}} & \cdots & r_{\hat{V}_q, z_q^{(2)}} \end{bmatrix} \tag{10-35}$$

where $r_{\hat{U}_i, z_k^{(1)}}$ and $r_{\hat{V}_i, z_k^{(1)}}$ are the sample correlation coefficients between the quantities with subscripts.

Using (10-32) with standardized observations, we obtain

Total (standardized) sample variance in first set

$$= \text{tr}(\mathbf{R}_{11}) = \text{tr}(\hat{\mathbf{a}}_{\mathbf{z}}^{(1)}\hat{\mathbf{a}}_{\mathbf{z}}^{(1)\prime} + \hat{\mathbf{a}}_{\mathbf{z}}^{(2)}\hat{\mathbf{a}}_{\mathbf{z}}^{(2)\prime} + \cdots + \hat{\mathbf{a}}_{\mathbf{z}}^{(p)}\hat{\mathbf{a}}_{\mathbf{z}}^{(p)\prime}) = p \tag{10-36a}$$

Total (standardized) sample variance in second set

$$= \text{tr}(\mathbf{R}_{22}) = \text{tr}(\hat{\mathbf{b}}_{\mathbf{z}}^{(1)}\hat{\mathbf{b}}_{\mathbf{z}}^{(1)\prime} + \hat{\mathbf{b}}_{\mathbf{z}}^{(2)}\hat{\mathbf{b}}_{\mathbf{z}}^{(2)\prime} + \cdots + \hat{\mathbf{b}}_{\mathbf{z}}^{(q)}\hat{\mathbf{b}}_{\mathbf{z}}^{(q)\prime}) = q \tag{10-36b}$$

Since the correlations in the first $r < p$ columns of $\hat{\mathbf{A}}_{\mathbf{z}}^{-1}$ and $\hat{\mathbf{B}}_{\mathbf{z}}^{-1}$ involve only the sample canonical variates $\hat{U}_1, \hat{U}_2, \ldots, \hat{U}_r$ and $\hat{V}_1, \hat{V}_2, \ldots, \hat{V}_r$, respectively, we define

the contributions of the first r canonical variates to the total (standardized) sample variances as

$$\operatorname{tr}(\hat{\mathbf{a}}_{\mathbf{z}}^{(1)}\hat{\mathbf{a}}_{\mathbf{z}}^{(1)\prime} + \hat{\mathbf{a}}_{\mathbf{z}}^{(2)}\hat{\mathbf{a}}_{\mathbf{z}}^{(2)\prime} + \cdots + \hat{\mathbf{a}}_{\mathbf{z}}^{(r)}\hat{\mathbf{a}}_{\mathbf{z}}^{(r)\prime}) = \sum_{i=1}^{r}\sum_{k=1}^{p} r^2_{\hat{U}_i, z_k^{(1)}}$$

and

$$\operatorname{tr}(\hat{\mathbf{b}}_{\mathbf{z}}^{(1)}\hat{\mathbf{b}}_{\mathbf{z}}^{(1)\prime} + \hat{\mathbf{b}}_{\mathbf{z}}^{(2)}\hat{\mathbf{b}}_{\mathbf{z}}^{(2)\prime} + \cdots + \hat{\mathbf{b}}_{\mathbf{z}}^{(r)}\hat{\mathbf{b}}_{\mathbf{z}}^{(r)\prime}) = \sum_{i=1}^{r}\sum_{k=1}^{p} r^2_{\hat{V}_i, z_k^{(2)}}$$

The *proportions* of total (standardized) sample variances "explained by" the first r canonical variates then become

$$R^2_{\mathbf{z}^{(1)}|\hat{U}_1,\hat{U}_2,\ldots,\hat{U}_r} = \begin{pmatrix}\text{proportion of total standardized}\\ \text{sample variance in first set}\\ \text{explained by } \hat{U}_1, \hat{U}_2, \ldots, \hat{U}_r\end{pmatrix}$$

$$= \frac{\operatorname{tr}(\hat{\mathbf{a}}_{\mathbf{z}}^{(1)}\hat{\mathbf{a}}_{\mathbf{z}}^{(1)\prime} + \cdots + \hat{\mathbf{a}}_{\mathbf{z}}^{(r)}\hat{\mathbf{a}}_{\mathbf{z}}^{(r)\prime})}{\operatorname{tr}(\mathbf{R}_{11})}$$

$$= \frac{\sum_{i=1}^{r}\sum_{k=1}^{p} r^2_{\hat{U}_i, z_k^{(1)}}}{p} \tag{10-37}$$

and

$$R^2_{\mathbf{z}^{(2)}|\hat{V}_1,\hat{V}_2,\ldots,\hat{V}_r} = \begin{pmatrix}\text{proportion of total standardized}\\ \text{sample variance in second set}\\ \text{explained by } \hat{V}_1, \hat{V}_2, \ldots, \hat{V}_r\end{pmatrix}$$

$$= \frac{\operatorname{tr}(\hat{\mathbf{b}}_{\mathbf{z}}^{(1)}\hat{\mathbf{b}}_{\mathbf{z}}^{(1)\prime} + \cdots + \hat{\mathbf{b}}_{\mathbf{z}}^{(r)}\hat{\mathbf{b}}_{\mathbf{z}}^{(r)\prime})}{\operatorname{tr}(\mathbf{R}_{22})}$$

$$= \frac{\sum_{i=1}^{r}\sum_{k=1}^{q} r^2_{\hat{V}_i, z_k^{(2)}}}{q}$$

Descriptive measures (10-37) provide some indication of how well the canonical variates represent their respective sets. They provide single-number descriptions of the matrices of errors. In particular,

$$\frac{1}{p}\operatorname{tr}[\mathbf{R}_{11} - \hat{\mathbf{a}}_{\mathbf{z}}^{(1)}\hat{\mathbf{a}}_{\mathbf{z}}^{(1)\prime} - \hat{\mathbf{a}}_{\mathbf{z}}^{(2)}\hat{\mathbf{a}}_{\mathbf{z}}^{(2)\prime} - \cdots - \hat{\mathbf{a}}_{\mathbf{z}}^{(r)}\hat{\mathbf{a}}_{\mathbf{z}}^{(r)\prime}] = 1 - R^2_{\mathbf{z}^{(1)}|\hat{U}_1,\hat{U}_2,\ldots,\hat{U}_r}$$

$$\frac{1}{q}\operatorname{tr}[\mathbf{R}_{22} - \hat{\mathbf{b}}_{\mathbf{z}}^{(1)}\hat{\mathbf{b}}_{\mathbf{z}}^{(1)\prime} - \hat{\mathbf{b}}_{\mathbf{z}}^{(2)}\hat{\mathbf{b}}_{\mathbf{z}}^{(2)\prime} - \cdots - \hat{\mathbf{b}}_{\mathbf{z}}^{(r)}\hat{\mathbf{b}}_{\mathbf{z}}^{(r)\prime}] = 1 - R^2_{\mathbf{z}^{(2)}|\hat{V}_1,\hat{V}_2,\ldots,\hat{V}_r}$$

according to (10-36) and (10-37).

Example 10.7 (Calculating proportions of sample variance explained by canonical variates) Consider the job characteristic–job satisfaction data discussed in Example 10.5. Using the table of sample correlation coefficients presented in that example, we find that

$$R^2_{\mathbf{z}^{(1)}|\hat{U}_1} = \frac{1}{5}\sum_{k=1}^{5} r^2_{\hat{U}_1, z_k^{(1)}} = \frac{1}{5}[(.83)^2 + (.74)^2 + \cdots + (.85)^2] = .58$$

$$R^2_{\mathbf{z}^{(2)}|\hat{V}_1} = \frac{1}{7}\sum_{k=1}^{7} r^2_{\hat{V}_1, z_k^{(2)}} = \frac{1}{7}[(.75)^2 + (.65)^2 + \cdots + (.50)^2] = .37$$

The first sample canonical variate $\hat{U}_1$ of the job characteristics set accounts for 58% of the set's total sample variance. The first sample canonical variate $\hat{V}_1$ of the job satisfaction set explains 37% of the set's total sample variance. We might thus infer that $\hat{U}_1$ is a "better" representative of its set than $\hat{V}_1$ is of its set. The interested reader may wish to see how well $\hat{U}_1$ and $\hat{V}_1$ reproduce the correlation matrices $\mathbf{R}_{11}$ and $\mathbf{R}_{22}$, respectively. [See (10-29).] ■

10.6 Large Sample Inferences

When $\boldsymbol{\Sigma}_{12} = \mathbf{0}$, $\mathbf{a}'\mathbf{X}^{(1)}$ and $\mathbf{b}'\mathbf{X}^{(2)}$ have covariance $\mathbf{a}'\boldsymbol{\Sigma}_{12}\mathbf{b} = 0$ for all vectors $\mathbf{a}$ and $\mathbf{b}$. Consequently, all the canonical correlations must be zero, and there is no point in pursuing a canonical correlation analysis. The next result provides a way of testing $\boldsymbol{\Sigma}_{12} = \mathbf{0}$, for large samples.

Result 10.3. Let

$$\mathbf{X}_j = \begin{bmatrix} \mathbf{X}_j^{(1)} \\ \cdots\cdots \\ \mathbf{X}_j^{(2)} \end{bmatrix}, \qquad j = 1, 2, \ldots, n$$

be a random sample from an $N_{p+q}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ population with

$$\boldsymbol{\Sigma} = \left[\begin{array}{c|c} \underset{(p\times p)}{\boldsymbol{\Sigma}_{11}} & \underset{(p\times q)}{\boldsymbol{\Sigma}_{12}} \\ \hline \underset{(q\times p)}{\boldsymbol{\Sigma}_{21}} & \underset{(q\times q)}{\boldsymbol{\Sigma}_{22}} \end{array}\right]$$

Then the likelihood ratio test of H_0: $\boldsymbol{\Sigma}_{12} = \underset{(p\times q)}{\mathbf{0}}$ versus H_1: $\boldsymbol{\Sigma}_{12} \neq \underset{(p\times q)}{\mathbf{0}}$ rejects H_0 for large values of

$$-2 \ln \Lambda = n \ln\left(\frac{|\mathbf{S}_{11}||\mathbf{S}_{22}|}{|\mathbf{S}|}\right) = -n \ln \prod_{i=1}^{p} (1 - \widehat{\rho_i^{*2}}) \qquad (10\text{-}38)$$

where

$$\mathbf{S} = \left[\begin{array}{c|c} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \hline \mathbf{S}_{21} & \mathbf{S}_{22} \end{array}\right]$$

is the unbiased estimator of $\boldsymbol{\Sigma}$. For large n, the test statistic (10-38) is approximately distributed as a chi-square random variable with pq d.f.

Proof. See Kshirsagar [8].

The likelihood ratio statistic (10-38) compares the sample generalized variance under H_0, namely,

$$\begin{vmatrix} \mathbf{S}_{11} & \mathbf{0} \\ \mathbf{0}' & \mathbf{S}_{22} \end{vmatrix} = |\mathbf{S}_{11}||\mathbf{S}_{22}|$$

with the unrestricted generalized variance $|\mathbf{S}|$.

Bartlett [3] suggests replacing the multiplicative factor n in the likelihood ratio statistic with the factor $n - 1 - \frac{1}{2}(p + q + 1)$ to improve the χ^2 approximation to the sampling distribution of $-2 \ln \Lambda$. Thus, for n and $n - (p$ large, we

Reject H_0: $\boldsymbol{\Sigma}_{12} = \mathbf{0}$ $(\rho_1^* = \rho_2^* = \cdots = \rho_p^* = 0)$ at significance level α if

$$-\left(n - 1 - \frac{1}{2}(p + q + 1)\right) \ln \prod_{i=1}^{p} (1 - \widehat{\rho_i^{*2}}) > \chi_{pq}^2(\alpha) \qquad (10\text{-}3$$

where $\chi_{pq}^2(\alpha)$ is the upper (100α)th percentile of a chi-square distribution with pq d.f.

If the null hypothesis H_0: $\boldsymbol{\Sigma}_{12} = \mathbf{0}$ $(\rho_1^* = \rho_2^* = \cdots = \rho_p^* = 0)$ is rejected, it is natural to examine the "significance" of the individual canonical correlations. Since the canonical correlations are ordered from the largest to the smallest, we can begin by assuming that the first canonical correlation is nonzero and the remaining $p -$ canonical correlations are zero. If this hypothesis is rejected, we assume that the first two canonical correlations are nonzero, but the remaining $p - 2$ canonical correlations are zero, and so forth.

Let the implied sequence of hypotheses be

$$H_0^k: \rho_1^* \neq 0, \rho_2^* \neq 0, \ldots, \rho_k^* \neq 0, \rho_{k+1}^* = \cdots = \rho_p^* = 0$$

$$H_1^k: \rho_i^* \neq 0, \text{ for some } i \geq k + 1$$

Bartlett [2] has argued that the kth hypothesis in (10-40) can be tested by the likelihood ratio criterion. Specifically,

Reject $H_0^{(k)}$ at significance level α if

$$-\left(n - 1 - \frac{1}{2}(p + q + 1)\right) \ln \prod_{i=k+1}^{p} (1 - \widehat{\rho}_i^{*2}) > \chi^2_{(p-k)(q-k)}(\alpha) \qquad (10\text{-}41)$$

where $\chi^2_{(p-k)(q-k)}(\alpha)$ is the upper (100α)th percentile of a chi-square distribution with $(p - k)(q - k)$ d.f. We point out that the test statistic in (10-41) involves $\prod_{i=k+1}^{p} (1 - \widehat{\rho}_i^{*2})$, the "residual" after the first k sample canonical correlations have been removed from the total criterion $\Lambda^{2/n} = \prod_{i=1}^{p} (1 - \widehat{\rho}_i^{*2})$.

If the members of the sequence $H_0, H_0^{(1)}, H_0^{(2)}$, and so forth, are tested one at a time until $H_0^{(k)}$ is not rejected for some k, the overall significance level is not α and, in fact, would be difficult to determine. Another defect of this procedure is the tendency it induces to conclude that a null hypothesis is correct simply because it is not rejected.

To summarize, the overall test of significance in Result 10.3 is useful for multivariate normal data. The sequential tests implied by (10-41) should be interpreted with caution and are, perhaps, best regarded as rough guides for selecting the number of important canonical variates.

Example 10.8 (Testing the significance of the canonical correlations for the job satisfaction data) Test the significance of the canonical correlations exhibited by the job characteristics–job satisfaction data introduced in Example 10.5.

All the test statistics of immediate interest are summarized in the table on page 566. From Example 10.5, $n = 784$, $p = 5$, $q = 7$, $\widehat{\rho}_1^* = .55$, $\widehat{\rho}_2^* = .23$, $\widehat{\rho}_3^* = .12$, $\widehat{\rho}_4^* = .08$, and $\widehat{\rho}_5^* = .05$.

Assuming multivariate normal data, we find that the first two canonical correlations, ρ_1^* and ρ_2^*, appear to be nonzero, although with the very large sample size, small deviations from zero will show up as statistically significant. From a practical point of view, the second (and subsequent) sample canonical correlations can probably be ignored, since (1) they are reasonably small in magnitude and (2) the corresponding canonical variates explain *very* little of the sample variation in the variable sets $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$. ■

The distribution theory associated with the sample canonical correlations and the sample canonical variate coefficients is extremely complex (apart from the $p = 1$ and $q = 1$ situations), even in the null case, $\boldsymbol{\Sigma}_{12} = \mathbf{0}$. The reader interested in the distribution theory is referred to Kshirsagar [8].

Test Results

Null hypothesis	Observed test statistic (Barlett correction)	Degrees of freedom	Upper 1% point of χ^2 distribution	Conclusion
1. $H_0: \boldsymbol{\Sigma}_{12} = \mathbf{0}$	$-\left(n - 1 - \frac{1}{2}(p + q + 1)\right) \ln \prod_{i=1}^{5} (1 - \widehat{\rho_i^{*2}})$	$pq = 5(7) = 35$	$\chi^2_{35}(.01) = 57$	Reject H_0.
(all $\rho_i^* = 0$)	$= -\left(784 - 1 - \frac{1}{2}(5 + 7 + 1)\right) \ln(.6453)$			
	$= 340.1$			
2. $H_0^{(1)}: \rho_1^* \neq 0,$	$-\left(n - 1 - \frac{1}{2}(p + q + 1)\right) \ln \prod_{i=2}^{5} (1 - \widehat{\rho_i^{*2}})$	$(p - 1)(q - 1) = 24$	$\chi^2_{24}(.01) = 42.98$	Reject H_0.
$\rho_2^* = \cdots = \rho_5^* = 0$	$= 60.4$			
3. $H_0^{(2)}: \rho_1^* \neq 0, \rho_2^* \neq 0,$	$-\left(n - 1 - \frac{1}{2}(p + q + 1)\right) \ln \prod_{i=3}^{5} (1 - \widehat{\rho_i^{*2}})$	$(p - 2)(q - 2) = 15$	$\chi^2_{15}(.01) = 30.58$	Do not reject H_0.
$\rho_3^* = \cdots = \rho_5^* = 0$	$= 18.2$			

6-2. Discrimination and Classification

- Discrimination and classification are multivariate techniques concerned with *separating* distinct sets of objects (or observations) and with *allocating* new objects (observations) to previously defined group.

  Goal 1. To describe, either graphically (in three or fewer dimensions) or algebraically, the differential features of objects (observations) from several known collections (population). We try to find *discriminants* whose numerical values are such that the collections are separated as much as possible.

  Goal 2. To sort objects (observations) into two or more labeled classes. The emphasis is on deriving a rule that can be used to optimally assign *new* objects to the labeled classes.

Separation and Classification for Two populations

Populations π_1 and π_2	Measured variables $\mathbf{X}$
1. Solvent and distressed property-liability insurance companies.	Total assets, cost of stocks and bonds, market value of stocks and bonds, loss expenses, surplus, amount of premiums written.
2. Nonulcer dyspeptics (those with upset stomach problems) and controls ("normal").	Measures of anxiety, dependence, guilt, perfectionism.
3. *Federalist Papers* written by James Madison and those written by Alexander Hamilton.	Frequencies of different words and lengths of sentences.
4. Two species of chickweed.	Sepal and petal length, petal cleft depth, bract length, scarious tip length, pollen diameter.
5. Purchasers of a new product and laggards (those "slow" to purchase).	Education, income, family size, amount of previous brand switching.
6. Successful or unsuccessful (fail to graduate) college students.	Entrance examination scores, high school grade-point average, number of high school activities.
7. Males and females.	Anthropological measurements, like circumference and volume on ancient skulls.
8. Good and poor credit risks.	Income, age, number of credit cards, family size.
9. Alcoholics and nonalcoholics.	Activity of monoamine oxidase enzyme, activity of adenylate cyclase enzyme.

- Allocation or classification rules are usually developed from *learning* samples. Measured characteristics of randomly selected object *known* to come from each of the two populations are examined for differences.

- Why we *know* that some observations belong to a particular population, but we are unsure about others.
  - Incomplete knowledge of future performance
  - *Perfect* information requires destroying the object.
  - Unavailable or expensive information.

Example 11.1 (Discriminating owners from nonowners of riding mowers) Consider two groups in a city: π_1, riding-mower owners, and π_2, those without riding mowers—that is, nonowners. In order to identify the best sales prospects for an intensive sales campaign, a riding-mower manufacturer is interested in classifying families as prospective owners or nonowners on the basis of x_1 = income and x_2 = lot size. Random samples of $n_1 = 12$ current owners and $n_2 = 12$ current nonowners yield the values in Table 11.1.

Table 11.1

π_1: Riding-mower owners		π_2: Nonowners	
x_1 (Income in \$1000s)	x_2 (Lot size in 1000 ft^2)	x_1 (Income in \$1000s)	x_2 (Lot size in 1000 ft^2)
90.0	18.4	105.0	19.6
115.5	16.8	82.8	20.8
94.8	21.6	94.8	17.2
91.5	20.8	73.2	20.4
117.0	23.6	114.0	17.6
140.1	19.2	79.2	17.6
138.0	17.6	89.4	16.0
112.8	22.4	96.0	18.4
99.0	20.0	77.4	16.4
123.0	20.8	63.0	18.8
81.0	22.0	81.0	14.0
111.0	20.0	93.0	14.8

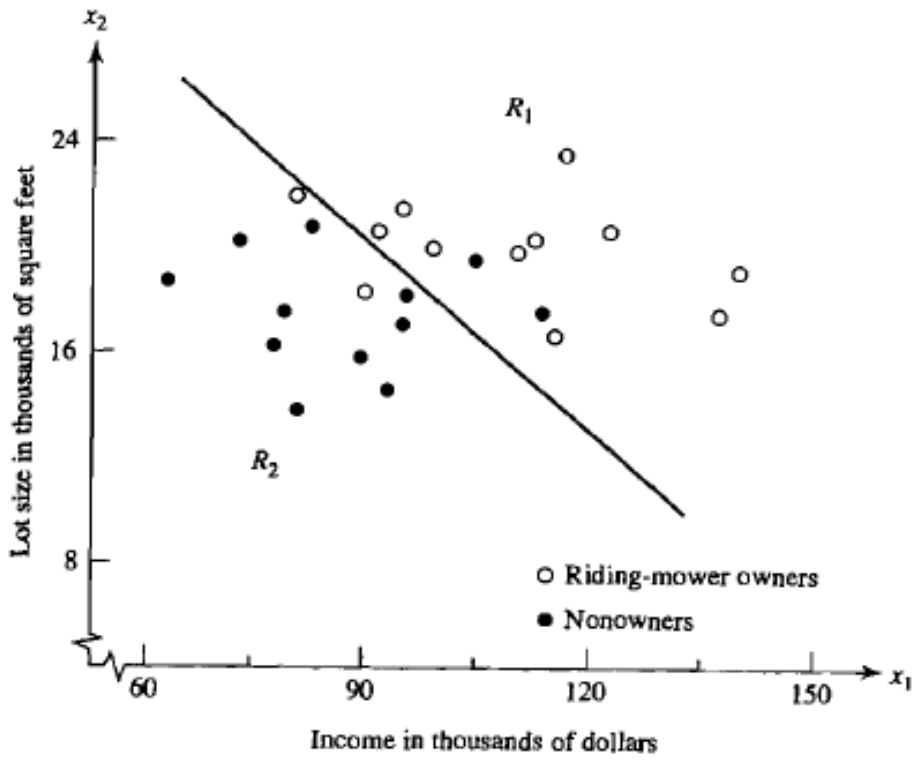

Figure 11.1 Income and lot size for riding-mower owners and nonowners.

- Let $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ be the probability density functions associated with $p \times 1$ vector random variable $\boldsymbol{X}$ for the populations π_1 and π_2 respectively.

- An object with associate measurements **x** must be assigned to either π_1 or π_2.

- Let Ω be the complete sample space, let R_1 be that set of **x** values for which we classify objects as π_1 and $R_2 = \Omega - R_1$ be the remaining **x** values for which we classify objects as π_2. So $R_1 \cup R_2 = \Omega$ and $R_1 \cap R_2 = \emptyset$.

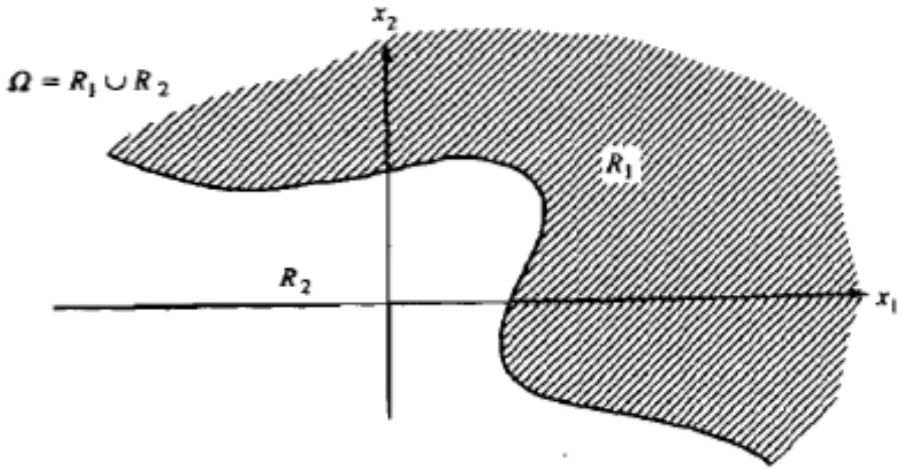

Figure 11.2 Classification regions for two populations.

- The conditional probability , $P(2|1)$, of classifying an object as π_2 when, in fact, it is from π_1 is

$$P(2|1) = P(\boldsymbol{X} \in R_2|\pi_1) = \int_{R_2=\Omega-R_1} f_1(\mathbf{x})d\mathbf{x}$$

- Similar, the conditional probability , $P(1|2)$, of classifying an object as π_1 when, in fact, it is from π_2 is

$$P(1|2) = P(\boldsymbol{X} \in R_1|\pi_2) = \int_{R_1=\Omega-R_2} f_1(\mathbf{x})d\mathbf{x}$$

.

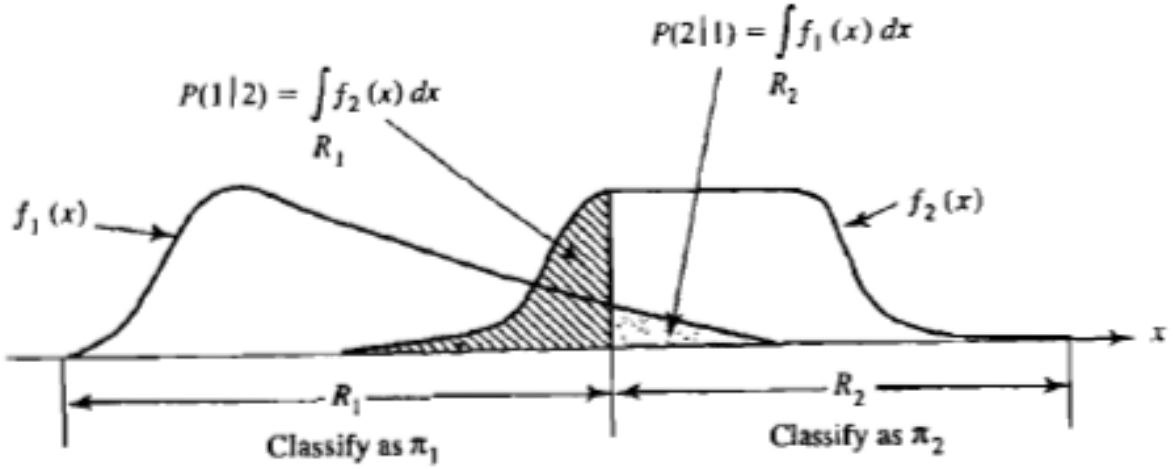

Figure 11.3 Misclassification probabilities for hypothetical classification regions when $p = 1$.

Let p_1 be the prior probability of π_1 and p_2 be the prior probability of π_2, where $p_1 + p_2 = 1$. Then

$$
\begin{aligned}
& P(\text{observation is correlectly classified as } \pi_1) \\
= \quad & P(\boldsymbol{X} \in R_1 | \pi_1) P(\pi_1) = P(1|1) p_1, \\
& P(\text{observation is misclassified as } \pi_1) \\
= \quad & P(\boldsymbol{X} \in R_1 | \pi_2) P(\pi_2) = P(1|2) p_2, \\
& P(\text{observation is correlectly classified as } \pi_2) \\
= \quad & P(\boldsymbol{X} \in R_2 | \pi_2) P(\pi_2) = P(2|2) p_2, \\
& P(\text{observation is misclassified as } \pi_2) \\
= \quad & P(\boldsymbol{X} \in R_2 | \pi_1) P(\pi_1) = P(2|1) p_1,
\end{aligned}
$$

- The costs of misclassification is defined by a cost matrix

	π_1	π_2
π_1	0	$c(2\|1)$
π_2	$c(1\|2)$	0

- *Expected cost of misclassification* (ECM)

$$\text{ECM} = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2$$

A reasonable classification rule should have an ECM as small, or nearly as small, as possible.

Results 6-2.1 The region R_1 and R_2 that minimize the ECM are defined by the value **x** for which the following inequalities hold:

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$\left(\begin{array}{c}\text{density}\\ \text{ratio}\end{array}\right) \geq \left(\begin{array}{c}\text{cost}\\ \text{ratio}\end{array}\right)\left(\begin{array}{c}\text{prior probability}\\ \text{ratio}\end{array}\right)$$

$$R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$\left(\begin{array}{c}\text{density}\\ \text{ratio}\end{array}\right) < \left(\begin{array}{c}\text{cost}\\ \text{ratio}\end{array}\right)\left(\begin{array}{c}\text{prior probability}\\ \text{ratio}\end{array}\right)$$

Special Cases of Minimum Expected Cost Regions

(a) $p_2/p_1 = 1$ (equal prior probabilities)

$$R_1: \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(1|2)}{c(2|1)} \qquad R_2: \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{c(1|2)}{c(2|1)}$$

(b) $c(1|2)/c(2|1) = 1$ (equal misclassification costs)

$$R_1: \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{p_2}{p_1} \qquad R_2: \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{p_2}{p_1} \tag{11-7}$$

(c) $p_2/p_1 = c(1|2)/c(2|1) = 1$ or $p_2/p_1 = 1/(c(1|2)/c(2|1))$
(equal prior probabilities and equal misclassification costs)

$$R_1: \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq 1 \qquad R_2: \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < 1$$

Example 11.2 (Classifying a new observation into one of the two populations) A researcher has enough data available to estimate the density functions $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ associated with populations π_1 and π_2, respectively. Suppose $c(2|1) = 5$ units and $c(1|2) = 10$ units. In addition, it is known that about 20% of *all* objects (for which the measurements $\mathbf{x}$ can be recorded) belong to π_2. Thus, the prior probabilities are $p_1 = .8$ and $p_2 = .2$.

Given the prior probabilities and costs of misclassification, we can use (11-6) to derive the classification regions R_1 and R_2. Specifically, we have

$$R_1: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left(\frac{10}{5}\right)\left(\frac{.2}{.8}\right) = .5$$

$$R_2: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left(\frac{10}{5}\right)\left(\frac{.2}{.8}\right) = .5$$

Suppose the density functions evaluated at a new observation $\mathbf{x}_0$ give $f_1(\mathbf{x}_0) = .3$ and $f_2(\mathbf{x}_0) = .4$. Do we classify the new observation as π_1 or π_2? To answer the question, we form the ratio

$$\frac{f_1(\mathbf{x}_0)}{f_2(\mathbf{x}_0)} = \frac{.3}{.4} = .75$$

and compare it with .5 obtained before. Since

$$\frac{f_1(\mathbf{x}_0)}{f_2(\mathbf{x}_0)} = .75 > \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) = .5$$

we find that $\mathbf{x}_0 \in R_1$ and classify it as belonging to π_1. ■

Other classification procedures:

- Choose R_1 and R_2 to minimize the *total probability of misclassification* (TPM).

$$\text{TPM} = p_1 \int_{R_2} f_1(\mathbf{x}) d\mathbf{x} + p_2 \int_{R_1} f_2(\mathbf{x}) d\mathbf{x}.$$

- Allocate a new observation $\mathbf{x}_0$ to the population with the largest *posterior* probability $P(\pi_i|\mathbf{x}_0)$.

$$P(\pi_1|\mathbf{x}_0) = \frac{p_1 f_1(\mathbf{x}_0)}{p_1 f_1(\mathbf{x}_0) + p_2 f_2(\mathbf{x}_0)},$$

$$P(\pi_2|\mathbf{x}_0) = 1 - P(\pi_1|\mathbf{x}_0) = \frac{p_2 f_2(\mathbf{x}_0)}{p_1 f_1(\mathbf{x}_0) + p_2 f_2(\mathbf{x}_0)}.$$

Classifying an observation $\mathbf{x}_0$ as π_1 when $P(\pi_1|\mathbf{x}_0) > P(\pi_2|\mathbf{x}_0)$ is equivalent to using the (b) rule for total probability of misclassification.

Classification with Two Multivariate Normal Populations

Assume $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are multivariate normal densities , the first with mean vector $\boldsymbol{\mu}_1$ and covariance Σ_1 and the second with mean $\boldsymbol{\mu}_2$ and covariance Σ_2.

Suppose that joint densities of $\boldsymbol{X}' = [X_1, X_2, \ldots, X_p]$ for population π_1 and π_2 are given by

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)'\Sigma(\mathbf{x} - \boldsymbol{\mu}_i)\right] \quad \text{for } i = 1, 2.$$

Result 6-2.2. Let the populations π_1 and π_2 be described by multivariate normal densities of the form above . Then the allocation rule that minimizes the ECM is as follows:
Allocate $\mathbf{x}_0$ to π_1 if

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}\mathbf{x}_0 - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\Sigma^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right],$$

Allocate $\mathbf{x}_0$ to π_2 otherwise.

The Estimated Minimum ECM Rule for Two Normal Population

Allocate $\mathbf{x}_0$ to π_1 if

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{pooled}^{-1}\mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\mathbf{S}_{pooled}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right],$$

Allocate $\mathbf{x}_0$ to π_2 otherwise.

Example 11.3 (Classification with two normal populations—common Σ and equal costs) This example is adapted from a study [4] concerned with the detection of hemophilia A carriers. (See also Exercise 11.32.)

To construct a procedure for detecting potential hemophilia A carriers, blood samples were assayed for two groups of women and measurements on the two variables,

$$X_1 = \log_{10}(\text{AHF activity})$$
$$X_2 = \log_{10}(\text{AHF-like antigen})$$

recorded. ("AHF" denotes antihemophilic factor.) The first group of $n_1 = 30$ women were selected from a population of women who did not carry the hemophilia gene. This group was called the *normal* group. The second group of $n_2 = 22$ women was selected from known hemophilia A carriers (daughters of hemophiliacs, mothers with more than one hemophilic son, and mothers with one hemophilic son and other hemophilic relatives). This group was called the *obligatory carriers*. The pairs of observations (x_1, x_2) for the two groups are plotted in Figure 11.4. Also shown are estimated contours containing 50% and 95% of the probability for bivariate normal distributions centered at $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$, respectively. Their common covariance matrix was taken as the pooled sample covariance matrix $\mathbf{S}_{\text{pooled}}$. In this example, bivariate normal distributions seem to fit the data fairly well.

The investigators (see [4]) provide the information

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} -.0065 \\ -.0390 \end{bmatrix}, \qquad \bar{\mathbf{x}}_2 = \begin{bmatrix} -.2483 \\ .0262 \end{bmatrix}$$

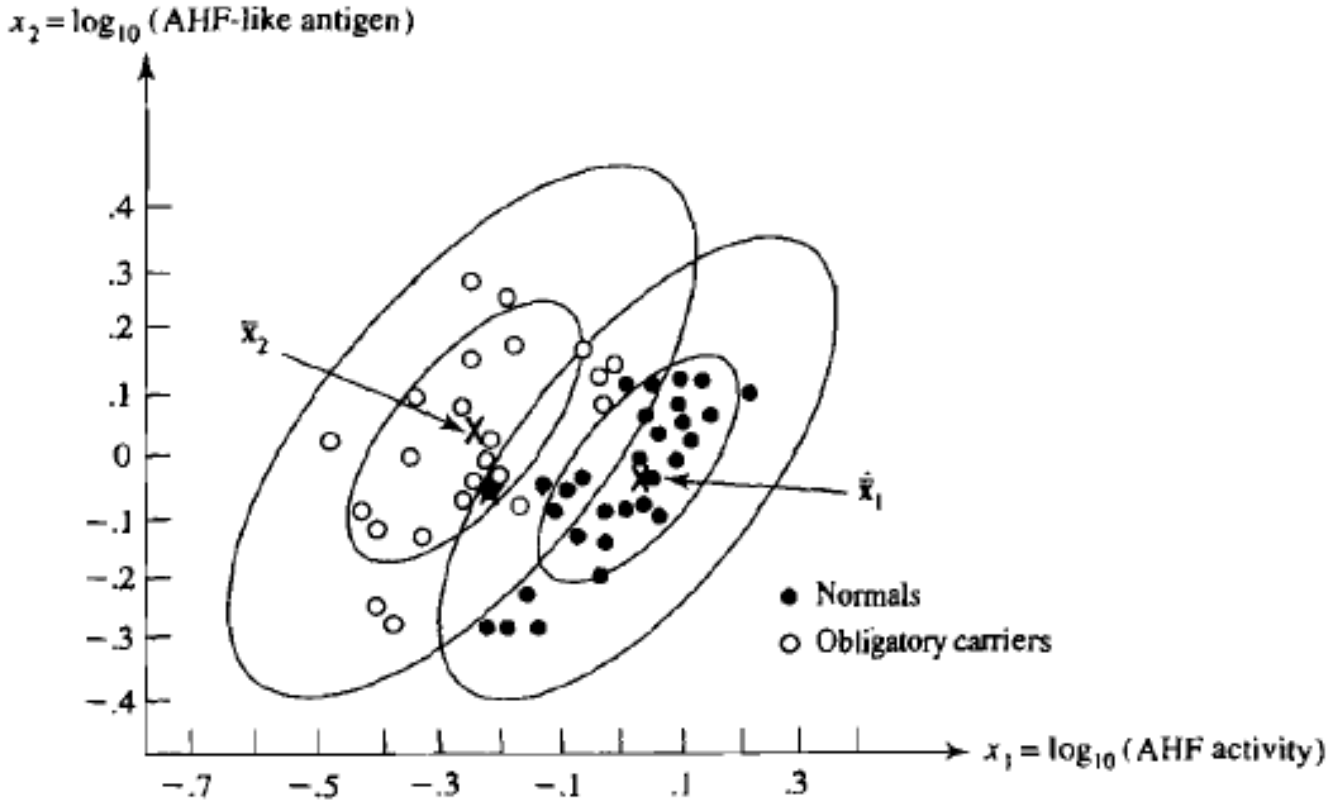

Figure 11.4 Scatter plots of [$\log_{10}$(AHF activity), $\log_{10}$(AHF-like antigen)] for the normal group and obligatory hemophilia A carriers.

and

$$\mathbf{S}_{\text{pooled}}^{-1} = \begin{bmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{bmatrix}$$

Therefore, the equal costs and equal priors discriminant function [see (11-19)] is

$$\begin{aligned} \hat{y} = \hat{\mathbf{a}}'\mathbf{x} &= [\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2]'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x} \\ &= [.2418 \quad -.0652]\begin{bmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= 37.61x_1 - 28.92x_2 \end{aligned}$$

Moreover,

$$\bar{y}_1 = \hat{\mathbf{a}}'\bar{\mathbf{x}}_1 = [37.61 \quad -28.92]\begin{bmatrix} -.0065 \\ -.0390 \end{bmatrix} = .88$$

$$\bar{y}_2 = \hat{\mathbf{a}}'\bar{\mathbf{x}}_2 = [37.61 \quad -28.92]\begin{bmatrix} -.2483 \\ .0262 \end{bmatrix} = -10.10$$

and the midpoint between these means [see (11-20)] is

$$\hat{m} = \tfrac{1}{2}(\bar{y}_1 + \bar{y}_2) = \tfrac{1}{2}(.88 - 10.10) = -4.61$$

Measurements of AHF activity and AHF-like antigen on a woman who may be a hemophilia A carrier give $x_1 = -.210$ and $x_2 = -.044$. Should this woman be classified as π_1 (normal) or π_2 (obligatory carrier)?

Using (11-18) with equal costs and equal priors so that $\ln(1) = 0$, we obtain

$$\text{Allocate } \mathbf{x}_0 \text{ to } \pi_1 \text{ if } \hat{y}_0 = \hat{\mathbf{a}}'\mathbf{x}_0 \geq \hat{m} = -4.61$$

$$\text{Allocate } \mathbf{x}_0 \text{ to } \pi_2 \text{ if } \hat{y}_0 = \hat{\mathbf{a}}'\mathbf{x}_0 < \hat{m} = -4.61$$

where $\mathbf{x}'_0 = [-.210, -.044]$. Since

$$\hat{y}_0 = \hat{\mathbf{a}}'\mathbf{x}_0 = [37.61 \quad -28.92]\begin{bmatrix} -.210 \\ -.044 \end{bmatrix} = -6.62 < -4.61$$

we classify the woman as π_2, an obligatory carrier. The new observation is indicated by a star in Figure 11.4. We see that it falls within the estimated .50 probability contour of population π_2 and about on the estimated .95 probability contour of population π_1. Thus, the classification is not clear cut.

Suppose now that the prior probabilities of group membership are known. For example, suppose the blood yielding the foregoing x_1 and x_2 measurements is drawn from the maternal first cousin of a hemophiliac. Then the genetic chance of being a hemophilia A carrier in this case is .25. Consequently, the prior probabilities of group membership are $p_1 = .75$ and $p_2 = .25$. Assuming, somewhat unrealistically, that the costs of misclassification are equal, so that $c(1|2) = c(2|1)$, and using the classification statistic

$$\hat{w} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}_{\text{pooled}}\mathbf{x}_0 - \tfrac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}_{\text{pooled}}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$

or $\hat{w} = \hat{\mathbf{a}}'\mathbf{x}_0 - \hat{m}$ with $\mathbf{x}'_0 = [-.210, -.044]$, $\hat{m} = -4.61$, and $\hat{\mathbf{a}}'\mathbf{x}_0 = -6.62$, we have

$$\hat{w} = -6.62 - (-4.61) = -2.01$$

Applying (11-18), we see that

$$\hat{w} = -2.01 < \ln\left[\frac{p_2}{p_1}\right] = \ln\left[\frac{.25}{.75}\right] = -1.10$$

and we classify the woman as π_2, an obligatory carrier. ■

Scaling

- The coefficient vector $\hat{\mathbf{a}} = \mathbf{S}_{pooled}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ is unique only up to a multiplicative constant, so for $c \neq 0$, any vector $c\hat{\mathbf{a}}$ will also serve as discriminant coefficients.

- The vector $\hat{\mathbf{a}}$ is frequently "scaled" or "normalized" to ease the interpretation of its elements.

- (1) Set $\hat{\mathbf{a}}^* = \hat{\mathbf{a}}/\sqrt{\hat{\mathbf{a}}'\hat{\mathbf{a}}}$.
  (2) Set $\hat{\mathbf{a}}^* = \hat{\mathbf{a}}/\hat{a}_1$.

Fisher's Approach to Classification with Two populations

- Fisher's idea was to transform the multivariate observations **x** to univariate observation y such that y's derived from populations π_1 and π_2 were separated as much as possible.

- Fisher suggested taking linear combinations of **x** to create y's because they are simple enough functions of the **x** to be handled easily.

- Fisher's approach does not assume that the populations are normal, but implicitly assume that the population covariance matrices are equal.

Result 6-2.3. The linear combinations $\hat{y} = \hat{\mathbf{a}}'\mathbf{x} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{pooled}^{-1}\mathbf{x}$ maximize the ratio

$$\frac{\begin{pmatrix}\text{squared distance}\\ \text{between sample mean of } y\end{pmatrix}}{(\text{sample variance of } y)} = \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2}$$

$$= \frac{(\hat{\mathbf{a}}'\bar{\mathbf{x}}_1 - \hat{\mathbf{a}}'\bar{\mathbf{x}}_2)^2}{\hat{\mathbf{a}}'\mathbf{S}_{pooled}\hat{\mathbf{a}}}$$

$$= \frac{(\hat{\mathbf{a}}'\mathbf{d})^2}{\hat{\mathbf{a}}'\mathbf{S}_{pooled}\hat{\mathbf{a}}}$$

over all possible coefficient vectors $\hat{\mathbf{a}}$ where $\mathbf{d} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$. The maximum of the ratio is

$$D^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{pooled}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2).$$

Example 11.4 (Fisher's linear discriminant for the hemophilia data) Consider the detection of hemophilia A carriers introduced in Example 11.3. Recall that the equal costs and equal priors linear discriminant function was

$$\hat{y} = \hat{\mathbf{a}}'\mathbf{x} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x} = 37.61x_1 - 28.92x_2$$

This linear discriminant function is Fisher's linear function, which maximally separates the two populations, and the maximum separation in the samples is

$$\begin{aligned} D^2 &= (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \\ &= [.2418, \quad -.0652]\begin{bmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{bmatrix}\begin{bmatrix} .2418 \\ -.0652 \end{bmatrix} \\ &= 10.98 \end{aligned}$$

■

An Allocation Rule Based on Fisher's Discriminant Function

Allocate $\mathbf{x}_0$ to π_1 if

$$\hat{y}_0 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\mathbf{S}_{pooled}^{-1}\mathbf{x}_0 \geq \hat{m} = \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\mathbf{S}_{pooled}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$

or

$$\hat{y}_0 - \hat{m} \geq 0.$$

Allocate $\mathbf{x}_0$ to π_2 if

$$\hat{y}_0 < \hat{m} \quad \text{or} \quad \hat{y}_0 - \hat{m} < 0.$$

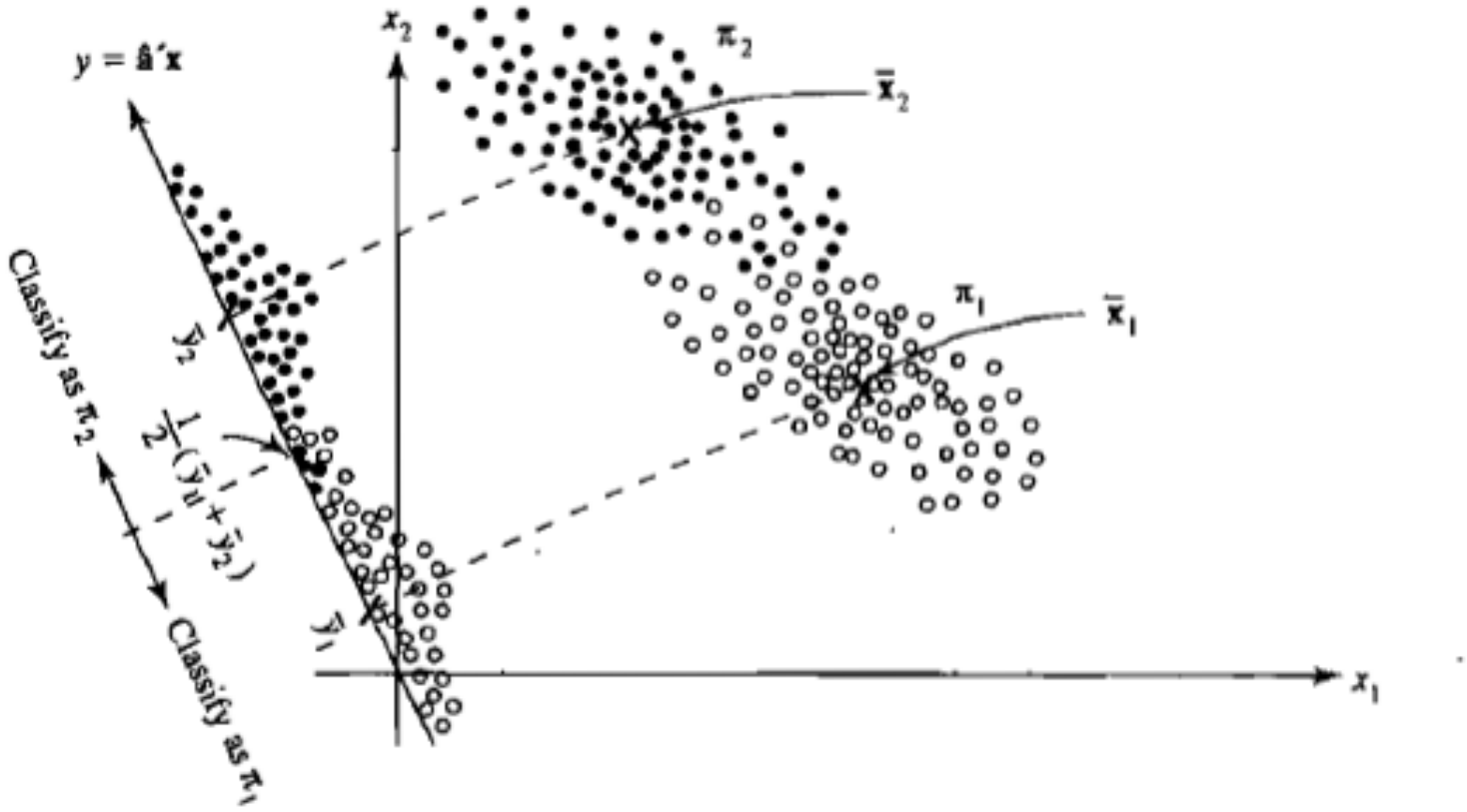

Figure 11.5 A pictorial representation of Fisher's procedure for two populations with $p = 2$.

Classification of Normal Population When $\Sigma_1 \neq \Sigma_2$

Result 6-2.4. Let the populations π_1 and π_2 be described by multivariate normal densities with mean vectors and covariance matrices $\boldsymbol{\mu}_1$, Σ_1 and $\boldsymbol{\mu}_2$, Σ_2, respectively. The allocation rule that minimizes the expected cost of misclassification is given by
Allocate $\mathbf{x}_0$ to π_1 if

$$-\frac{1}{2}\mathbf{x}_0'(\Sigma_1^{-1} - \Sigma_2^{-1})\mathbf{x}_0 + (\boldsymbol{\mu}_1'\Sigma_1^{-1} - \boldsymbol{\mu}_2'\Sigma_2^{-1})\mathbf{x}_0 - k \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right],$$

where $k = \frac{1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + \frac{1}{2}(\boldsymbol{\mu}_1'\Sigma_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2'\Sigma_2^{-1}\boldsymbol{\mu}_2)$. Allocate $\mathbf{x}_0$ to π_2 otherwise.

Quadratic Classification Rule
(Normal Populations with Unequal Covariance Matrices)

Allocate $\mathbf{x}_0$ to π_1 if

$$-\frac{1}{2}\mathbf{x}_0'(\mathbf{S}_1^{-1}-\mathbf{S}_2^{-1})\mathbf{x}_0+(\bar{\mathbf{x}}_1'\mathbf{S}_1^{-1}-\bar{\mathbf{x}}_2'\mathbf{S}_2^{-1})\mathbf{x}_0-k\geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right].$$

Allocate $\mathbf{x}_0$ to π_2 otherwise.

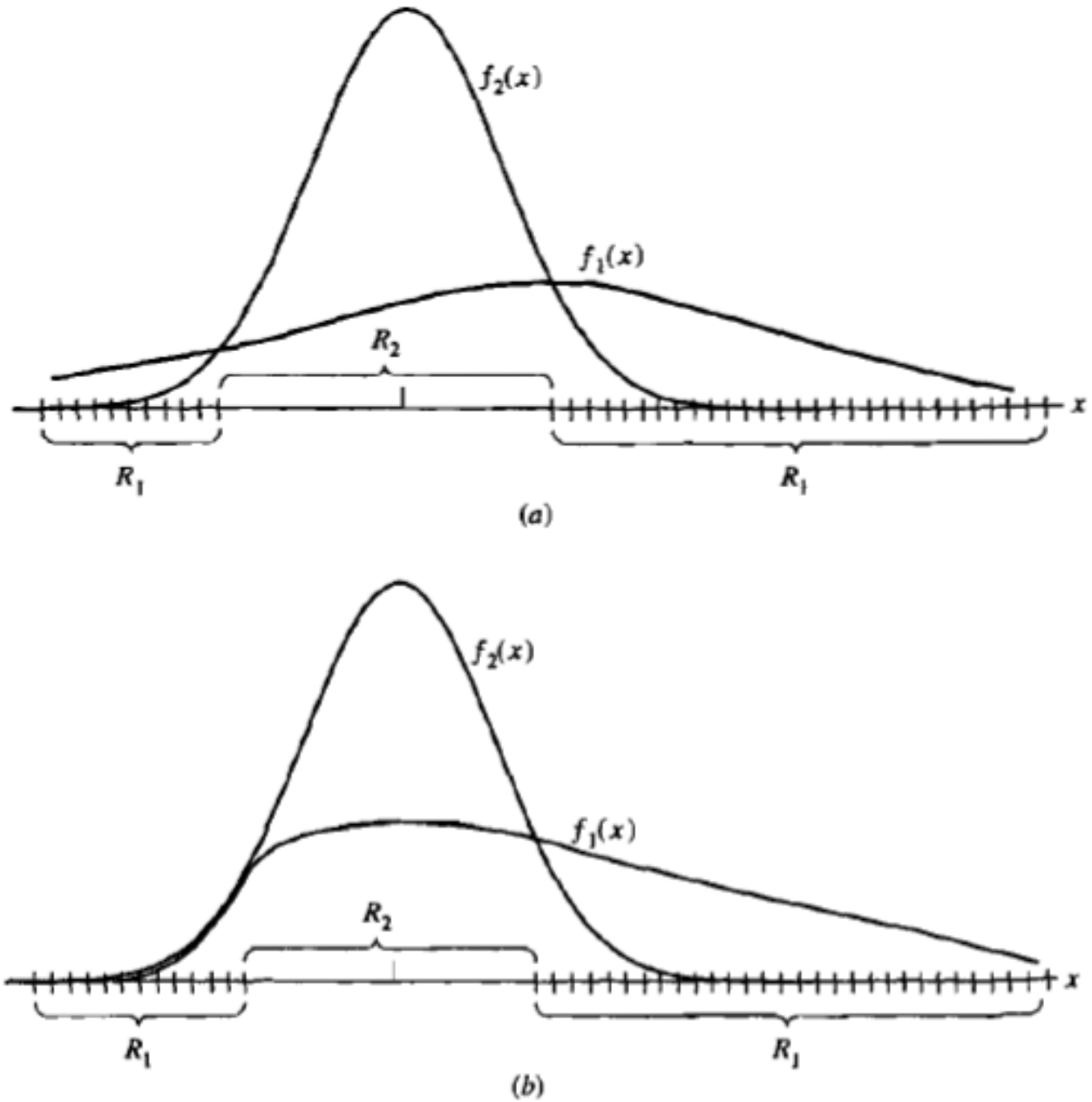

Figure 11.6 Quadratic rules for (a) two normal distribution with unequal variances and (b) two distributions, one of which is nonnormal—rule not appropriate.

Evaluating Classification Functions

- The total probability of misclassification

$$\text{TPM} = p_1 \int_{R_2} f_1(\mathbf{x}) d\mathbf{x} + p_2 \int_{R_1} f_2(\mathbf{x}) d\mathbf{x}$$

- *Optimum error rate* (OER): the smallest value of TPM, obtained by a judicious choice of R_1 and R_2.

- R_1 and R_2 for OER are determined by the rule of Minimum Expected Cost Regions with equal misclassification costs.

Example 11.5 (Calculating misclassification probabilities) Let us derive an expression for the optimum error rate when $p_1 = p_2 = \frac{1}{2}$ and $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are the multivariate normal densities in (11-10).

Now, the minimum ECM and minimum TPM classification rules coincide when $c(1|2) = c(2|1)$. Because the prior probabilities are also equal, the minimum TPM classification regions are defined for normal populations by (11-12), with $\ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] = 0$. We find that

$$R_1: \quad (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \tfrac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq 0$$

$$R_2: \quad (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \tfrac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) < 0$$

These sets can be expressed in terms of $y = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x} = \mathbf{a}'\mathbf{x}$ as

$$R_1(y): \quad y \geq \tfrac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$$

$$R_2(y): \quad y < \tfrac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$$

But Y is a linear combination of normal random variables, so the probability densities of Y, $f_1(y)$ and $f_2(y)$, are univariate normal (see Result 4.2) with means and a variance given by

$$\mu_{1Y} = \mathbf{a}'\boldsymbol{\mu}_1 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1$$

$$\mu_{2Y} = \mathbf{a}'\boldsymbol{\mu}_2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2$$

$$\sigma_Y^2 = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \Delta^2$$

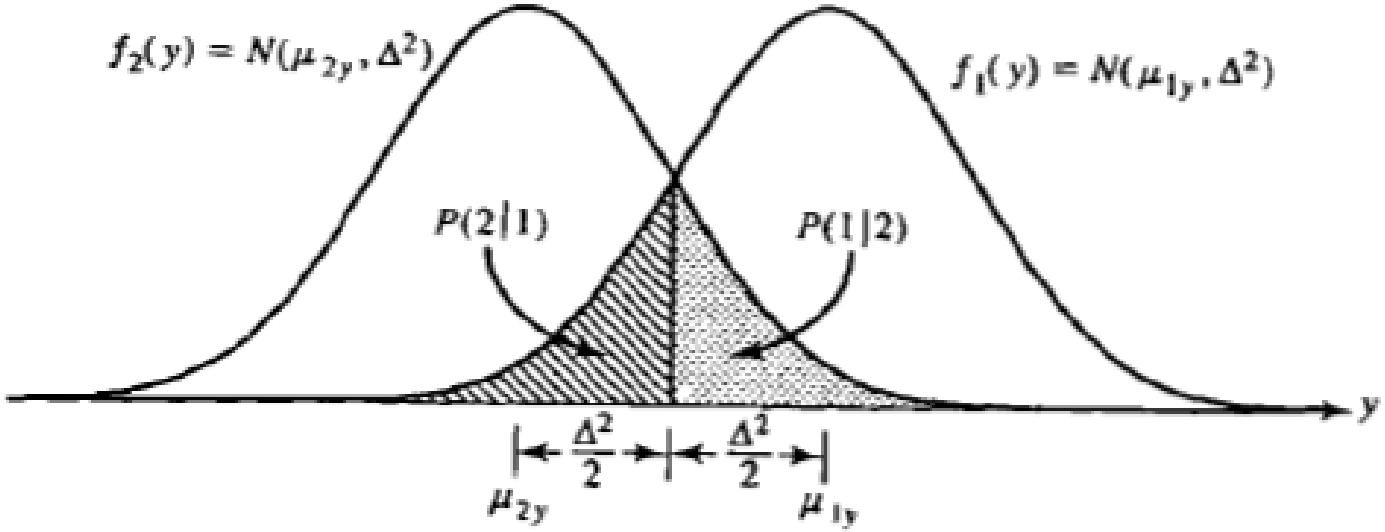

Figure 11.7 The misclassification probabilities based on Y.

Now,

$$\text{TPM} = \tfrac{1}{2}P[\text{misclassifying a } \pi_1 \text{ observation as } \pi_2] + \tfrac{1}{2}P[\text{misclassifying a } \pi_2 \text{ observation as } \pi_1]$$

But, as shown in Figure 11.7

$$\begin{aligned} P[\text{misclassifying a } \pi_1 \text{ observation as } \pi_2] &= P(2|1) \\ &= P[Y < \tfrac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)] \\ &= P\left(\frac{Y - \mu_{1Y}}{\sigma_Y} < \frac{\tfrac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1}{\Delta}\right) \\ &= P\left(Z < \frac{-\frac{1}{2}\Delta^2}{\Delta}\right) = \Phi\left(\frac{-\Delta}{2}\right) \end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable. Similarly,

$$\begin{aligned} P[\text{misclassifying a } \pi_2 \text{ observation as } \pi_1] &= P(1|2) = P[Y \geq \tfrac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)] \\ &= P\left(Z \geq \frac{\Delta}{2}\right) = 1 - \Phi\left(\frac{\Delta}{2}\right) = \Phi\left(\frac{-\Delta}{2}\right) \end{aligned}$$

Therefore, the optimum error rate is

$$\text{OER} = \text{minimum TPM} = \frac{1}{2}\Phi\left(\frac{-\Delta}{2}\right) + \frac{1}{2}\Phi\left(\frac{-\Delta}{2}\right) = \Phi\left(\frac{-\Delta}{2}\right) \qquad (11\text{-}31)$$

If, for example, $\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = 2.56$, then $\Delta = \sqrt{2.56} = 1.6$, and, using Table 1 in the appendix, we obtain

$$\text{Minimum TPM} = \Phi\left(\frac{-1.6}{2}\right) = \Phi(-.8) = .2119$$

The optimal classification rule here will incorrectly allocate about 21% of the items to one population or the other. ■

- **Actual error rate** (AER):

$$\text{AER} = p_1 \int_{\hat{R}_2} f_1(\mathbf{x}) d\mathbf{x} + p_2 \int_{\hat{R}_1} f_2(\mathbf{x}) d\mathbf{x}$$

- **Apparent error rate** (APER): the fraction of observations in the training sample that are misclassified by the sample classification function.

The apparent error rate can be easily calculated from the *confusion matrix*, which shows actual versus predicted group membership. For n_1 observations from π_1 and n_2 observations from π_2, the confusion matrix has the form

		Predicted membership π_1	Predicted membership π_2		
Actual membership	π_1	n_{1C}	$n_{1M} = n_1 - n_{1C}$	n_1	(11-33)
	π_2	$n_{2M} = n_2 - n_{2C}$	n_{2C}	n_2	

where

n_{1C} = number of π_1 items correctly classified as π_1 items
n_{1M} = number of π_1 items misclassified as π_2 items
n_{2C} = number of π_2 items correctly classified
n_{2M} = number of π_2 items misclassified

The apparent error rate is then

$$\text{APER} = \frac{n_{1M} + n_{2M}}{n_1 + n_2} \tag{11-34}$$

which is recognized as the *proportion* of items in the training set that are misclassified.

Example 11.6 (Calculating the apparent error rate) Consider the classification regions R_1 and R_2 shown in Figure 11.1 for the riding-mower data. In this case, observations northeast of the solid line are classified as π_1, mower owners; observations southwest of the solid line are classified as π_2, nonowners. Notice that some observations are misclassified. The confusion matrix is

		Predicted membership: π_1: riding-mower owners	Predicted membership: π_2: nonowners	
Actual membership	π_1: riding-mower owners	$n_{1C} = 10$	$n_{1M} = 2$	$n_1 = 12$
	π_2: nonowners	$n_{2M} = 2$	$n_{2C} = 10$	$n_2 = 12$

The apparent error rate, expressed as a percentage, is

$$\text{APER} = \left(\frac{2 + 2}{12 + 12}\right)100\% = \left(\frac{4}{24}\right)100\% = 16.7\%$$

■

- APER tends to underestimate the AER, and the problem does not disappear unless the sample sizes n_1 and n_2 are very large.

- Essentially, this optimistic estimate occurs because the data used to build the classification function are also used to evaluate it.

- Error-rate estimates can be constructed that are better than the apparent error rate, remain relatively easy to calculate, and do not require distributional assumption.
  - Split the total sample into training sample and a validation sample. Shortcoming: 1. Requires large samples, 2. valuable information may be lost.
  - Lachenbruch's "holdout" procedure.

Lachenbruch's "holdout" procedure

1. Start with the π_1 group of observations. Omit one observation from this group, and develop a classification function based on the remaining $n_1 - 1, n_2$ observations.

2. Classify the "holdout" observation, using the function constructed in Step 1.

3. Repeat Step 1 and 2 until all of the π_1 observations are classified, Let $n_{1M}^{(H)}$ be the number of holdout (H) observations misclassified in this group.

4. Repeat Step 1 through 3 for the π_2 observations, Let $n_{2M}^{(H)}$ be the number of holdout observations misclassified in this group.

$$\hat{P}(2|1) = \frac{n_{1M}^{(H)}}{n_1}, \quad \hat{P}(1|2) = \frac{n_{2M}^{(H)}}{n_2}$$

and

$$\hat{\text{E}}(\text{AER}) = \frac{n_{1M}^{(H)} + n_{2M}^{(H)}}{n_1 + n_2}$$

Example 11.7 Calculating an estimate of the error rate using the holdout procedure) We shall illustrate Lachenbruch's holdout procedure and the calculation of error rate estimates for the equal costs and equal priors version of (11-18). Consider the following data matrices and descriptive statistics. (We shall assume that the $n_1 = n_2 = 3$ bivariate observations were selected randomly from two populations π_1 and π_2 with a common covariance matrix.)

$$\mathbf{X}_1 = \begin{bmatrix} 2 & 12 \\ 4 & 10 \\ 3 & 8 \end{bmatrix}; \quad \bar{\mathbf{x}}_1 = \begin{bmatrix} 3 \\ 10 \end{bmatrix}, \quad 2\mathbf{S}_1 = \begin{bmatrix} 2 & -2 \\ -2 & 8 \end{bmatrix}$$

$$\mathbf{X}_2 = \begin{bmatrix} 5 & 7 \\ 3 & 9 \\ 4 & 5 \end{bmatrix}; \quad \bar{\mathbf{x}}_2 = \begin{bmatrix} 4 \\ 7 \end{bmatrix}, \quad 2\mathbf{S}_2 = \begin{bmatrix} 2 & -2 \\ -2 & 8 \end{bmatrix} .$$

The pooled covariance matrix is

$$\mathbf{S}_{\text{pooled}} = \frac{1}{4}(2\mathbf{S}_1 + 2\mathbf{S}_2) = \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix}$$

Using $\mathbf{S}_{\text{pooled}}$, the rest of the data, and Rule (11-18) with equal costs and equal priors, we may classify the sample observations. You may then verify (see Exercise 11.19) that the confusion matrix is

		Classify as:	
		π_1	π_2
True population:	π_1	2	1
	π_2	1	2

and consequently,

$$\text{APER(apparent error rate)} = \frac{2}{6} = .33$$

Holding out the first observation $\mathbf{x}_H' = [2, 12]$ from $\mathbf{X}_1$, we calculate

$$\mathbf{X}_{1H} = \begin{bmatrix} 4 & 10 \\ 3 & 8 \end{bmatrix}; \quad \bar{\mathbf{x}}_{1H} = \begin{bmatrix} 3.5 \\ 9 \end{bmatrix}; \quad \text{and} \quad 1\mathbf{S}_{1H} = \begin{bmatrix} .5 & 1 \\ 1 & 2 \end{bmatrix}$$

The new pooled covariance matrix, $\mathbf{S}_{H,\text{pooled}}$, is

$$\mathbf{S}_{H,\text{pooled}} = \frac{1}{3}[1\mathbf{S}_{1H} + 2\mathbf{S}_2] = \frac{1}{3}\begin{bmatrix} 2.5 & -1 \\ -1 & 10 \end{bmatrix}$$

with inverse[8]

$$\mathbf{S}_{H,\text{pooled}}^{-1} = \frac{1}{8}\begin{bmatrix} 10 & 1 \\ 1 & 2.5 \end{bmatrix}$$

It is computationally quicker to classify the holdout observation $\mathbf{x}_{1H}$ on the basis of its squared distances from the group means $\bar{\mathbf{x}}_{1H}$ and $\bar{\mathbf{x}}_2$. This procedure is equivalent to computing the value of the linear function $\hat{y} = \hat{\mathbf{a}}_H'\mathbf{x}_H = (\bar{\mathbf{x}}_{1H} - \bar{\mathbf{x}}_2)'\mathbf{S}_{H,\text{pooled}}^{-1}\mathbf{x}_H$ and comparing it to the midpoint $\hat{m}_H = \frac{1}{2}(\bar{\mathbf{x}}_{1H} - \bar{\mathbf{x}}_2)'\mathbf{S}_{H,\text{pooled}}^{-1}(\bar{\mathbf{x}}_{1H} + \bar{\mathbf{x}}_2)$. [See (11-19) and (11-20).]

Thus with $\mathbf{x}_H' = [2, 12]$ we have

$$\text{Squared distance from } \bar{\mathbf{x}}_{1H} = (\mathbf{x}_H - \bar{\mathbf{x}}_{1H})' \mathbf{S}^{-1}_{H,\text{pooled}} (\mathbf{x}_H - \bar{\mathbf{x}}_{1H})$$

$$= [2 - 3.5 \quad 12 - 9] \frac{1}{8} \begin{bmatrix} 10 & 1 \\ 1 & 2.5 \end{bmatrix} \begin{bmatrix} 2 & -3.5 \\ 12 & -9 \end{bmatrix} = 4.5$$

$$\text{Squared distance from } \bar{\mathbf{x}}_2 = (\mathbf{x}_H - \bar{\mathbf{x}}_2)' \mathbf{S}^{-1}_{H,\text{pooled}} (\mathbf{x}_H - \bar{\mathbf{x}}_2)$$

$$= [2 - 4 \quad 12 - 7] \frac{1}{8} \begin{bmatrix} 10 & 1 \\ 1 & 2.5 \end{bmatrix} \begin{bmatrix} 2 & -4 \\ 12 & -7 \end{bmatrix} = 10.3$$

Since the distance from $\mathbf{x}_H$ to $\bar{\mathbf{x}}_{1H}$ is smaller than the distance from $\mathbf{x}_H$ to $\bar{\mathbf{x}}_2$, we classify $\mathbf{x}_H$ as a π_1 observation. In this case, the classification is correct.

If $\mathbf{x}_H' = [4, 10]$ is withheld, $\bar{\mathbf{x}}_{1H}$ and $\mathbf{S}^{-1}_{H,\text{pooled}}$ become

$$\bar{\mathbf{x}}_{1H} = \begin{bmatrix} 2.5 \\ 10 \end{bmatrix} \quad \text{and} \quad \mathbf{S}^{-1}_{H,\text{pooled}} = \frac{1}{8} \begin{bmatrix} 16 & 4 \\ 4 & 2.5 \end{bmatrix}$$

We find that

$$(\mathbf{x}_H - \bar{\mathbf{x}}_{1H})'\mathbf{S}_{H,\text{pooled}}^{-1}(\mathbf{x}_H - \bar{\mathbf{x}}_{1H}) = [4 - 2.5 \quad 10 - 10]\frac{1}{8}\begin{bmatrix}16 & 4\\ 4 & 2.5\end{bmatrix}\begin{bmatrix}4 - 2.5\\ 10 - 10\end{bmatrix}$$
$$= 4.5$$
$$(\mathbf{x}_H - \bar{\mathbf{x}}_2)'\mathbf{S}_{H,\text{pooled}}^{-1}(\mathbf{x}_H - \bar{\mathbf{x}}_2) = [4 - 4 \quad 10 - 7]\frac{1}{8}\begin{bmatrix}16 & 4\\ 4 & 2.5\end{bmatrix}\begin{bmatrix}4 - 4\\ 10 - 7\end{bmatrix}$$
$$= 2.8$$

and consequently, we would incorrectly assign $\mathbf{x}_H' = [4, 10]$ to π_2. Holding out $\mathbf{x}_H' = [3, 8]$ leads to incorrectly assigning this observation to π_2 as well. Thus, $n_{1M}^{(H)} = 2$.

Turning to the second group, suppose $\mathbf{x}_H' = [5, 7]$ is withheld. Then

$$\mathbf{X}_{2H} = \begin{bmatrix}3 & 9\\ 4 & 5\end{bmatrix}; \quad \bar{\mathbf{x}}_{2H} = \begin{bmatrix}3.5\\ 7\end{bmatrix}; \text{ and } \quad 1\mathbf{S}_{2H} = \begin{bmatrix}.5 & -2\\ -2 & 8\end{bmatrix}$$

The new pooled covariance matrix is

$$\mathbf{S}_{H,\text{pooled}} = \frac{1}{3}[2\mathbf{S}_1 + 1\mathbf{S}_{2H}] = \frac{1}{3}\begin{bmatrix}2.5 & -4\\ -4 & 16\end{bmatrix}$$

with inverse

$$\mathbf{S}_{H,\text{pooled}}^{-1} = \frac{3}{24}\begin{bmatrix}16 & 4\\ 4 & 2.5\end{bmatrix}$$

We find that

$$(\mathbf{x}_H - \bar{\mathbf{x}}_1)'\mathbf{S}_{H,\text{pooled}}^{-1}(\mathbf{x}_H - \bar{\mathbf{x}}_1) = [5 - 3 \quad 7 - 10]\frac{3}{24}\begin{bmatrix}16 & 4\\ 4 & 2.5\end{bmatrix}\begin{bmatrix}5 - 3\\ 7 - 10\end{bmatrix}$$
$$= 4.8$$
$$(\mathbf{x}_H - \bar{\mathbf{x}}_{2H})'\mathbf{S}_{H,\text{pooled}}^{-1}(\mathbf{x}_H - \bar{\mathbf{x}}_{2H}) = [5 - 3.5 \quad 7 - 7]\frac{3}{24}\begin{bmatrix}16 & 4\\ 4 & 2.5\end{bmatrix}\begin{bmatrix}5 - 3.5\\ 7 - 7\end{bmatrix}$$
$$= 4.5$$

and $\mathbf{x}_H' = [5, 7]$ is correctly assigned to π_2.

When $\mathbf{x}_H' = [3, 9]$ is withheld,

$$(\mathbf{x}_H - \bar{\mathbf{x}}_1)'\mathbf{S}_{H,\text{pooled}}^{-1}(\mathbf{x}_H - \bar{\mathbf{x}}_1) = [3 - 3 \quad 9 - 10]\frac{3}{24}\begin{bmatrix}10 & 1\\ 1 & 2.5\end{bmatrix}\begin{bmatrix}3 - 3\\ 9 - 10\end{bmatrix}$$
$$= .3$$
$$(\mathbf{x}_H - \bar{\mathbf{x}}_{2H})'\mathbf{S}_{H,\text{pooled}}^{-1}(\mathbf{x}_H - \bar{\mathbf{x}}_{2H}) = [3 - 4.5 \quad 9 - 6]\frac{3}{24}\begin{bmatrix}10 & 1\\ 1 & 2.5\end{bmatrix}\begin{bmatrix}3 - 4.5\\ 9 - 6\end{bmatrix}$$
$$= 4.5$$

and $\mathbf{x}_H' = [3, 9]$ is incorrectly assigned to π_1. Finally, withholding $\mathbf{x}_H' = [4, 5]$ leads to correctly classifying this observation as π_2. Thus, $n_{2M}^{(H)} = 1$.

An estimate of the expected actual error rate is provided by

$$\hat{E}(\text{AER}) = \frac{n_{1M}^{(H)} + n_{2M}^{(H)}}{n_1 + n_2} = \frac{2 + 1}{3 + 3} = .5$$

Hence, we see that the apparent error rate APER = .33 is an optimistic measure of performance. Of course, in practice, sample sizes are larger than those we have considered here, and the difference between APER and $\hat{E}(\text{AER})$ may not be as large. ■

If you are interested in pursuing the approaches to estimating classification error rates, see [23].

The next example illustrates a difficulty that can arise when the variance of the discriminant is not the same for both populations.

Example 11.8 (Classifying Alaskan and Canadian salmon) The salmon fishery is a valuable resource for both the United States and Canada. Because it is a limited resource, it must be managed efficiently. Moreover, since more than one country is involved, problems must be solved equitably. That is, Alaskan commercial fishermen cannot catch too many Canadian salmon and vice versa.

These fish have a remarkable life cycle. They are born in freshwater streams and after a year or two swim into the ocean. After a couple of years in salt water, they return to their place of birth to spawn and die. At the time they are about to return as mature fish, they are harvested while still in the ocean. To help regulate catches, samples of fish taken during the harvest must be identified as coming from Alaskan or Canadian waters. The fish carry some information about their birthplace in the growth rings on their scales. Typically, the rings associated with freshwater growth are smaller for the Alaskan-born than for the Canadian-born salmon. Table 11.2 gives the diameters of the growth ring regions, magnified 100 times, where

X_1 = diameter of rings for the first-year freshwater growth (hundredths of an inch)

X_2 = diameter of rings for the first-year marine growth (hundredths of an inch)

In addition, females are coded as 1 and males are coded as 2.

Training samples of sizes $n_1 = 50$ Alaskan-born and $n_2 = 50$ Canadian-born salmon yield the summary statistics

$$\bar{\mathbf{x}}_1 = \begin{bmatrix}98.380\\ 429.660\end{bmatrix}, \quad \mathbf{S}_1 = \begin{bmatrix}260.608 & -188.093\\ -188.093 & 1399.086\end{bmatrix}$$

$$\bar{\mathbf{x}}_2 = \begin{bmatrix}137.460\\ 366.620\end{bmatrix}, \quad \mathbf{S}_2 = \begin{bmatrix}326.090 & 133.505\\ 133.505 & 893.261\end{bmatrix}$$

Table 11.2 Salmon Data (Growth-Ring Diameters)

Alaskan			Canadian		
Gender	Freshwater	Marine	Gender	Freshwater	Marine
2	108	368	1	129	420
1	131	355	1	148	371
1	105	469	1	179	407
2	86	506	2	152	381
1	99	402	2	166	377
2	87	423	2	124	389
1	94	440	1	156	419
2	117	489	2	131	345
2	79	432	1	140	362
1	99	403	2	144	345
1	114	428	2	149	393
2	123	372	1	108	330
1	123	372	1	135	355
2	109	420	2	170	386
2	112	394	1	152	301
1	104	407	1	153	397
2	111	422	1	152	301
2	126	423	2	136	438
2	105	434	2	122	306
1	119	474	1	148	383
1	114	396	2	90	385
2	100	470	1	145	337
2	84	399	1	123	364
2	102	429	2	145	376
2	101	469	2	115	354
2	85	444	2	134	383
1	109	397	1	117	355
2	106	442	2	126	345
1	82	431	1	118	379
2	118	381	2	120	369
1	105	388	1	153	403
1	121	403	2	150	354
1	85	451	1	154	390
1	83	453	1	155	349
1	53	427	2	109	325
1	95	411	2	117	344
1	76	442	1	128	400
1	95	426	1	144	403
2	87	402	2	163	370
1	70	397	2	145	355
2	84	511	1	133	375
2	91	469	1	128	383
1	74	451	2	123	349
2	101	474	1	144	373
1	80	398	2	140	388

(continues on next page)

Table 11.2 *(continued)*

Alaskan			Canadian		
Gender	Freshwater	Marine	Gender	Freshwater	Marine
1	95	433	2	150	339
2	92	404	2	124	341
1	99	481	1	125	346
2	94	491	1	153	352
1	87	480	1	108	339

Gender Key: 1 = female; 2 = male.
Source: Data courtesy of K. A. Jensen and B. Van Alen of the State of Alaska Department of Fish and Game.

The data appear to satisfy the assumption of bivariate normal distributions (see Exercise 11.31), but the covariance matrices may differ. However, to illustrate a point concerning misclassification probabilities, we will use the linear classification procedure.

The classification procedure, using equal costs and equal prior probabilities, yields the holdout estimated error rates

		Predicted membership	
		π_1: Alaskan	π_2: Canadian
Actual membership	π_1: Alaskan	44	6
	π_2: Canadian	1	49

based on the linear classification function [see (11-19) and (11-20)]

$$\hat{w} = \hat{y} - \hat{m} = -5.54121 - .12839x_1 + .05194x_2$$

There is some difference in the sample standard deviations of $\hat{w}$ for the two populations:

	n	Sample Mean	Sample Standard Deviation
Alaskan	50	4.144	3.253
Canadian	50	−4.147	2.450

Although the overall error rate (7/100, or 7%) is quite low, there is an unfairness here. It is less likely that a Canadian-born salmon will be misclassified as Alaskan born, rather than vice versa. Figure 11.8, which shows the two normal densities for the linear discriminant $\hat{y}$, explains this phenomenon. Use of the

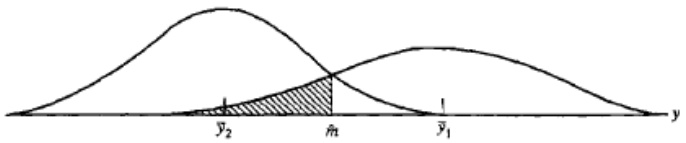

Figure 11.8 Schematic of normal densities for linear discriminant—salmon data.

midpoint between the two sample means does not make the two misclassification probabilities equal. It clearly penalizes the population with the largest variance. Thus, blind adherence to the linear classification procedure can be unwise. ■

Classification with Serval Populations

The Minimum Expected Cost of Misclassification Method

Let $f_i(\mathbf{x})$ be the density associated with population $\pi_i, i = 1, 2, \ldots, g$. Let p_i be the prior probability of population $\pi_i, i = 1, \ldots, g$, and $c(k|i)$ be the cost of allocating an item to π_k when, in fact, it belongs to π_i, for $k, i = 1, \ldots, g$. For $k = i$, $c(i|i) = 0$. Finally, let R_k be the set of $\mathbf{x}'s$ classed as π_k and

$$P(k|i) = \int_{R_k} f_i(\mathbf{x}) d\mathbf{x}$$

for $k, i = 1, 2, \ldots, g$ with $P(i|i) = 1 - \sum_{k=1, k \neq i}^{g} P(k|i)$.

- The conditional expected cost of misclassifying an $\mathbf{x}$ from π_1 int π_2, or $\pi_3, \ldots, \pi_g$ is

$$\text{ECM}(1) = P(2|1)c(2|1) + \ldots + P(g|1)c(g|1) = \sum_{k=2}^{g} P(k|1)c(k|1).$$

- Multiplying each conditional ECM by its prior probability and summing gives the overall ECM:

$$\begin{aligned} \text{ECM} &= p_1\text{ECM}(1) + \cdots + p_g\text{ECM}(g) \\ &= \sum_{i=1}^{g} p_i \left(\sum_{k=1, k\neq i}^{g} P(k|i)c(k|i) \right). \end{aligned}$$

- **Result 6-2.5** The classification regions that minimize the above ECM are defined by allocating

$$\sum_{i=1, i\neq k}^{g} p_i f_i(\mathbf{x}) c(k|i)$$

is smallest. If a tie occurs, $\mathbf{x}$ can be assigned to any of the tied populations.

If $c(i|k)$ for any $i \neq k$ are same, allocate $\mathbf{x}$ to π_k if

$$p_k f_k(\mathbf{x}) > p_i f_i(\mathbf{x}), \quad \text{for all } i \neq k.$$

Classification with Normal Populations

- When the

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)'\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right], i = 1, 2, \ldots, g,$$

If further the misclassification costs are all equals, $c(k|i) = 1, k \neq i$, then

Allocate $\mathbf{x}$ to π_k if

$$\ln p_k f_k(\mathbf{x}) = \ln p_k - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)'\Sigma_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) = \max_i \ln p_i f_i(\mathbf{x}).$$

- Define the sample quadratic discrimination score $\hat{d}_i^Q(\mathbf{x})$ as

$$\hat{d}_i^Q(\mathbf{x}) = -\frac{1}{2}\ln|S_i| - \frac{1}{2}(\mathbf{x} - \bar{(}\mathbf{x})_i)'\Sigma_k^{-1}(\mathbf{x} - \bar{(}\mathbf{x})_i) + \ln p_i, i = 1, 2, \ldots, g.$$

Then allocate $\mathbf{x}$ to π_k if the quadratic score $\hat{d}_k^Q(\mathbf{x})$ is the largest of $\hat{d}_1^Q(\mathbf{x}), \ldots, \hat{d}_g^Q(\mathbf{x})$.

- If the population covariance matrices Σ_i are equal, then define the sample linear discrimination score as

$$\hat{d}_i(\mathbf{x}) = \bar{\mathbf{x}}_i \mathbf{S}_{pooled}^{-1} \mathbf{x} - \frac{1}{2} \mathbf{S}_{pooled}^{-1} \bar{\mathbf{x}}_i + \ln p_i, \text{for } i = 1, 2, \ldots, g.$$

  Then, allocate $\mathbf{x}$ to π_i if the linear discrimination score $\hat{d}_k(\mathbf{x})$ is the a largest of $\hat{d}_1(\mathbf{x}), \ldots, \hat{d}_g(\mathbf{x})$. where

$$\mathbf{S}_{pooled} = \frac{1}{n_1 + n_2 + \ldots + n_g - g}((n_1 -)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2 + \cdots + (n_g - 1)\mathbf{S}_g).$$

Fisher's Method for Discrimiating among Several Populations

- The motivation behind the Fisher discriminant analysis is the need to obtain a reasonable representation of the populations that involves only a few linear combinations of the observations, such as $\mathbf{a}_1'\mathbf{x}, \mathbf{a}_2'\mathbf{x}$ and $\mathbf{a}_3'\mathbf{x}$.

- The approach has several advantages when one is interested in separating several population for (1) visual inspection or (2) graphical descriptive purposes.

- Assume that $p \times p$ population covariance matrices are equal and of full rank. That is $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_g = \Sigma$.

- Define

$$\bar{\mathbf{x}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{x}_{ij}, i = 1, \ldots, g, \text{and } \bar{\mathbf{x}} = \frac{1}{g}\sum_{i=1}^{g}\mathbf{x}_i.$$

$$\mathbf{B} = \sum_{i=1}^{g}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})'$$

and

$$\mathbf{W} = \sum_{i=1}^{g}(n_i - 1)\mathbf{S}_i = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)'.$$

Fisher's Sample Linear Discriminants

Let $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_s > 0$ denote the $s \leq \min(g-1, p)$ nonzero eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$ and $\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_s$ be the corresponding eigenvectors (scaled so that $\hat{\mathbf{e}}'\mathbf{S}_{pooled}\hat{\mathbf{e}} = 1$). Then the vector of coefficients $\hat{\mathbf{a}}$ that maximizes the ratio

$$\frac{\hat{\mathbf{a}}'\mathbf{B}\hat{\mathbf{a}}}{\hat{\mathbf{a}}'\mathbf{W}\hat{\mathbf{a}}} = \frac{\hat{\mathbf{a}}'\left(\sum_{i=1}^{g}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})'\right)\hat{\mathbf{a}}}{\hat{\mathbf{a}}'\left(\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)'\right)\hat{\mathbf{a}}}$$

is given by $\hat{\mathbf{a}}_1 = \hat{\mathbf{e}}_1$. The linear combination $\hat{\mathbf{a}}_1'\mathbf{x}$ is, called the sample first discriminant. The choice $\hat{\mathbf{a}}_2 = \hat{\mathbf{e}}_2$ produces the sample second discriminant, $\hat{\mathbf{a}}_2'\mathbf{x}$, and continuing, we obtain $\hat{\mathbf{a}}_k\mathbf{x} = \hat{\mathbf{e}}_k'\mathbf{x}$, the sample kth discriminant, $k \leq s$.

Let

$$d_i(\mathbf{x}) = \mu_i' \Sigma^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_i' \Sigma^{-1} \boldsymbol{\mu}_i + \ln p_i$$

or, equivalently

$$d_i(\mathbf{x}) - \frac{1}{2} \mathbf{x}' \Sigma^{-1} \mathbf{x} = -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) + \ln p_i.$$

Result 6-2.6. Let $y_j = \mathbf{a}_j'\mathbf{x}$ where $\mathbf{a}_j = \Sigma^{-1/2}\mathbf{e}_j$ and $\mathbf{e}_j$ is an eigenvector of $\Sigma^{-1/2}\mathbf{B}_{\boldsymbol{\mu}}\Sigma^{-1/2}$. Then

$$\begin{aligned}\sum_{j=1}^{p}(y_j - \mu_{iY_j})^2 &= \sum_{j=1}^{p}[\mathbf{a}_j'(\mathbf{x}-\boldsymbol{\mu}_i)]^2 = (\mathbf{x}-\boldsymbol{\mu}_i)'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\\ &= -2d_i(\mathbf{x}) + \mathbf{x}'\Sigma^{-1}\mathbf{x} + 2\ln p_i\end{aligned}$$

If $\lambda_1 \geq \cdot \geq \lambda_s > 0 = \lambda_{s+1} = \cdot = \lambda_p$, $\sum_{j=s+1}^{p}(y_j - \mu_{iY_j})^2$ is constant for all populations, $i = 1, 2, \ldots, g$ so only the first s discriminants y_j, or $\sum_{j=1}^{s}(y_j - \mu_{iY_j})^2$, contribute to the classification.

Fisher's Classification Procedure Based on Sample Discriminations

Allocate $\mathbf{x}$ to π_k if

$$\sum_{j=1}^{r}(\hat{y}_j - \bar{y}_{kj})^2 = \sum_{j=1}^{r}[\hat{\mathbf{a}}_j'(\mathbf{x} - \bar{\mathbf{x}}_k)]^2 \leq \sum_{j=1}^{r}[\hat{\mathbf{a}}_j'(\mathbf{x} - \bar{\mathbf{x}}_j)]^2 \quad mbox forall\ i \neq k$$

where $\hat{\mathbf{a}}_j$ is the corresponding eigenvectors of $\mathbf{W}^{-1}\mathbf{B}$, $\bar{y}_{kj} = \hat{\mathbf{a}}_j'\bar{\mathbf{x}}_k$ and $r \leq s$.