8 Principal Components Analysis

Further reading. The paper Xoplaki et al. (2000) shows a nice application of the principal components and canonical correlation analysis method (see Chapter 9).

8.1 Introduction

Example (Weather Report, p. 1-6)**.** The data set consists of sunshine (three variables), air temperature (six variables), heating degree days (HDD) and precipitation data (five variables). Figure 8.1 shows a scatterplot matrix of several variables. With the help of principal component analysis we want to reduce the number of variables, without loosing a lot of information.

A principal component analysis is concerned with explaining the variance-covariance structure of a set of variables through a few linear combinations of these variables. Its general objectives are (1) data reduction and (2) interpretation.

Although p components are required to reproduce the total system variability, often much of this variability can be accounted for by a small number k of the principal components. If so, there is almost as much information in the k components as there is in the original p variables. The k principal components can then replace the initial p variables, and the original data set, consisting of n measurements on p variables, is reduced to a data set consisting of n measurements on k principal components.

Analyses of principal components are more of a means to an end rather than an end in themselves, because they frequently serve as intermediate steps in much larger investigations.

8.2 Population Principal Components

Algebraically, principal components are particular linear combinations of the p random variables $X_1, \ldots, X_p$. Geometrically, these linear combinations represent the selection of a new coordinate system obtained by rotating the original system with $X_1, \ldots, X_p$ as the coordinate axes. The new axes represent the directions with maximum variability and provide a simpler and more parsimonious description of the covariance structure.

As we shall see, principal components depend solely on the covariance matrix $\boldsymbol{\Sigma}$ (or the correlation matrix $\boldsymbol{\rho}$ of $X_1, \ldots, X_p$). Their development does not require a multivariate normal assumption. On the other hand, principal components derived for multivariate normal populations have useful interpretations in terms of the constant density ellipsoids. Further, inferences can be done from the sample components when the population is multivariate normal.

Let the random vector $\boldsymbol{X}' = (X_1, \ldots, X_p)$ have the covariance matrix $\boldsymbol{\Sigma}$ with eigen-

values $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$. Consider the linear combinations

$$\begin{aligned} Y_1 &= \boldsymbol{a}_1'\boldsymbol{X} = a_{11}X_1 + \ldots + a_{1p}X_p \\ &\vdots \\ Y_p &= \boldsymbol{a}_p'\boldsymbol{X} = a_{p1}X_1 + \ldots + a_{pp}X_p. \end{aligned}$$

Then, we find

$$\mathrm{Var}(Y_i) = \boldsymbol{a}_i'\boldsymbol{\Sigma}\boldsymbol{a}_i, \qquad i = 1, \ldots, p \tag{8.1}$$

$$\mathrm{Cov}(Y_i, Y_k) = \boldsymbol{a}_i'\boldsymbol{\Sigma}\boldsymbol{a}_k, \qquad i, k = 1, \ldots, p \tag{8.2}$$

The principal components are those uncorrelated linear combinations $Y_1, \ldots, Y_p$ whose variances in (8.1) are as large as possible.

Definition 8.2.1. We define

$$\begin{aligned} \text{First principal component} &= \text{linear combination } \boldsymbol{a}_1'\boldsymbol{X} \text{ that maximizes} \\ &\quad \mathrm{Var}(\boldsymbol{a}_1'\boldsymbol{X}) \text{ subject to } \boldsymbol{a}_1'\boldsymbol{a}_1 = 1 \\ \text{Second principal component} &= \text{linear combination } \boldsymbol{a}_2'\boldsymbol{X} \text{ that maximizes} \\ &\quad \mathrm{Var}(\boldsymbol{a}_2'\boldsymbol{X}) \text{ subject to } \boldsymbol{a}_2'\boldsymbol{a}_2 = 1 \text{ and} \\ &\quad \mathrm{Cov}(\boldsymbol{a}_1'\boldsymbol{X}, \boldsymbol{a}_2'\boldsymbol{X}) = 0 \\ &\vdots \\ i\text{th principal component} &= \text{linear combination } \boldsymbol{a}_i'\boldsymbol{X} \text{ that maximizes} \\ &\quad \mathrm{Var}(\boldsymbol{a}_i'\boldsymbol{X}) \text{ subject to } \boldsymbol{a}_i'\boldsymbol{a}_i = 1 \text{ and} \\ &\quad \mathrm{Cov}(\boldsymbol{a}_i'\boldsymbol{X}, \boldsymbol{a}_k'\boldsymbol{X}) = 0 \text{ for } k < i. \end{aligned}$$

Proposition 8.2.2. *Let $\boldsymbol{\Sigma}$ be the covariance matrix associated with the random vector $\boldsymbol{X}' = (X_1, \ldots, X_p)$. Let $\boldsymbol{\Sigma}$ have eigenvalue-eigenvector pairs $(\lambda_1, \boldsymbol{e}_1), \ldots, (\lambda_p, \boldsymbol{e}_p)$ where $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$. Then the ith principal component is given by*

$$Y_i = \boldsymbol{e}_i'\boldsymbol{X} = \sum_{j=1}^{p} e_{ij}X_j, \qquad i = 1, \ldots, p. \tag{8.3}$$

With these choices

$$\begin{aligned} \mathrm{Var}(Y_i) &= \boldsymbol{e}_i'\boldsymbol{\Sigma}\boldsymbol{e}_i = \lambda_i, \qquad i = 1, \ldots, p \\ \mathrm{Cov}(Y_i, Y_k) &= \boldsymbol{e}_i'\boldsymbol{\Sigma}\boldsymbol{e}_k = 0, \qquad i \neq k. \end{aligned}$$

Remark. If some λ_i are equal, the choices of the corresponding coefficient vectors, $\boldsymbol{e}_i$, and hence Y_i, are not unique.

Proposition 8.2.3. *Let* $\boldsymbol{X}' = (X_1, \ldots, X_p)$ *as in Proposition 8.2.2. Then*

$$tr(\boldsymbol{\Sigma}) = \sum_{i=1}^{p} \sigma_{ii} = \sum_{i=1}^{p} \mathrm{Var}(X_i) = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} \mathrm{Var}(Y_i) \tag{8.4}$$

and the proportion of total population variance due to the k*th principal component*

$$\frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i}, \qquad k = 1, \ldots, p.$$

Remark. If most of the total population variance, for large p, can be attributed to the first one, two or three components, then these components can "replace" the original p variables without much loss of information.

Remark. The magnitude of e_{ik} – also principal component loading – measures the importance of the kth variable to the ith principal component and thus is a useful basis for interpretation. A large coefficient (in absolute value) corresponds to a high loading, while a coefficient near zero has a low loading.

Remark. One important use of principal components is interpreting the original data in terms of the principal components. The images of the original data under the principal components transformation are referred to as principal component scores.

Proposition 8.2.4. *If* $Y_1 = \boldsymbol{e}_1'\boldsymbol{X}, \ldots, Y_p = \boldsymbol{e}_p'\boldsymbol{X}$ *are the principal components obtained from the covariance matrix* $\boldsymbol{\Sigma}$ *then*

$$\rho_{Y_i, X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}, \qquad i, k = 1, \ldots, p.$$

Remark. Suppose $\boldsymbol{X} \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$. Then we have $c^2 = \sum_{i=1}^{p} y_i^2/\lambda_i$ and this equation defines an ellipsoid with axes $y_1, \ldots, y_p$ lying in directions of $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p$, respectively.

8.2.1 Principal Components obtained from Standardized Values

Principal components may also be obtained for the standardized variables

$$Z_i = \frac{X_i - \mu_i}{\sqrt{\sigma_{ii}}}, \qquad i = 1, \ldots, p \tag{8.5}$$

or

$$\boldsymbol{Z} = \left(\mathbf{V}^{1/2}\right)^{-1} (\boldsymbol{X} - \boldsymbol{\mu}), \tag{8.6}$$

where the diagonal standard deviation matrix $\mathbf{V}^{1/2}$ is defined as

$$\mathbf{V}^{1/2} = \begin{pmatrix} \sqrt{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}}. \end{pmatrix}.$$

We find $\mathrm{E}(\boldsymbol{Z}) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{Z}) = \left(\mathbf{V}^{-1/2}\right)^{-1} \boldsymbol{\Sigma} \left(\mathbf{V}^{1/2}\right)^{-1} = \boldsymbol{\rho}$.

Remark. The eigenvalue-eigenvector pairs derived from $\boldsymbol{\Sigma}$ are in general not the same as the ones derived from $\boldsymbol{\rho}$.

Proposition 8.2.5. *The ith principal component of the standardized variables* $\boldsymbol{Z}' = (Z_1, \ldots, Z_p)$ *with* $\mathrm{Cov}(\boldsymbol{Z}) = \boldsymbol{\rho}$*, is given by*

$$Y_i = \boldsymbol{e}_i' \boldsymbol{Z} = \boldsymbol{e}_i' \left(\mathbf{V}^{1/2}\right)^{-1} (\boldsymbol{X} - \boldsymbol{\mu}), \qquad i = 1, \ldots, p.$$

Moreover

$$\sum_{i=1}^{p} \mathrm{Var}(Y_i) = \sum_{i=1}^{p} \mathrm{Var}(Z_i) = p$$

and

$$\rho_{Y_i, Z_k} = e_{ik} \sqrt{\lambda_i}, \qquad i, k = 1, \ldots, p.$$

In this case $(\lambda_1, \boldsymbol{e}_1), \ldots, (\lambda_p, \boldsymbol{e}_p)$ *are the eigenvalue-eigenvector pairs of* $\boldsymbol{\rho}$ *with* $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$.

Remark. Variables should be standardized if they are measured on scales with widely differing ranges or if the units of measurements are not commensurate.

8.3 Summarizing Sample Variation by Principal Components

We now study the problem of summarizing the variation in n measurements on p variables with a few judiciously chosen linear combinations. Suppose the data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ represent n independent drawings from some p-dimensional population with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. These data yield the sample mean vector $\overline{\boldsymbol{x}}$, sample covariance matrix $\mathbf{S}$ and the sample correlation matrix $\mathbf{R}$.

The uncorrelated combinations with the largest variances will be called the sample principal components. The sample principal components (PC) are defined as those linear

combinations which have maximum sample variance. Specifically,

$$\begin{aligned}
\text{First sample PC} &= \text{linear combination } \boldsymbol{a}_1'\boldsymbol{x}_j \text{ that maximizes the} \\
&\quad \text{sample variance } \mathrm{Var}(\boldsymbol{a}_1'\boldsymbol{x}_j) \text{ subject to } \boldsymbol{a}_1'\boldsymbol{a}_1 = 1 \\
\text{Second sample PC} &= \text{linear combination } \boldsymbol{a}_2'\boldsymbol{x}_j \text{ that maximizes the} \\
&\quad \text{sample variance } \mathrm{Var}(\boldsymbol{a}_2'\boldsymbol{x}_j) \text{ subject to } \boldsymbol{a}_2'\boldsymbol{a}_2 = 1 \text{ and} \\
&\quad \mathrm{Cov}(\boldsymbol{a}_1'\boldsymbol{x}_j, \boldsymbol{a}_2'\boldsymbol{x}_j) = 0 \\
&\vdots \\
i\text{th sample PC} &= \text{linear combination } \boldsymbol{a}_i'\boldsymbol{x}_j \text{ that maximizes the} \\
&\quad \text{sample variance } \mathrm{Var}(\boldsymbol{a}_i'\boldsymbol{x}_j) \text{ subject to } \boldsymbol{a}_i'\boldsymbol{a}_i = 1 \text{ and} \\
&\quad \mathrm{Cov}(\boldsymbol{a}_i'\boldsymbol{x}_j, \boldsymbol{a}_k'\boldsymbol{x}_j) = 0, \quad k < i.
\end{aligned}$$

Proposition 8.3.1. *If* $\mathbf{S} = \{s_{ik}\}$ *is the* $p \times p$ *sample covariance matrix with eigenvalue-eigenvector pairs* $(\hat{\lambda}_1, \hat{\boldsymbol{e}}_1), \ldots, (\hat{\lambda}_p, \hat{\boldsymbol{e}}_p)$ *the* i*th sample principal component is given by*

$$\hat{y}_i = \hat{\boldsymbol{e}}_i'\boldsymbol{x} = \hat{e}_{i1}x_1 + \ldots + \hat{e}_{ip}x_p, \qquad i = 1, \ldots, p,$$

where $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_p \geq 0$ *and* $\boldsymbol{x}$ *is any observation on the variable* $X_1, \ldots, X_p$*. Also,*

$$\begin{aligned}
\textit{sample variance } (\hat{y}_k) &= \hat{\lambda}_k, \qquad k = 1, \ldots, p, \\
\textit{sample covariance } (\hat{y}_i, \hat{y}_k) &= 0, \qquad i \neq k.
\end{aligned}$$

In addition

$$\textit{total sample variance} = \sum_{i=1}^{p} s_{ii} = \hat{\lambda}_1 + \ldots + \hat{\lambda}_p$$

and

$$r(\hat{y}_i, x_k) = \frac{\hat{e}_{ik}\sqrt{\hat{\lambda}_i}}{\sqrt{\sigma_{kk}}}, \qquad i, k = 1, \ldots, p.$$

Number of Principal Components

How many components to retain? There is no definitive answer to this question. A useful visual aid to determining an appropriate number of principal components is a scree plot. With the eigenvalues ordered from largest to smallest, a scree plot is a plot of $\hat{\lambda}_i$ versus i. To determine the appropriate number of components, we look for an elbow (bend) in the scree plot. The number of components is taken to be the point at which the remaining eigenvalues are relatively small and all about the same size.

Remark. An unusually small value for the last eigenvalue from either the sample covariance or correlation matrix can indicate an unnoticed linear dependency in the data set and should therefore not be routinely ignored.

Example (Weather Report, p. 1-6)**.** Figure 8.2 shows the results of the principal component analysis calculated with the correlation matrix.

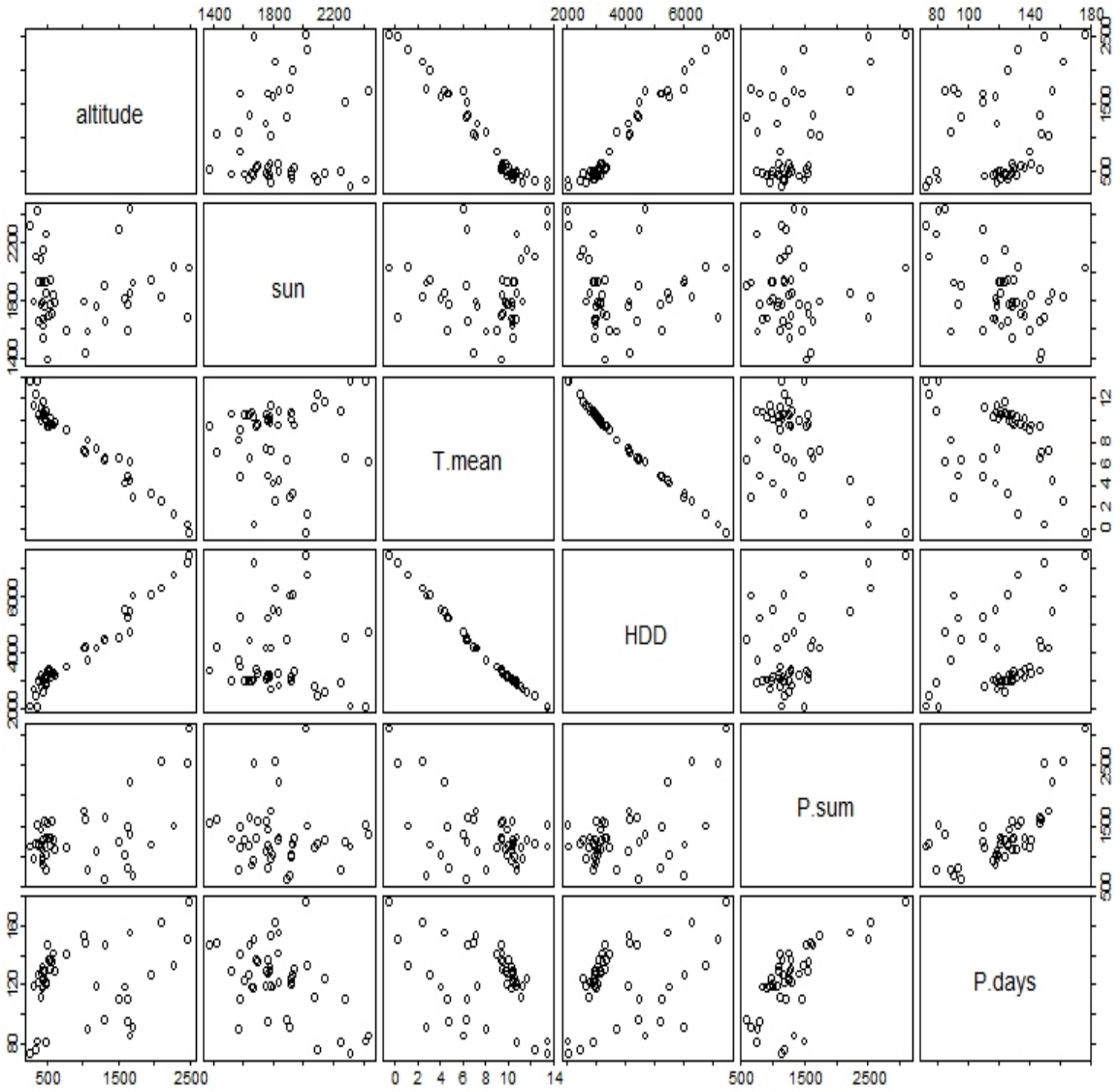

Figure 8.1: Scatterplot matrix of some variables of the Weather Report, p. 1-6.

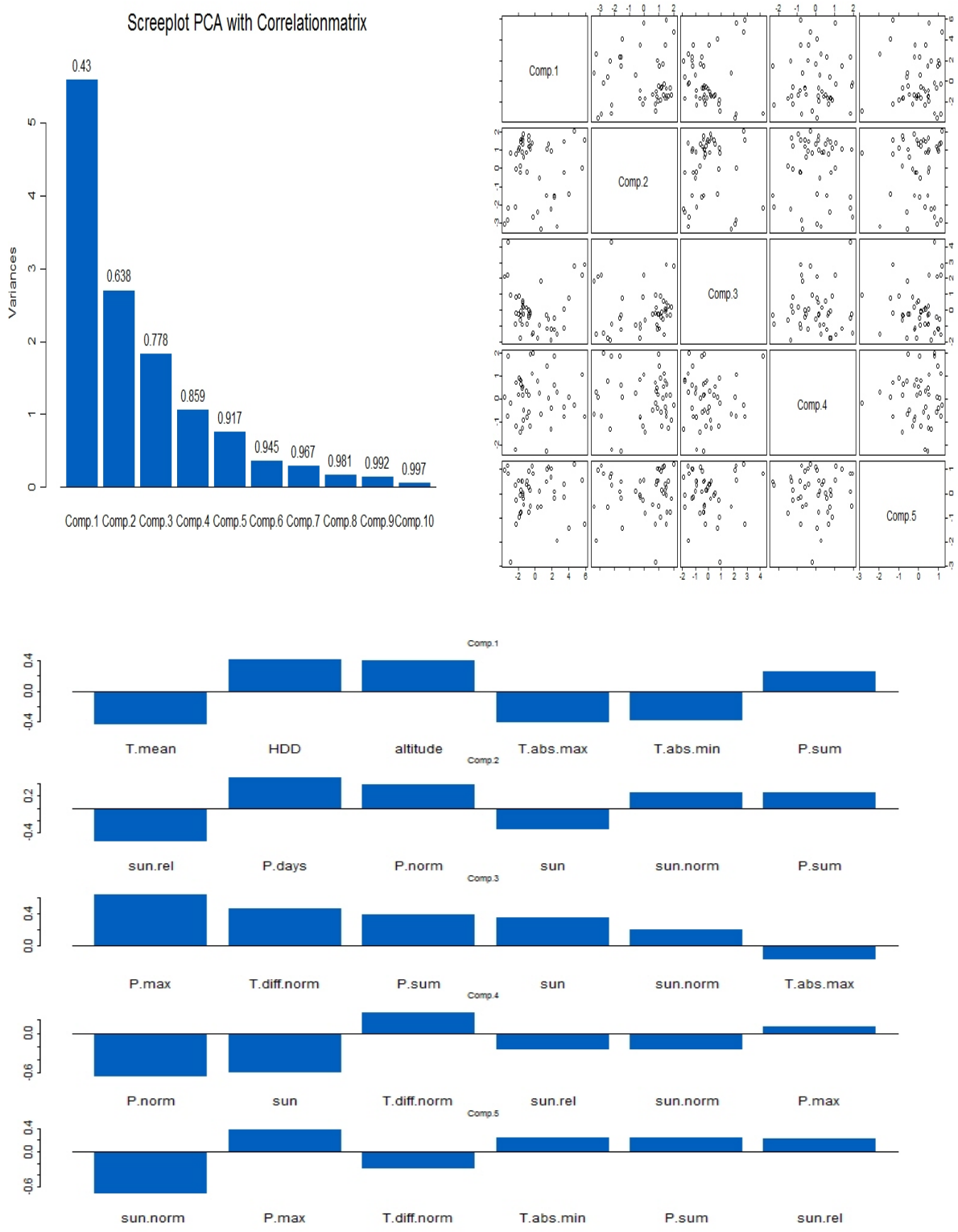

Figure 8.2: Screeplot of the principal components analysis (top left), scatterplot matrix of the scores calculated with the correlation matrix (top right) and the loadings of the variables (bottom). Data set: Weather Report, p. 1-6.

9 Canonical Correlation Analysis

Further reading. Read again the paper Xoplaki et al. (2000) to see how the canonical correlation analysis method can be applied in climate sciences.

9.1 Introduction

Example (Soil Evaporation, p. 1-8)**.** The observed variables are maximum (maxst), minimum (minst), and average soil temperature (avst); maximum (maxat), minimum (minat), and average air temperature (avat); maximum (maxh), minimum (minh), and average relative humidity (avh); total wind in miles per day (wind) and the daily amount of evaporation from the soil (evap). The three "average" measurements are integrated: average soil temperature is the integrated area under the daily soil temperature curve, average air temperature is the integrated area under the daily air temperature curve, and average relative humidity is the integrated area under the daily relative humidity curve.

We want to find the association between the soil variables (maxst, minst and avst) and the air variables (maxat, minat, avat, maxh, minh, avh, wind).

Canonical correlation analysis (CCA) seeks to identify and quantify the associations between two sets of variables. Canonical correlation analysis focuses on the correlation between a linear combination of the variables in one set and a linear combination of the variables in another set. The idea is first to determine the pair of linear combinations having the largest correlation. Next, we determine the pair of linear combinations having the largest correlation among all pairs uncorrelated with the initially selected pair, and so on. The pairs of linear combinations are called the canonical variables, and their correlations are called canonical correlations. The canonical correlations measure the strength of association between the two sets of variables. The maximization aspect of the technique represents an attempt to concentrate a high-dimensional relationship between two sets of variables into a few pairs of canonical variables.

9.2 Canonical Variates and Canonical Correlations

We are interested in measures of association between two groups of variables. The first group, of p variables, is represented by the $(p \times 1)$ random vector $\boldsymbol{X}^{(1)}$. The second group, of q variables, is represented by the $(q \times 1)$ random vector $\boldsymbol{X}^{(2)}$. We assume, in the theoretical development, that $\boldsymbol{X}^{(1)}$ represents the smaller set, so that $p \leq q$.

For the random vector $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$, let

$$\begin{aligned} \mathrm{E}(\boldsymbol{X}^{(1)}) &= \boldsymbol{\mu}^{(1)}, \qquad \mathrm{Cov}(\boldsymbol{X}^{(1)}) = \boldsymbol{\Sigma}_{11} \\ \mathrm{E}(\boldsymbol{X}^{(2)}) &= \boldsymbol{\mu}^{(2)}, \qquad \mathrm{Cov}(\boldsymbol{X}^{(2)}) = \boldsymbol{\Sigma}_{22} \\ \mathrm{Cov}(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}) &= \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}'. \end{aligned}$$

It will be convenient to consider $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ jointly, so we find that the random vector

$$\underset{((p+q)\times 1)}{\boldsymbol{X}} = \begin{pmatrix} \boldsymbol{X}^{(1)} \\ \hline \boldsymbol{X}^{(2)} \end{pmatrix}$$

has mean vector

$$\underset{((p+q)\times 1)}{\boldsymbol{\mu}} = \mathrm{E}(\boldsymbol{X}) = \begin{pmatrix} \boldsymbol{\mu}^{(1)} \\ \hline \boldsymbol{\mu}^{(2)} \end{pmatrix}$$

and covariance matrix

$$\begin{aligned} \boldsymbol{\Sigma} &= \mathrm{E}(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})' \\ &= \underset{\substack{((p+q)\times p) \\ ((p+q)\times q)}}{\left(\begin{array}{cc} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right)} \begin{matrix} (p\times(p+q)) \\ (q\times(p+q)) \end{matrix} \end{aligned}$$

The covariances between pairs of variables from different sets – one variable from $\boldsymbol{X}^{(1)}$, one variable from $\boldsymbol{X}^{(2)}$ – are contained in $\boldsymbol{\Sigma}_{12}$ or, equivalently, in $\boldsymbol{\Sigma}_{21}$. That is, the pq elements of $\boldsymbol{\Sigma}_{12}$ measure the association between the two sets. When p and q are relatively large, interpreting the elements of $\boldsymbol{\Sigma}_{12}$ collectively is ordinarily hopeless. Moreover, it is often linear combinations of variables that are interesting and useful for predictive or comparative purposes. The main task of canonical correlation analysis is to summarize the associations between the $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ sets in terms of a few carefully chosen covariances (or correlations) rather than the pq covariances in $\boldsymbol{\Sigma}_{12}$.

Linear combinations provide simple summary measures of a set of variables. Set

$$\begin{aligned} U &= \boldsymbol{a}'\boldsymbol{X}^{(1)} \\ V &= \boldsymbol{b}'\boldsymbol{X}^{(2)} \end{aligned}$$

for some pair of coefficient vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. We obtain

$$\begin{aligned} \mathrm{Var}(U) &= \boldsymbol{a}'\mathrm{Cov}(\boldsymbol{X}^{(1)})\boldsymbol{a} = \boldsymbol{a}'\boldsymbol{\Sigma}_{11}\boldsymbol{a} \\ \mathrm{Var}(V) &= \boldsymbol{b}'\mathrm{Cov}(\boldsymbol{X}^{(2)})\boldsymbol{b} = \boldsymbol{b}'\boldsymbol{\Sigma}_{22}\boldsymbol{b} \\ \mathrm{Cov}(U, V) &= \boldsymbol{a}'\mathrm{Cov}(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)})\boldsymbol{b} = \boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{b}. \end{aligned}$$

We shall seek coefficient vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ such that

$$\mathrm{Cor}(U,V) = \frac{\boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{b}}{\sqrt{\boldsymbol{a}'\boldsymbol{\Sigma}_{11}\boldsymbol{a}}\sqrt{\boldsymbol{b}'\boldsymbol{\Sigma}_{22}\boldsymbol{b}}} \tag{9.1}$$

is as large as possible.

Definition 9.2.1. We define:

First pair of canonical variables (first canonical variate pair): pair of linear combinations U_1, V_1 having unit variances, which maximizes the correlation (9.1)

Second pair of canonical variables (second canonical variate pair): pair of linear combinations U_2, V_2 having unit variances, which maximize the correlation (9.1) among all choices that are uncorrelated with the first pair of canonical variables.

At the kth step,

The kth pair of canonical variables (kth canonical variate pair): pair of linear combinations U_k, V_k having unit variances, which maximize the correlation (9.1) among all choices uncorrelated with the previous $k-1$ canonical variable pairs.

The correlation between the kth pair of canonical variables is called the kth canonical correlation.

Proposition 9.2.2. *Suppose $p \leq q$ and let the random vectors $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ have $\mathrm{Cov}(\boldsymbol{X}^{(1)}) = \boldsymbol{\Sigma}_{11}$, $\mathrm{Cov}(\boldsymbol{X}^{(2)}) = \boldsymbol{\Sigma}_{22}$ and $\mathrm{Cov}(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}) = \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}'$, where $\boldsymbol{\Sigma}$ has full rank. For coefficient vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, form the linear combinations $U = \boldsymbol{a}'\boldsymbol{X}^{(1)}$ and $V = \boldsymbol{b}'\boldsymbol{X}^{(2)}$. Then*

$$\max_{\boldsymbol{a},\boldsymbol{b}} \mathrm{Cor}(U,V) = \rho_1^*$$

is attained by the linear combinations

$$U_1 = \underbrace{\boldsymbol{e}_1'\boldsymbol{\Sigma}_{11}^{-1/2}}_{\boldsymbol{a}_1'}\boldsymbol{X}^{(1)} \quad \text{and} \quad V_1 = \underbrace{\boldsymbol{f}_1'\boldsymbol{\Sigma}_{22}^{-1/2}}_{\boldsymbol{b}_1'}\boldsymbol{X}^{(2)}.$$

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

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

maximizes

$$\mathrm{Cor}(U_k, V_k) = \rho_k^*$$

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

Here $(\rho_1^*)^2 \geq \ldots \geq (\rho_p^*)^2$ are the eigenvalues of $\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}$, and $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p$ are the associated $(p \times 1)$ eigenvectors.

The canonical variates have the properties

$$\begin{aligned}
\mathrm{Var}(U_k) &= \mathrm{Var}(V_k) = 1 \\
\mathrm{Cov}(U_k, U_l) &= \mathrm{Cov}(V_k, V_l) = \mathrm{Cov}(U_k, V_l) = 0 \\
\mathrm{Cor}(U_k, U_l) &= \mathrm{Cor}(V_k, V_l) = \mathrm{Cor}(U_k, V_l) = 0.
\end{aligned}$$

for $k, l = 1, 2, \ldots, p$ with $k \neq l$.

Remark. The quantities $(\rho_1^*)^2 \geq \ldots \geq (\rho_p^*)^2$ are also the p largest eigenvalues of the matrix

$$\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}$$

with corresponding $(q \times 1)$ eigenvectors $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_p$. Each $\boldsymbol{f}_i$ is proportional to

$$\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{e}_i.$$

Remark. If the original variables are standardized with $\boldsymbol{Z}^{(1)} = (Z_1^{(1)}, \ldots, Z_p^{(1)})'$ and $\boldsymbol{Z}^{(2)} = (Z_1^{(2)}, \ldots, Z_q^{(2)})'$, the canonical variates are of the form

$$\begin{aligned} U_k &= \boldsymbol{a}_k'\boldsymbol{Z}^{(1)} = \boldsymbol{e}_k'\boldsymbol{\rho}_{11}^{-1/2}\boldsymbol{Z}^{(1)} \\ V_k &= \boldsymbol{b}_k'\boldsymbol{Z}^{(2)} = \boldsymbol{f}_k'\boldsymbol{\rho}_{22}^{-1/2}\boldsymbol{Z}^{(2)}. \end{aligned} \tag{9.2}$$

Here, $\mathrm{Cov}(\boldsymbol{Z}^{(1)}) = \boldsymbol{\rho}_{11}$, $\mathrm{Cov}(\boldsymbol{Z}^{(2)}) = \boldsymbol{\rho}_{22}$, $\mathrm{Cov}(\boldsymbol{Z}^{(1)}, \boldsymbol{Z}^{(2)}) = \boldsymbol{\rho}_{12} = \boldsymbol{\rho}_{21}'$, and $\boldsymbol{e}_k$ and $\boldsymbol{f}_k$ are the eigenvectors of $\boldsymbol{\rho}_{11}^{-1/2}\boldsymbol{\rho}_{12}\boldsymbol{\rho}_{22}^{-1}\boldsymbol{\rho}_{21}\boldsymbol{\rho}_{11}^{-1/2}$ and $\boldsymbol{\rho}_{22}^{-1/2}\boldsymbol{\rho}_{21}\boldsymbol{\rho}_{11}^{-1}\boldsymbol{\rho}_{12}\boldsymbol{\rho}_{22}^{-1/2}$, respectively. The canonical correlations, ρ_k^*, satisfy

$$\mathrm{Cor}(U_k, V_k) = \rho_k^*, \qquad k = 1, \ldots, p,$$

where $(\rho_1^*)^2 \geq \ldots \geq (\rho_p^*)^2$ are the nonzero eigenvalues of the matrix $\boldsymbol{\rho}_{11}^{-1/2}\boldsymbol{\rho}_{12}\boldsymbol{\rho}_{22}^{-1}\boldsymbol{\rho}_{21}\boldsymbol{\rho}_{11}^{-1/2}$ or, equivalently, the largest eigenvalues of $\boldsymbol{\rho}_{22}^{-1/2}\boldsymbol{\rho}_{21}\boldsymbol{\rho}_{11}^{-1}\boldsymbol{\rho}_{12}\boldsymbol{\rho}_{22}^{-1/2}$.

Remark. The canonical coefficients for the standardized variables,

$$Z_i^{(1)} = \frac{X_i^{(1)} - \mu_i^{(1)}}{\sqrt{\sigma_{ii}}},$$

are simply related to the canonical coefficients attached to the original variables $X_i^{(1)}$. Specifically, if $\boldsymbol{a}_k'$ is the coefficient vector for the kth canonical variate U_k, then $\boldsymbol{a}_k'\mathbf{V}_{11}^{1/2}$ is the coefficient vector for the kth canonical variate constructed from the standardized variables $\boldsymbol{Z}^{(1)}$. Here $\mathbf{V}_{11}^{1/2}$ is the diagonal matrix with the ith diagonal element $\sqrt{\sigma_{ii}} = \sqrt{\mathrm{Var}(X_i^{(1)})}$. Similarly, $\boldsymbol{b}_k'\mathbf{V}_{22}^{1/2}$ is the coefficient vector for the canonical variate constructed from the set of standardized variables $\boldsymbol{Z}^{(2)}$. The canonical correlations are unchanged by the standardization.

9.3 Interpreting the Population Canonical Variables

Canonical variables are, in general, artificial. That is, they have no physical meaning. If the original variables $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ are used, the canonical coefficients $\boldsymbol{a}$ and $\boldsymbol{b}$ have units proportional to those of the $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ sets. If the original variables are standardized to have zero means and unit variances, the canonical coefficients have no units of measurement, and they must be interpreted in terms of the standardized variables.

9.3.1 Identifying the Canonical Variables

Even though the canonical variables are artificial, they can be "identified" in terms of the subject-matter variables. Many times this identification is aided by computing the correlations between the canonical variates and the original variables. These correlations, however, must be interpreted with caution. They provide only univariate information, in the sense that they do not indicate how the original variables contribute jointly to the canonical analyses. For this reason, many investigators prefer to assess the contributions of the original variables directly from the standardized coefficients (9.2).

Let $\mathbf{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_p)'$ and $\mathbf{B} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_q)'$, so that the vectors of canonical variables are

$$\underset{(p\times 1)}{\boldsymbol{U}} = \underset{(p\times p)}{\mathbf{A}}\ \underset{(p\times 1)}{\boldsymbol{X}^{(1)}} \quad \text{and} \quad \underset{(q\times 1)}{\boldsymbol{V}} = \underset{(q\times q)}{\mathbf{B}}\ \underset{(q\times 1)}{\boldsymbol{X}^{(2)}}$$

where we are primarily interested in the first p canonical variables in $\boldsymbol{V}$.

Introducing the $(p \times p)$ diagonal matrix $\mathbf{V}_{11}^{-1/2}$ with kth diagonal element

$$\sigma_{kk}^{-1/2} = \left(\mathrm{Var}(X_k^{(1)})\right)^{-1/2},$$

we find

$$\begin{aligned} \underset{(p\times p)}{\boldsymbol{\rho}_{\boldsymbol{U},\boldsymbol{X}^{(1)}}} &= \mathrm{Cor}(\boldsymbol{U}, \boldsymbol{X}^{(1)}) = \mathrm{Cov}(\boldsymbol{U}, \mathbf{V}_{11}^{-1/2}\boldsymbol{X}^{(1)}) \\ &= \mathrm{Cov}(\mathbf{A}\boldsymbol{X}^{(1)}, \mathbf{V}_{11}^{-1/2}\boldsymbol{X}^{(1)}) = \mathbf{A}\boldsymbol{\Sigma}_{11}\mathbf{V}_{11}^{-1/2}. \end{aligned} \tag{9.3}$$

Similar calculations for the pairs $(\boldsymbol{U}, \boldsymbol{X}^{(2)})$, $(\boldsymbol{V}, \boldsymbol{X}^{(2)})$ and $(\boldsymbol{V}, \boldsymbol{X}^{(1)})$ yield

$$\begin{aligned} \underset{(p\times q)}{\boldsymbol{\rho}_{\boldsymbol{U},\boldsymbol{X}^{(2)}}} &= \mathbf{A}\boldsymbol{\Sigma}_{12}\mathbf{V}_{22}^{-1/2} \\ \underset{(q\times q)}{\boldsymbol{\rho}_{\boldsymbol{V},\boldsymbol{X}^{(2)}}} &= \mathbf{B}\boldsymbol{\Sigma}_{22}\mathbf{V}_{22}^{-1/2} \\ \underset{(q\times p)}{\boldsymbol{\rho}_{\boldsymbol{V},\boldsymbol{X}^{(1)}}} &= \mathbf{B}\boldsymbol{\Sigma}_{21}\mathbf{V}_{11}^{-1/2}, \end{aligned} \tag{9.4}$$

where $\mathbf{V}_{22}^{-1/2}$ is the $(q \times q)$ diagonal matrix with the ith diagonal element

$$\sigma_{ii}^{-1/2} = \left(\mathrm{Var}(X_i^{(2)})\right)^{-1/2}.$$

Canonical variables derived from standardized variables are sometimes interpreted by computing the correlations. Thus

$$\begin{aligned} \boldsymbol{\rho}_{\boldsymbol{U},\boldsymbol{Z}^{(1)}} &= \mathbf{A}_{\boldsymbol{Z}}\boldsymbol{\rho}_{11}, \qquad \boldsymbol{\rho}_{\boldsymbol{V},\boldsymbol{Z}^{(2)}} = \mathbf{B}_{\boldsymbol{Z}}\boldsymbol{\rho}_{22} \\ \boldsymbol{\rho}_{\boldsymbol{U},\boldsymbol{Z}^{(2)}} &= \mathbf{A}_{\boldsymbol{Z}}\boldsymbol{\rho}_{12}, \qquad \boldsymbol{\rho}_{\boldsymbol{V},\boldsymbol{Z}^{(1)}} = \mathbf{B}_{\boldsymbol{Z}}\boldsymbol{\rho}_{21} \end{aligned}$$

where $\mathbf{A}_{\boldsymbol{Z}}$ and $\mathbf{B}_{\boldsymbol{Z}}$ are the matrices whose rows contain the canonical coefficients for the $\boldsymbol{Z}^{(1)}$ and $\boldsymbol{Z}^{(2)}$ sets, respectively.

Remark. The correlations are unaffected by the standardization, since for example

$$\boldsymbol{\rho}_{\boldsymbol{U},\boldsymbol{X}^{(1)}} = \mathbf{A}\boldsymbol{\Sigma}_{11}\mathbf{V}_{11}^{-1/2} = \underbrace{\mathbf{A}\mathbf{V}_{11}^{1/2}}_{\mathbf{A}_{\boldsymbol{Z}}}\underbrace{\mathbf{V}_{11}^{-1/2}\boldsymbol{\Sigma}_{11}\mathbf{V}_{11}^{-1/2}}_{\boldsymbol{\rho}_{11}} = \boldsymbol{\rho}_{\boldsymbol{U},\boldsymbol{Z}^{(1)}}.$$

Remark. The correlations $\boldsymbol{\rho}_{\boldsymbol{U},\boldsymbol{X}^{(1)}}$ and $\boldsymbol{\rho}_{\boldsymbol{V},\boldsymbol{X}^{(2)}}$ can help supply meanings for the canonical variates. The spirit is the same as in principal component analysis when the correlations between the principal components and their associated variables may provide subject-matter interpretations for the components.

9.4 Sample Canonical Variates and Sample Canonical Correlations

A random sample of n observations on each of the $(p+q)$ variables $\boldsymbol{X}^{(1)}$, $\boldsymbol{X}^{(2)}$ can be assembled into the $n \times (p+q)$ data matrix

$$\mathbf{X} = (\mathbf{X}^{(1)} \mid \mathbf{X}^{(2)})$$

$$= \begin{pmatrix} x_{11}^{(1)} & \cdots & x_{1p}^{(1)} & \vdots & x_{11}^{(2)} & \cdots & x_{1q}^{(2)} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1}^{(1)} & \cdots & x_{np}^{(1)} & \vdots & x_{n1}^{(2)} & \cdots & x_{nq}^{(2)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^{(1)\prime} & \vdots & \boldsymbol{x}_1^{(2)\prime} \\ \vdots & \vdots & \vdots \\ \boldsymbol{x}_n^{(1)\prime} & \vdots & \boldsymbol{x}_n^{(2)\prime} \end{pmatrix}.$$

We find $\overline{\boldsymbol{x}} = \begin{pmatrix} \overline{\boldsymbol{x}}^{(1)} \\ \overline{\boldsymbol{x}}^{(2)} \end{pmatrix}$ where $\overline{\boldsymbol{x}}^{(i)} = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{x}_j^{(i)}$, $i = 1, 2$ and

$$\mathbf{S} = \underset{((p+q)\times p)}{\begin{pmatrix} \mathbf{S}_{11} \\ \mathbf{S}_{21} \end{pmatrix}} \quad \underset{((p+q)\times q)}{\begin{pmatrix} \mathbf{S}_{12} \\ \mathbf{S}_{22} \end{pmatrix}} \begin{matrix} (p\times(p+q)) \\ (q\times(p+q)) \end{matrix} \qquad \text{with} \qquad \mathbf{S}_{12} = \mathbf{S}_{21}'$$

and

$$\mathbf{S}_{kl} = \frac{1}{n-1}\sum_{j=1}^{n}(\boldsymbol{x}_j^{(k)} - \overline{\boldsymbol{x}}^{(k)})(\boldsymbol{x}_j^{(l)} - \overline{\boldsymbol{x}}^{(l)})', \qquad k, l = 1, 2. \tag{9.5}$$

The linear combinations $\hat{U} = \hat{\boldsymbol{a}}'\boldsymbol{x}^{(1)}$ and $\hat{V} = \hat{\boldsymbol{b}}'\boldsymbol{x}^{(2)}$ have sample correlation

$$r_{\hat{U},\hat{V}} = \frac{\hat{\boldsymbol{a}}'\mathbf{S}_{12}\hat{\boldsymbol{b}}}{\sqrt{\hat{\boldsymbol{a}}'\mathbf{S}_{11}\hat{\boldsymbol{a}}}\sqrt{\hat{\boldsymbol{b}}'\mathbf{S}_{22}\hat{\boldsymbol{b}}}}. \tag{9.6}$$

The first pair of sample canonical variates is the pair of linear combinations $\hat{U}_1$, $\hat{V}_1$ having unit sample variances that maximizes the ratio (9.6).

In general, the kth pair of sample canonical variates is the pair of linear combinations $\hat{U}_k$, $\hat{V}_k$ having unit sample variances that maximizes the ratio (9.6) among those linear combinations uncorrelated with the previous $k-1$ sample canonical variates.

The sample correlation between $\hat{U}_k$ and $\hat{V}_k$ is called the kth sample canonical correlation.

Example (Soil Evaporation, p. 1-8)**.** Table 9.1 and Figure 9.1 show the results of the canonical correlation analysis of the soil and the air variables.

Table 9.1: Results of the canonical correlation analysis. Data set: Soil Evaporation, p. 1-8.

```
> cc.airsoil
$cor:
[1] 0.9624326 0.7604630 0.5963187

$xcoef:
              [,1]          [,2]          [,3]          [,4]          [,5]          [,6]          [,7]
[1,] -0.00810700994 -0.0131838494 -0.0161725757  0.02517594372  0.04412103124 -0.0579086504 -0.0005754910
[2,]  0.00216809356 -0.0440762196 -0.0176049769  0.01160575703 -0.08327800431 -0.0336352188 -0.0366291212
[3,] -0.00470274969  0.0086711986  0.0003125829 -0.00638888876  0.00565562936  0.0235464153  0.0041274285
[4,]  0.01681633315 -0.0389196441  0.0407784120  0.12108216355 -0.05106799333 -0.0318368973  0.0450208056
[5,]  0.01080944196 -0.0294491540  0.0218271660 -0.00541588722  0.02269311835 -0.0180959750  0.0190990783
[6,] -0.00218121839  0.0096778587 -0.0110017742  0.00148441044  0.00070780115  0.0098347719 -0.0065376621
[7,]  0.00001586807  0.0005137182 -0.0004222563  0.00007431056  0.00004439516 -0.0004388761  0.0009371029


$ycoef:
             [,1]         [,2]        [,3]
[1,] -0.023889982  0.079116806 -0.02793426
[2,]  0.006589414  0.007931868 -0.13448882
[3,] -0.001183075 -0.026269281  0.02816080

$xcenter:
    maxat    minat  avat     maxh minh     avh     wind
 90.73913 70.06522 190.5 94.71739 48.5 396.913 277.6739

$ycenter:
    maxst    minst     avst
 87.56522 71.26087 173.5217
```

Proposition 9.4.1. *Let* $\widehat{\rho_1^*}^2 \geq \ldots \geq \widehat{\rho_p^*}^2$ *be the* p *ordered eigenvalues of*

$$\mathbf{S}_{11}^{-1/2}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-1/2}$$

with corresponding eigenvectors $\hat{\boldsymbol{e}}_1, \ldots, \hat{\boldsymbol{e}}_p$*, where the* $\mathbf{S}_{kl}$ *are defined in* (9.5) *and* $p \leq q$*. Let* $\hat{\boldsymbol{f}}_1, \ldots, \hat{\boldsymbol{f}}_p$ *be the eigenvectors of*

$$\mathbf{S}_{22}^{-1/2}\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1/2},$$

where the first p *eigenvectors* $\hat{\boldsymbol{f}}$ *may be obtained from*

$$\hat{\boldsymbol{f}}_k = \frac{1}{\widehat{\rho_k^*}}\mathbf{S}_{22}^{-1/2}\mathbf{S}_{21}\mathbf{S}_{11}^{-1/2}\hat{\boldsymbol{e}}_k, \quad k = 1, \ldots, p.$$

Then the k*th sample canonical variate pair is*

$$\hat{U}_k = \underbrace{\hat{\boldsymbol{e}}_k'\mathbf{S}_{11}^{-1/2}}_{=\hat{\boldsymbol{a}}_k'}\boldsymbol{x}^{(1)} \qquad \hat{V}_k = \underbrace{\hat{\boldsymbol{f}}_k'\mathbf{S}_{22}^{-1/2}}_{=\hat{\boldsymbol{b}}_k'}\boldsymbol{x}^{(2)}$$

where $\boldsymbol{x}^{(1)}$ *and* $\boldsymbol{x}^{(2)}$ *are the values of the variables* $\boldsymbol{X}^{(1)}$ *and* $\boldsymbol{X}^{(2)}$ *for a particular experimental unit.*

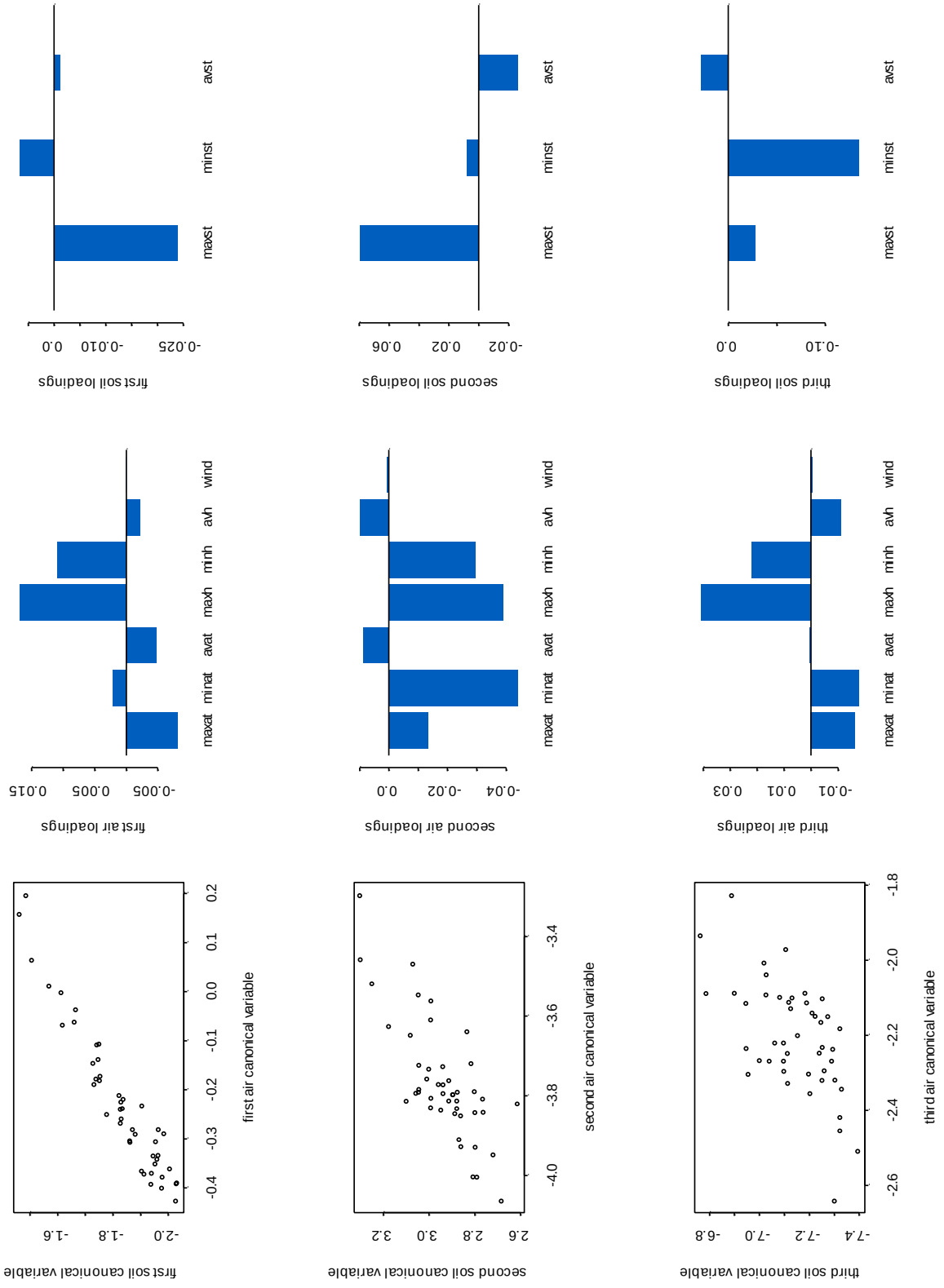

Figure 9.1: Results of the canonical correlation analysis. Data set: Soil Evaporation, p. 1-8.

Remark. The interpretation of $\hat{U}_k$ and $\hat{V}_k$ is often aided by computing the sample correlations between the canonical variates and the variables in the sets $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$. We define the matrices

$$\underset{(p\times p)}{\hat{\mathbf{A}}} = \underset{(p\times 1)}{(\hat{\boldsymbol{a}}_1} \quad , \ldots , \quad \underset{(p\times 1)}{\hat{\boldsymbol{a}}_p)'}$$

$$\underset{(q\times q)}{\widehat{\mathbf{B}}} = \underset{(q\times 1)}{(\hat{\boldsymbol{b}}_1} \quad , \ldots , \quad \underset{(q\times 1)}{\hat{\boldsymbol{b}}_q)'}$$

whose rows are the coefficient vectors for the sample canonical variates. We find

$$\underset{(p\times 1)}{\hat{\boldsymbol{U}}} = \underset{(p\times p)}{\hat{\mathbf{A}}} \; \underset{(p\times 1)}{\boldsymbol{x}^{(1)}} \qquad \text{and} \qquad \underset{(q\times 1)}{\hat{\boldsymbol{V}}} = \underset{(q\times q)}{\widehat{\mathbf{B}}} \; \underset{(q\times 1)}{\boldsymbol{x}^{(2)}}$$

and can define

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

and corresponding to (9.3) and (9.4) we find

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

where $\mathbf{D}_{11}^{-1/2}$ and $\mathbf{D}_{22}^{-1/2}$ are the $(p \times p)$ and $(q \times q)$ diagonal matrix with ith diagonal element

$$\left(\text{sample variance}(x_i^{(1)})\right)^{-1/2} \text{ and } \left(\text{sample variance}(x_i^{(2)})\right)^{-1/2},$$

respectively.

9.5 Canonical Correlation Analysis applied to Fields and Forecasting with Canonical Correlation Analysis

Source: Wilks (2006) pp. 519-522 (see next pages).

12.2 CCA Applied to Fields

12.2.1 Translating Canonical Vectors to Maps

Canonical correlation analysis is usually most interesting for atmospheric data when applied to fields. Here the spatially distributed observations (either at gridpoints or observing locations) are encoded into the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in the same way as for PCA. That is, even though the data may pertain to a two- or three-dimensional field, each location is numbered sequentially and pertains to one element of the corresponding data vector. It is not necessary for the spatial domains encoded into $\boldsymbol{x}$ and $\boldsymbol{y}$ to be the same, and indeed in the applications of CCA that have appeared in the literature they are usually different.

As is the case with the use of PCA with spatial data, it is often informative to plot maps of the canonical vectors by associating the magnitudes of their elements and the geographic locations to which they pertain. In this context the canonical vectors are sometimes called canonical patterns, since the resulting maps show spatial patterns of the ways in which the original variables contribute to the canonical variables. Examining the pairs of maps formed by corresponding vectors $\boldsymbol{a}_m$ and $\boldsymbol{b}_m$ can be informative about the nature of the relationship between variations in the data over the two domains encoded in $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. Figures 12.2 and 12.3 show examples of maps of canonical vectors.

It can also be informative to plot pairs of maps of the homogeneous (Equation 12.7) or heterogeneous correlations (Equation 12.8). Each of these vectors contain correlations between an underlying data field and one of the canonical variables, and these correlations can also be plotted at the corresponding locations. Figure 12.1, from Wallace *et al.* (1992), shows one such pair of homogeneous correlation patterns. Figure 12.1a shows the spatial distribution of correlations between a canonical variable v, and the values of the corresponding data $\boldsymbol{x}$ that contains values of average December-February sea-surface temperatures (SSTs) in the north Pacific Ocean. This canonical variable accounts for 18% of the total variance of the SSTs in the data set analyzed (Equation 12.12). Figure 12.1b shows the spatial distribution of the correlations for the corresponding canonical variable w, that pertains to average hemispheric 500 mb heights $\boldsymbol{y}$ during the same winters included in the SST data in $\boldsymbol{x}$. This canonical variable accounts for 23% of the total variance of the winter hemispheric height variations. The correlation pattern in Figure 12.1a corresponds to either cold water in the central north Pacific and warm water along the west coast of North America, or warm water in the central north Pacific and cold water along the west coast of North America. The pattern of 500 mb height correlations in Figure 12.1b is remarkably similar to the PNA pattern (cf. Figures 11.10b and 3.28).

The correlation between the two time series v and w is the canonical correlation $r_C = 0.79$. Because v and w are well correlated, these figures indicate that cold SSTs in the central Pacific simultaneously with warm SSTs in the northeast Pacific (relatively large positive v) tend to coincide with a 500 mb ridge over northwestern North America and a 500 mb trough over southeastern North America (relatively large positive w). Similarly, warm water in the central north Pacific and cold water in the northwestern Pacific (relatively large negative v) are associated with the more zonal PNA flow (relatively large negative w).

12.2.2 Combining CCA with PCA

The sampling properties of CCA can be poor when the available data are few relative to the dimensionality of the data vectors. The result can be that sample estimates for

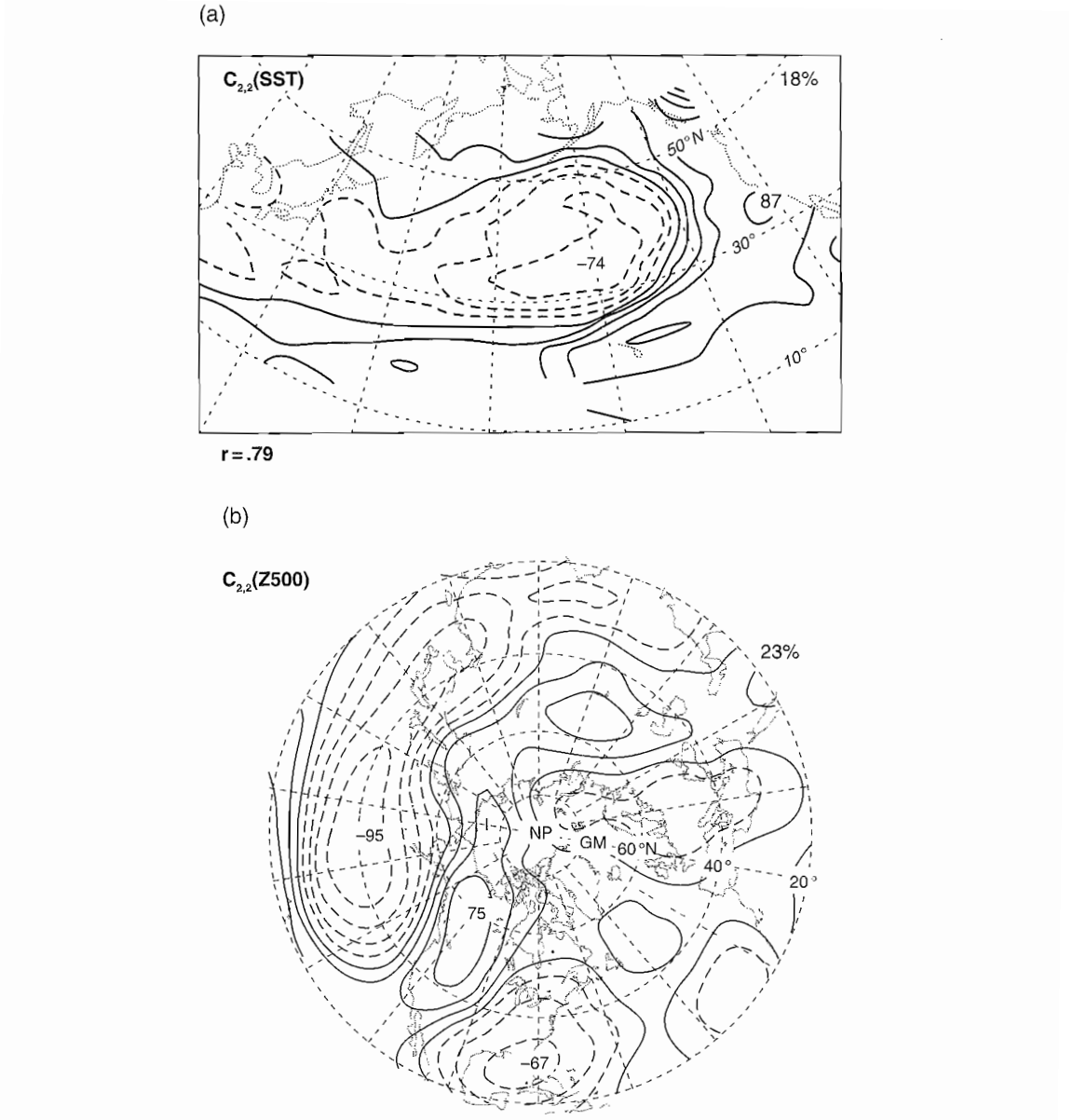

FIGURE 12.1 Homogeneous correlation maps for a pair of canonical variables pertaining to (a) average winter sea-surface temperatures (SSTs) in the northern Pacific Ocean, and (b) hemispheric winter 500 mb heights. The pattern of SST correlation in the left-hand panel (and its negative) are associated with the PNA pattern of 500 mb height correlations shown in the right-hand panel. The canonical correlation for this pair of canonical variables is 0.79. From Wallace *et al.* (1992).

CCA parameters may be unstable (i.e., exhibit large variations from batch to batch) for small samples (e.g., Bretherton *et al.* 1992; Cherry 1996; Friedrerichs and Hense 2003). Friedrerichs and Hense (2003) describe, in the context of atmospheric data, both conventional parametric tests and resampling tests to help assess whether sample canonical correlations may be spurious sampling artifacts. These tests examine the null hypothesis that all the underlying population canonical correlations are zero.

Relatively small sample sizes are common when analyzing time series of atmospheric fields. In CCA, it is not uncommon for there to be fewer observations n than the dimensions I and J of the data vectors, in which case the necessary matrix inversions

cannot be computed (see Section 12.3). However, even if the sample sizes are large enough to carry through the calculations, sample CCA statistics are erratic unless $n \gg M$. Barnett and Preisendorfer (1987) suggested that a remedy for this problem is to prefilter the two fields of raw data using separate PCAs before subjecting them to a CCA, and this has become a conventional procedure. Rather than directly correlating linear combinations of the fields $\boldsymbol{x}'$ and $\boldsymbol{y}'$, the CCA operates on the vectors $\boldsymbol{u}_x$ and $\boldsymbol{u}_y$, which consist of the leading principal components of $\boldsymbol{x}'$ and $\boldsymbol{y}'$. The truncations for these two PCAs (i.e., the dimensions of the vectors $\boldsymbol{u}_x$ and $\boldsymbol{u}_y$) need not be the same, but should be severe enough for the larger of the two to be substantially smaller than the sample size n. Livezey and Smith (1999) provide some guidance for the subjective choices that need to be made in this approach.

This combined PCA/CCA approach is not always best, and can be inferior if important information is discarded when truncating the PCA. In particular, there is no guarantee that the most strongly correlated linear combinations of $\boldsymbol{x}$ and $\boldsymbol{y}$ will be well related to the leading principal components of one field or the other.

10 Discrimination and Classification

10.1 Introduction

Discrimination and classification are multivariate techniques concerned with separating distinct sets of objects (or observations) and with allocating new objects (observations) to previously defined groups. Discriminant analysis is rather exploratory in nature. As a separative procedure, it is often employed on a one-time basis in order to investigate observed differences when causal relationships are not well understood. Classification procedures are less exploratory in the sense that they lead to well-defined rules, which can be used for assigning new objects. Classification ordinarily requires more problem structure than discrimination does.

Thus, the immediate goals of discrimination and classification, respectively, are as follows:

- Discrimination (or separation): describes the differential features of objects (observations) from several known collections (populations). We try to find "discriminants" whose numerical values are such that the collection is separated as much as possible.
- Classification (or allocation): sorts 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.

In practice, the two goals frequently overlap, and the distinction between separation and allocation becomes blurred.

10.2 Separation and Classification for Two Populations

Example (El Niño)**.** Consider the data set in Table 10.1. With discriminant analysis we want to classify the years in the data set as either El Niño or non-El Niño, on the basis of the corresponding temperature and pressure data.

It is convenient to label the classes π_1 and π_2. The objects are ordinarily separated or classified on the basis of measurements on, for instance, p associated random variables $\boldsymbol{X}' = (X_1, \ldots, X_p)$. The observed values of $\boldsymbol{X}$ differ to some extent from one class to the other. We can think of the totality of values from the first class as being the population $\boldsymbol{x}$ values for π_1 and those from the second class as the population of $\boldsymbol{x}$ values for π_2. These two populations can then be described by probability density functions $f_1(\boldsymbol{x})$ and

Table 10.1: June climate data for Guayaquil, Ecuador, 1951-1970. Source: Wilks (2006).

Year	Temp. (°C)	Prec. (mm)	Pressure (hPa)	El Nino yes=1, no=0
1951	26.1	43	1009.5	1
1952	24.5	10	1010.9	0
1953	24.8	4	1010.7	1
1954	24.5	0	1011.2	0
1955	24.1	2	1011.9	0
1956	24.3	NA	1011.2	0
1957	26.4	31	1009.3	1
1958	24.9	0	1011.1	0
1959	23.7	0	1012.0	0
1960	23.5	0	1011.4	0
1961	24.0	2	1010.9	0
1962	24.1	3	1011.5	0
1963	23.7	0	1011.0	0
1964	24.3	4	1011.2	0
1965	26.6	15	1009.9	1
1966	24.6	2	1012.5	0
1967	24.8	0	1011.1	0
1968	24.4	1	1011.8	0
1969	26.8	127	1009.3	1
1970	25.2	2	1010.6	0

$f_2(\boldsymbol{x})$, and consequently, we can talk of assigning observations to populations or objects to classes interchangeably.

A good classification procedure should:

- result in a few misclassifications. In other words, the chances, or probabilities, of misclassification should be small
- take into account that one class or population has a greater likelihood of occurrence than another because one of the two populations is relatively much larger than the other

  Example. A randomly selected firm should be classified as nonbankrupt unless the data overwhelmingly favors bankruptcy.

- whenever possible, account for the costs associated with misclassification.

Classification rules cannot usually provide an error-free method of assignment. This is because there may not be a clear distinction between the measured characteristics of

the populations; that is, the groups may overlap. It is then possible, for example, to incorrectly classify a π_2 object as belonging to π_1 or a π_1 object as belonging to π_2.

Let $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$ be the probability density functions associated with the $(p \times 1)$ random vector $\boldsymbol{X}$ for the populations π_1 and π_2, respectively. An object with associated measurements $\boldsymbol{x}$ must be assigned to either π_1 or π_2. Let Ω be the sample space and R_1 that set of $\boldsymbol{x}$ values for which we classify objects as π_1 and $R_2 = \Omega - R_1$ be the remaining $\boldsymbol{x}$ values for which we classify objects as π_2. Since every object must be assigned to one and only one of the two populations, the sets R_1 and R_2 are mutually exclusive and exhaustive. For $p = 2$, we might have a case like the one pictured in Figure 10.1.

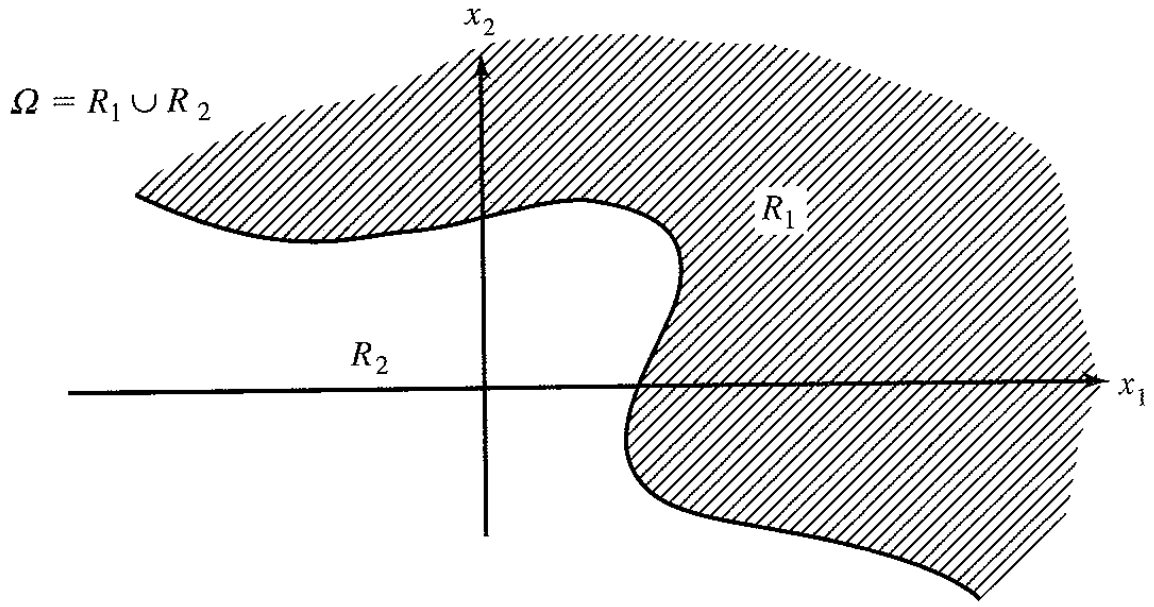

Figure 10.1: Classification regions for two populations. Source: Johnson and Wichern (2007).

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

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

Similarly, the conditional probability, $P(1|2)$, of classifying an object as π_1 when it is really from π_2 is

$$P(1|2) = P(\boldsymbol{X} \in R_1|\pi_2) = \int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x}.$$

This is illustrated in Figure 10.2 for the univariate case, $p = 1$.

Let $p_1 = P(\pi_1)$ be the prior probability of π_1 and $p_2 = P(\pi_2)$ be the prior probability of π_2, where $p_1 + p_2 = 1$. Then the overall probabilities of correctly or incorrectly

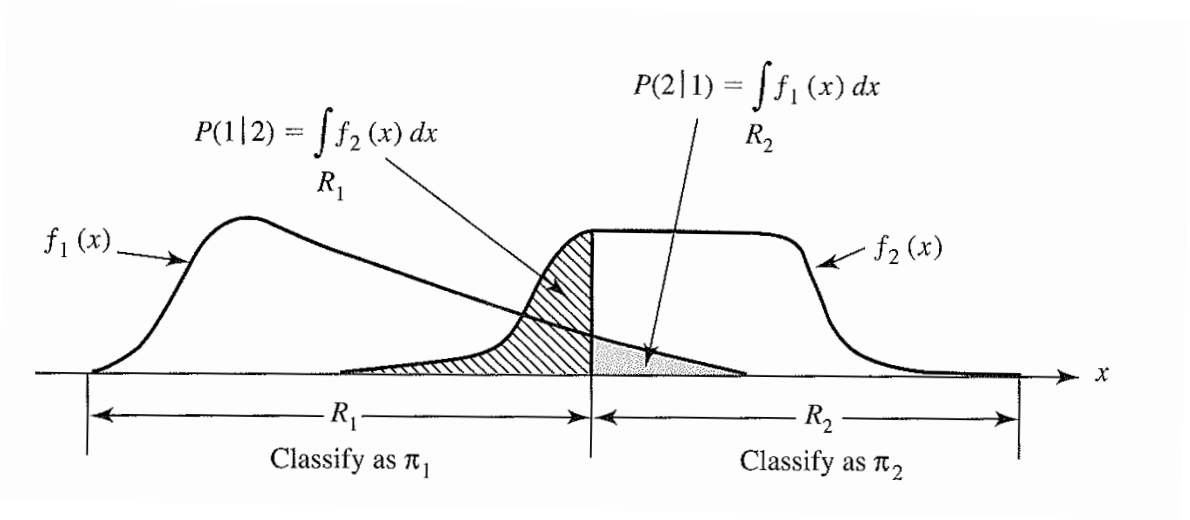

Figure 10.2: Misclassification probabilities for hypothetical classification regions. Source: Johnson and Wichern (2007).

classifying objects can be derived as the product of the prior and conditional classification probabilities:

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

Classification schemes are often evaluated in terms of their misclassification probabilities, but this ignores misclassification costs. A rule that ignores costs may cause problems. The costs of misclassification can be defined by a cost matrix (Table 10.2)

The costs are zero for correct classification, $c(1|2)$ when an observation from π_2 is incorrectly classified as π_1, and $c(2|1)$ when a π_1 observation is incorrectly classified as π_2.

For any rule, the average, or expected cost of misclassification (ECM) is provided by multiplying the off-diagonal entries in Table 10.2 by their probabilities of occurrence.

Table 10.2: Costs of misclassification

		classify as:	
		π_1	π_2
true population:	π_1	0	$c(2\|1)$
	π_2	$c(1\|2)$	0

Consequently,

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

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

Proposition 10.2.1. *The regions R_1 and R_2 that minimize the ECM are defined by the values $\boldsymbol{x}$ for which the following inequalities hold:*

$$\begin{aligned} R_1: \quad & \underbrace{\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})}}_{\textit{density ratio}} \geq \underbrace{\frac{c(1|2)}{c(2|1)}}_{\textit{cost ratio}} \underbrace{\frac{p_2}{p_1}}_{\substack{\textit{prior prob-}\\ \textit{ability ratio}}} \\ R_2: \quad & \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1} \end{aligned} \tag{10.1}$$

Remark. It is interesting to consider the classification regions defined in (10.1) for some special cases:

- When the prior probabilities are unknown, they are often taken to be equal, i.e. $p_2/p_1 = 1$, and the minimum ECM rule involves comparing the ratio of the population densities to the ratio of the appropriate misclassification costs;
- equal misclassification costs: $c(1|2)/c(2|1) = 1$;
- equal prior probabilities and equal misclassification costs: $p_2/p_1 = c(1|2)/c(2|1) = 1$.

If $\boldsymbol{x}_0$ is a new observation, it is assigned to π_1, if

$$\frac{f_1(\boldsymbol{x}_0)}{f_2(\boldsymbol{x}_0)} \geq \frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1}.$$

On the other hand, if

$$\frac{f_1(\boldsymbol{x}_0)}{f_2(\boldsymbol{x}_0)} < \frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1},$$

we assign $\boldsymbol{x_0}$ to π_2.

Remark. There are other criterias to derive "optimal" classification procedures. For example, choose R_1 and R_2 to minimize the total probability of misclassification (TPM):

$$\begin{aligned}\text{TPM} &= P(\text{misclassifying a } \pi_1 \text{ observation or misclassifying a } \pi_2 \text{ observation}) \\ &= P(\text{observation comes from } \pi_1 \text{ and is misclassified}) \\ &\quad + P(\text{observation comes from } \pi_2 \text{ and is misclassified}) \\ &= p_1 \int_{R_2} f_1(\boldsymbol{x})\, d\boldsymbol{x} + p_2 \int_{R_1} f_2(\boldsymbol{x})\, d\boldsymbol{x}.\end{aligned}$$

Mathematically, this problem is equivalent to minimizing the expected cost of misclassification when the costs of misclassification are equal.

10.2.1 Classification with Two Multivariate Normal Populations

Classification procedures based on normal populations predominate in statistical practice because of their simplicity and reasonably high efficiency across a wide variety of population models. We now assume that $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$ are multivariate normal densities, the first with mean vector $\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{\Sigma}_1$ and the second with mean vector $\boldsymbol{\mu}_2$ and covariance matrix $\boldsymbol{\Sigma}_2$. There are two cases to distinguish:

i) Case $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$

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

$$f_i(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-1/2(\boldsymbol{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right) \qquad \text{for } i = 1, 2. \tag{10.2}$$

- Suppose that $\boldsymbol{\mu_1}$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$ are known. Then, the minimum ECM regions in (10.1) become

$$\begin{aligned}R_1: \quad &\exp\left(-1/2(\boldsymbol{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + 1/2(\boldsymbol{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right) \\ &\qquad \geq \frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1} \\ R_2: \quad &\exp\left(-1/2(\boldsymbol{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + 1/2(\boldsymbol{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right) \\ &\qquad < \frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}\end{aligned}$$

Proposition 10.2.2. *Let the populations π_1 and π_2 be described by multivariate normal densities of the form* (10.2)*. Then the allocation rule that minimizes the ECM is as follows: Allocate $\boldsymbol{x}_0$ to π_1 if*

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_0 - 1/2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq \ln\left(\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}\right). \tag{10.3}$$

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

- Suppose that $\boldsymbol{\mu_1}$, $\boldsymbol{\mu_2}$, and $\boldsymbol{\Sigma}$ are unknown. Suppose then, that we have n_1 observations of the multivariate normal random vector $\boldsymbol{X}' = (X_1, \ldots, X_p)$ from π_1 and n_2 observations of this quantity from π_2, with $n_1 + n_2 - 2 \geq p$. Then the respective data matrices are

$$\underset{(n_1 \times p)}{\mathbf{X}_1} = \begin{pmatrix} \boldsymbol{x}'_{11} \\ \vdots \\ \boldsymbol{x}'_{1n_1} \end{pmatrix} = \begin{pmatrix} x_{111} & \ldots & x_{11p} \\ \vdots & \ddots & \vdots \\ x_{1n_11} & \ldots & x_{1n_1p} \end{pmatrix} \quad \text{and} \quad \underset{(n_2 \times p)}{\mathbf{X}_2} = \begin{pmatrix} \boldsymbol{x}'_{21} \\ \vdots \\ \boldsymbol{x}'_{2n_2} \end{pmatrix}.$$

From these data matrices, the sample mean vectors and covariance matrices are determined by

$$\underset{(p\times 1)}{\overline{\boldsymbol{x}}_1} = \frac{1}{n_1}\sum_{j=1}^{n_1} \boldsymbol{x}_{1j}, \qquad \underset{(p\times p)}{\mathbf{S}_1} = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(\boldsymbol{x}_{1j} - \overline{\boldsymbol{x}}_1)(\boldsymbol{x}_{1j} - \overline{\boldsymbol{x}}_1)'$$

$$\underset{(p\times 1)}{\overline{\boldsymbol{x}}_2} = \frac{1}{n_2}\sum_{j=1}^{n_2} \boldsymbol{x}_{2j}, \qquad \underset{(p\times p)}{\mathbf{S}_2} = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(\boldsymbol{x}_{2j} - \overline{\boldsymbol{x}}_2)(\boldsymbol{x}_{2j} - \overline{\boldsymbol{x}}_2)'$$

Since it is assumed that the parent populations have the same covariance matrix $\boldsymbol{\Sigma}$, the sample covariance matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ are combined (pooled) to derive a single, unbiased estimate of $\boldsymbol{\Sigma}$. In particular, the weighted average

$$\mathbf{S}_{\text{pooled}} = \frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)}\mathbf{S}_1 + \frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)}\mathbf{S}_2$$

is an unbiased estimate of $\boldsymbol{\Sigma}$ if the data matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ contain random samples from the populations π_1 and π_2, respectively.

Substituting $\overline{\boldsymbol{x}}_1$ for $\boldsymbol{\mu}_1$, $\overline{\boldsymbol{x}}_2$ for $\boldsymbol{\mu}_2$, and $\mathbf{S}_{\text{pooled}}$ for $\boldsymbol{\Sigma}$ in (10.3) gives the "sample" classification rule:

Proposition 10.2.3. *Allocate $\boldsymbol{x}_0$ to π_1 if*

$$(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)'\mathbf{S}^{-1}_{pooled}\boldsymbol{x}_0 - 1/2(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)'\mathbf{S}^{-1}_{pooled}(\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2) \geq \ln\left(\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}\right).$$

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

ii) Case $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$

As might be expected, the classification rules are more complicated when the population covariance matrices are unequal. For details see Johnson and Wichern (2007), pp. 593 – 595.

Remark. If the data are not multivariate normal, they can be transformed to data more nearly normal, and a test for the equality of covariance matrices can be conducted.

10.2.2 Fisher's Approach to Classification with Two Populations

Fisher's idea was to transform the multivariate observations $\boldsymbol{x}$ to univariate observations y such that the y's derived from population π_1 and π_2 were separated as much as possible. Fisher suggested taking linear combinations of $\boldsymbol{x}$ to create y's because they are simple enough functions of the $\boldsymbol{x}$ to be handled easily. Fisher's approach does not assume that the populations are normal. It does, however, implicitly assume that the population covariance matrices are equal.

Proposition 10.2.4. *The linear combination* $\hat{y} = \hat{\boldsymbol{a}}'\boldsymbol{x} = (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)'\mathbf{S}_{pooled}^{-1}\boldsymbol{x}$ *maximizes the ratio*

$$\frac{\textit{squared distance between sample means of } y}{\textit{sample variance of } y} = \frac{(\overline{y}_1 - \overline{y}_2)^2}{s_y^2} = \frac{(\hat{\boldsymbol{a}}'\overline{\boldsymbol{x}}_1 - \hat{\boldsymbol{a}}'\overline{\boldsymbol{x}}_2)^2}{\hat{\boldsymbol{a}}'\mathbf{S}_{pooled}\hat{\boldsymbol{a}}}$$

over all possible coefficient vectors $\hat{\boldsymbol{a}}$ *(see Figure 10.3). The maximum of the ratio is*

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

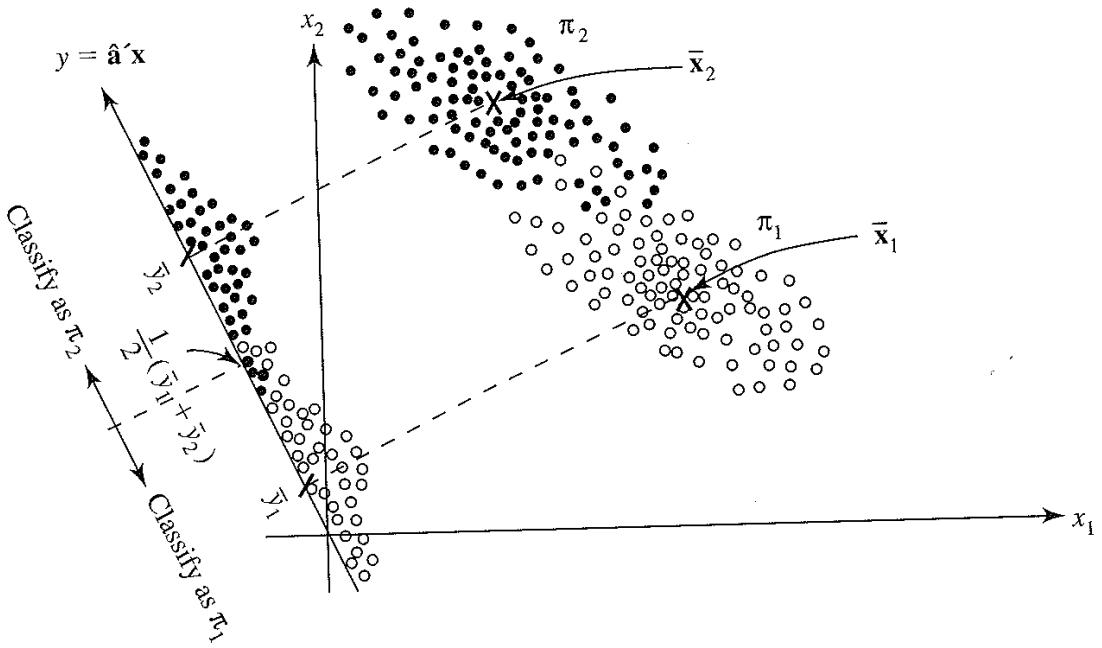

Figure 10.3: Fisher's procedure for two populations with $p = 2$. Source: Johnson and Wichern (2007).

Proposition 10.2.5. *An allocation rule based on Fisher's discriminant function: Allocate* $\boldsymbol{x}_0$ *to* π_1 *if*

$$\begin{aligned}\hat{y}_0 &= (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)' \mathbf{S}_{pooled}^{-1} \boldsymbol{x}_0 \\ &\geq \hat{m} = 1/2 (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)' \mathbf{S}_{pooled}^{-1} (\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2).\end{aligned}$$

Allocate $\boldsymbol{x}_0$ *to* π_2 *if*

$$\hat{y}_0 < \hat{m}.$$

Example (El Niño, p. 10-2)**.** In Figure 10.4 the Fisher discriminant function is shown for the pairs of variables pressure (Pres), temperature (Temp) and precipitation (Prec), temperature (Temp), respectively. There is one misclassification, namely the year 1953.

Remark. The performance of any classification procedure should always be checked. Ideally, there will be enough data available to provide for "training" samples and "validation" samples. The training samples can be used to develop the classification function, and the validation samples can be used to evaluate its performance.

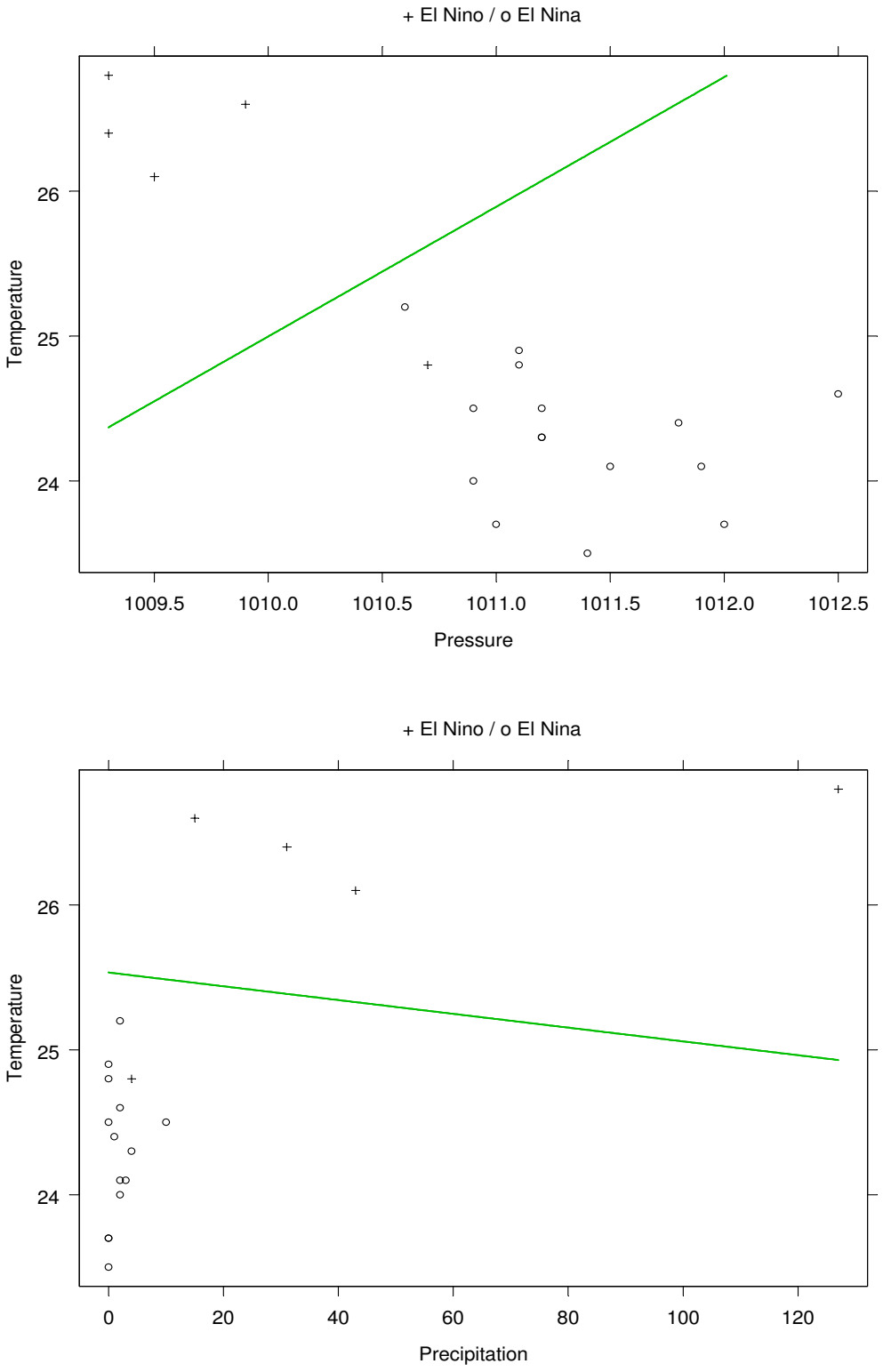

Figure 10.4: Discriminant function of the data for Guayaquil. Data set: Table 10.1, p. 10-2.