MULTIVARIATE ANALYSIS

© Richard L. Smith

Department of Statistics
University of North Carolina
Chapel Hill, NC 27599-3260

Email address: rls@email.unc.edu

VERSION 1.0

11 MAY 1999

PREFACE

These notes have been prepared in conjunction with the course Statistics 133, which I taught in Spring 1999. They are intended as informal notes introducing some of the theory of multivariate analysis and also showing how some of the better-known multivariate analysis techniques may be implemented in SPlus. The material draws on a number of standard references in the field; among those which I have used particularly heavily are Mardia, Kent and Bibby (1979), and Chatfield and Collins (1980).

Beginning with the 2000-2001 academic year, Statistics 133 is to become Statistics 185 with the new name "Time Series and Multivariate Analysis". A separate set of course notes is available for the "Time Series" section of Statistics 133.

Richard Smith
Chapel Hill
May 1999

TABLE OF CONTENTS

1. MULTIVARIATE NORMAL DISTRIBUTION THEORY

1.1 Definitions

Definition 1. A p-dimensional random vector X has a *multivariate normal distribution* with mean vector μ and nonsingular $p \times p$ covariance matrix Σ (notation: $X \sim MVN_p[\mu, \Sigma]$) if X has the density

$$f_X(x) = (2\pi)^{-p/2}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}. \tag{1.1}$$

Definition 2. $X \sim MVN_p[\mu, \Sigma]$ if and only if X has the characteristic function

$$\mathrm{E}\left(e^{it^TX}\right) = \exp\left(it^T\mu - \frac{1}{2}t^T\Sigma t\right). \tag{1.2}$$

Note that (1.1) makes sense only if Σ is nonsingular, whereas (1.2) is valid without any such restriction. For this reason, some people like to think of (1.2) as the correct definition. That (1.1) implies (1.2), when Σ is nonsingular, is shown by the following argument.

First, since Σ is a covariance matrix, it is necessarily symmetric and non-negative definite. Since we are assuming it is nonsingular, it must therefore be positive definite as well, which implies that Σ^{-1} exists and has a symmetric positive-definite square root matrix A, satisfying $A^2 = \Sigma^{-1}$. Let $Y = A(X - \mu)$. It is easily verified that the data transformation from X to Y has Jacobian $|A|$ (determinant of A), which is the same as $|\Sigma|^{-1/2}$, and that $(X-\mu)^T\Sigma^{-1}(X-\mu) = Y^TY$. Therefore, the density of Y is

$$f_Y(y) = (2\pi)^{-p/2}\exp\left(-\frac{1}{2}y^Ty\right). \tag{1.3}$$

If we write $Y = (Y_1, ..., Y_p)^T$ with corresponding values $y = (y_1, ..., y_p)^T$, then (1.3) is the same as

$$\prod_{j=1}^{p}\left\{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y_j^2\right)\right\}$$

which confirms that $Y_1, ..., Y_p$ are independent $N(0, 1)$ random variables. Hence it follows that for any $t = (t_1, ..., t_p)^T$,

$$\mathrm{E}(e^{it^TY}) = \mathrm{E}\{\exp(i\sum t_jY_j)\} = \prod_{j=1}^{p}\exp\left(-\frac{t_j^2}{2}\right) = \exp\left(-\frac{t^Tt}{2}\right).$$

However, writing $X = \mu + A^{-1}Y$, we have

$$\begin{aligned}\mathrm{E}\{e^{it^TX}\} &= e^{it^T\mu}\cdot\mathrm{E}\{e^{it^TA^{-1}Y}\}\\ &= \exp\left(it^T\mu - \frac{1}{2}t^TA^{-1}A^{-1}t\right)\\ &= \exp\left(it^T\mu - \frac{1}{2}t^T\Sigma t\right),\end{aligned}$$

consistent with (1.2).

Proposition 1. If $X \sim MVN_p[\mu, \Sigma]$ and $Y = AX + b$, where A is $q \times p$, b is $q \times 1$, then $Y \sim MVN_q[A\mu + b, A\Sigma A^T]$.

Proof. We use definition 2 of the MVN, since in many cases one of Σ or $A\Sigma A^T$ will be singular. But then

$$\begin{aligned}
\mathrm{E}\{e^{it^T Y}\} &= \mathrm{E}\{\exp(it^T AX + it^T b)\} \\
&= \exp\left(it^T A\mu - \frac{1}{2} t^T A\Sigma A^T t + it^T b\right) \\
&= \exp\left\{it^T (A\mu + b) - \frac{1}{2} t^T (A\Sigma A^T) t\right\},
\end{aligned}$$

which is as required.

Corollary 1. If $X \sim MVN_p[\mu, \Sigma]$ and $a \in \mathcal{R}^p$, then $a^T Y \sim N(a^T \mu, a^T \Sigma a)$.

Remark 1. The converse statement also holds, i.e. if $a^T X$ is (univariate) normal for any vector a, then X is multivariate normal. This is sometimes taken as an alternative definition of the multivariate normal distribution.

Proposition 2. If $X \sim MVN_p[\mu, \Sigma]$ with nonsingular Σ, and $U = (X - \mu)^T \Sigma^{-1} (X - \mu)$, then $U \sim \chi^2_p$.

Proof. Defining $Y = A(X - \mu)$ as before, we have $U = Y^T A^{-1} \Sigma^{-1} A^{-1} Y = Y^T Y = \sum_{j=1}^p Y_j^2$, and it is well known that the sum of squares of p independent standard normal variables has a χ^2_p distribution.

To avoid constantly repeating statements of the form "assuming the matrix Σ is nonsingular", we shall henceforth adopt the convention that whenever a definition or a proposition involves the inverse of a matrix, it is part of the statement that the inverse is assumed to exist.

Proposition 3. Suppose

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim MVN_p \left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right].$$

In other words, X is being partitioned into two vectors X_1 and X_2, whose dimensions are p_1 and p_2 say, with $p_1 + p_2 = p$, and μ and Σ are partitioned correspondingly. Then the conditional distribution of X_1 given $X_2 = x_2$ is

$$MVN_{p_1}[\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2),\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}].$$

Proof. Define $X_1^* = X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2$, $X_2^* = X_2$,

$$X^* = \begin{pmatrix} X_1^* \\ X_2^* \end{pmatrix} = AX, \qquad A = \begin{pmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{pmatrix},$$

where I is the identity matrix.

By Prop. 1, X^* is multivariate normal with mean

$$A\mu = \begin{pmatrix} \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2 \\ \mu_2 \end{pmatrix}$$

and covariance matrix

$$\begin{aligned} A\Sigma^{-1}A^T &= \begin{pmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} I & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21} & I \end{pmatrix} \\ &= \begin{pmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & \Sigma_{12} \\ 0 & \Sigma_{22} \end{pmatrix} \\ &= \begin{pmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}. \end{aligned}$$

This shows that X_1^* and X_2^* are uncorrelated, and hence *independent* with

$$\begin{aligned} X_1^* &\sim MVN_{p_1}[\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2,\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}] \\ X_2^* &\sim MVN_{p_2}[\mu_2,\ \Sigma_{22}]. \end{aligned}$$

To calculate the conditional distribution of X_1 given X_2, write

$$X_1 = X_1^* + \Sigma_{12}\Sigma_{22}^{-1}X_2.$$

The second term is constant while the first is independent of X_2 and therefore has the same distribution conditionally as it does unconditionally. This distribution is normal, hence X_1 is conditionally normal with conditional mean $\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2 + \Sigma_{12}\Sigma_{22}^{-1}X_2$ and covariance $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. This proves the result.

1.2 The Wishart Distribution

Definition 3. Let $X_1, ..., X_m$ be independent $MVN_p[0, \Sigma]$ and $M = \sum_{j=1}^m X_j X_j^T$. The M is said to have the Wishart distribution with m degrees of freedom and covariance matrix Σ, notation $M \sim W_p[\Sigma, m]$. The density is

$$f_M(M) = \frac{|M|^{(m-p-1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}M)\right\}}{2^{mp/2}\pi^{p(p-1)/4}|\Sigma|^{m/2}\prod_{j=1}^p \Gamma\left(\frac{m+1-j}{2}\right)}$$

with respect to Lebesgue measure on $\mathcal{R}^{p(p+1)/2}$, restricted to positive definite symmetric matrices M.

The case $\Sigma = I_p$ (the $p \times p$ identity matrix) is known as *standard Wishart.*

The precise form of the probability density function is very rarely used, the one exception to this statement being in Bayesian calculations where the Wishart distribution is often used as a conjugate prior for the inverse of a normal covariance matrix. Even in this case, for applied Bayesian calculations, it is not necessary to remember the constants needed to make the density integrate to 1.

Although the Wishart density is defined for any real positive m, in practice we usually restrict attention to cases when m is integer, and in that case, most of the important properties are derived from Definition 3 rather than from the form of the density function.

Proposition 4. If $M \sim W_p[\Sigma, m]$ and B is $p \times q$, then $B^T M B \sim W_q[B^T \Sigma B, m]$.

Proof. If $X_j \sim MVN_p[0, \Sigma]$ then $B^T X_j \sim MVN_q[0, B^T \Sigma B]$ by Prop. 1, and $B^T M B = \sum_j (B^T X_j)(B^T X_j)^T$. The result is then immediate from Definition 3.

Corollary 2. If $M \sim W_p[\Sigma, m]$ then $\Sigma^{-1/2} M \Sigma^{-1/2} \sim W_p[I_p, m]$.

Proposition 5. If $M \sim W_1[\sigma^2, m]$, then $M/\sigma^2 \sim \chi^2_m$.

Proof. $M/\sigma^2 = \sum_j (X_j/\sigma)^2$, which is the sum of squares of m independent $N[0,1]$ variables.

Corollary 3. If $M \sim W_p[\Sigma, m]$ and $a \in \mathcal{R}^p$ is such that $a^T \Sigma a \neq 0$, then

$$\frac{a^T M a}{a^T \Sigma a} \sim \chi^2_m.$$

Proof. Immediate from Prop. 4 and Prop. 5.

Note that it follows from Corollary 3 that all diagonal entries of a Wishart matrix have scaled χ^2_m distributions.

Up to this point, all the properties we have derived have been more or less immediate consequences of the definitions. Now, however, comes something much more subtle:

Proposition 6. If $M \sim W_p[\Sigma, m]$ and $a \in \mathcal{R}^p$ and $m > p - 1$, then

$$\frac{a^T \Sigma^{-1} a}{a^T M^{-1} a} \sim \chi^2_{m-p+1}. \qquad (1.4)$$

Before proving Prop. 6, we note

Proposition 7. Suppose we partition both A and A^{-1} as

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \qquad A^{-1} = \begin{pmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{pmatrix}.$$

Then the elements of A^{-1} may be given by

$$\begin{aligned} A^{11} &= (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}, \\ A^{12} &= -A^{11}A_{12}A_{22}^{-1}, \\ A^{21} &= -A_{22}^{-1}A_{21}A^{11}, \\ A^{22} &= (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}, \end{aligned}$$

assuming (as in earlier results) that all the inverse matrices given in these formulae actually exist.

The proof of Proposition 7 is essentially just direct verification by multiplying out the matrices and some algebraic rearrangement of the expressions that result. We omit all details of this.

Proof of Proposition 6. We divide this into two cases, doing it first for the case when $a^T = (1, 0, 0, 0..., 0)$, and then in general. The proof follows that given by Chatfield and Collins (1980), which reduces to showing that the result is equivalent to a standard one about least squares regression.

Suppose, then, $a^T = (1, 0, 0, 0..., 0)$. Partitioning both Σ and M by the first row and column, we write

$$\Sigma = \begin{pmatrix} \sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \qquad \Sigma^{-1} = \begin{pmatrix} \sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{pmatrix},$$

$$M = \begin{pmatrix} m_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}, \qquad M^{-1} = \begin{pmatrix} m^{11} & M^{12} \\ M^{21} & M^{22} \end{pmatrix},$$

where by Prop. 7,

$$\sigma^{11} = \frac{1}{\sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}, \qquad m^{11} = \frac{1}{m_{11} - M_{12}M_{22}^{-1}M_{21}}.$$

So in this case, the result reduces to showing

$$\frac{\sigma^{11}}{m^{11}} = \frac{m_{11} - M_{12}M_{22}^{-1}M_{21}}{\sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}} \sim \chi^2_{m-p+1}. \tag{1.5}$$

Suppose $X_i,\ i = 1, 2, ..., m$ are a sample of independent observations from $N(0, \Sigma)$. Write $X_i{=}(X_i^{(1)}, ..., X_j^{(p)})^T$ so that, in particular, $X_i^{(1)}$ is the first component of the vector

X_i. To prove (1.5), we proceed by *conditioning* on all the other components, i.e. on $\{X_i^{(j)},\ 2 \le j \le p,\ 1 \le i \le m\}$. With that conditioning, we may write

$$X_i^{(1)} = \sum_{j=2}^{p} \beta_j X_i^{(j)} + \epsilon_i \tag{1.6}$$

where $\{\epsilon_i\}$ are independent $N(0, \tau^2)$ random variables and $\{\beta_j,\ 2 \le j \le p\}$ are the regression coefficients of $X_i^{(1)}$ on $\{X_i^{(j)},\ 2 \le j \le p\}$. Prop. 3 shows that $\beta = (\beta_2, ..., \beta_p)^T = \Sigma_{12}\Sigma_{22}^{-1}$ and $\tau^2 = \sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. Suppose, however, we were interested in *estimating* β based on the regression equation (1.6). We form the standard sums of squares and products of the dependent and independent variables without centering, since we are assuming the overall mean of all the observations is 0. Writing (1.6) in the form $y = A\beta + \epsilon$, a typical entry of the matrix $A^T A$ is of the form

$$\sum_{i=1}^{m} X_i^{(j)} X_i^{(k)},\ 2 \le j \le p,\ 2 \le k \le p, \tag{1.7}$$

and the collection of all such entries is precisely the matrix M_{22}, the lower $(p-1) \times (p-1)$ submatrix of M. Similarly, the vector that corresponds to $A^T y$ in the regression model (1.6) is the submatrix M_{21}. Finally, the regression sum of squares $y^T y$ or $\sum_{i=1}^{m} X_i^{(1)^2}$ is just m_{11}. Therefore, in the regression equation (1.6), the least squares estimate of β is $M_{22}^{-1} M_{21}$, and the residual sum of squares is $m_{11} - M_{12}M_{22}^{-1}M_{21}$. Noting that we are estimating a $(p-1)$-dimensional parameter vector based on m observations, standard least squares regression theory shows that

$$\frac{m_{11} - M_{12}M_{22}^{-1}M_{21}}{m - p + 1}$$

is an unbiased estimator of τ^2, with $m - p + 1$ degrees of freedom, and a scaled χ^2_{m-p+1} distribution. In particular,

$$\frac{m_{11} - M_{12}M_{22}^{-1}M_{21}}{\tau^2} \sim \chi^2_{m-p+1}.$$

However, this result is precisely (1.5), expressed as a *conditional* distribution given $\{X_i^{(j)},\ 2 \le j \le p,\ 1 \le i \le m\}$.

The form of this result is that the conditional distribution of $m_{11} - M_{12}M_{22}^{-1}M_{21}$ does not depend on the values we are conditioning on, and therefore holds unconditionally as well. With that, we have proved (1.5) in the form originally stated, and hence have proved Prop. 6 in the case that $a^T = (1, 0, 0, 0..., 0)$.

To complete the proof of Prop. 6, we must show that the same result holds for any other a as well. Let A be a $p \times p$ nonsingular matrix whose first column is a. Provided

$a \neq 0$, we can always find such a matrix by taking the last $p-1$ columns of A to be linearly independent vectors in the space orthogonal to a. Let $B = A^{-1}$. By Prop. 4, $BMB^T \sim W_p[B\Sigma B^T, m]$. Now, $(BMB^T)^{-1} = A^T M^{-1} A$ and the top left hand entry of this matrix is precisely $a^T M^{-1} a$. Similarly, $(B\Sigma B^T)^{-1} = A^T \Sigma^{-1} A$ with top left hand entry $a^T \Sigma^{-1} a$. Therefore, by the result for the case $a^T = (1, 0, 0, 0..., 0)$, applied to the matrix BMB^T, we deduce (1.4), which is the result we want. With this, the proof of Prop. 6 is complete.

Remark 2. The result of Prop. 6 holds for any deterministic $a \in \mathcal{R}^p$, and therefore holds also for random a provided the distribution of a is independent of M. This note is critical in the next subsection.

1.3 Hotelling's T^2

Definition 4. Suppose X and S are independent such that

$$X \sim MVN_p[\mu, \Sigma], \qquad mS \sim W_p[\Sigma, m].$$

Then

$$T_p^2(m) = (X-\mu)^T S^{-1} (X-\mu)$$

is known as *Hotelling's T^2 statistic.* We shall see in Chapter 2 that this plays a similar role in multivariate analysis to that of the student's t distribution in univariate statistical theory.

The next proposition states the distribution of T^2:

Proposition 8.

$$\frac{m-p+1}{mp} T_p^2(m) \sim F_{p,m-p+1}, \quad \text{provided } m > p-1.$$

Proof. Define $M = mS$.

By Remark 2 following Prop. 6, the result of Prop. 6 holds if we replace the deterministic vector a by the random vector $X - \mu$, whose distribution is independent of M by assumption. Therefore, defining

$$R = m \cdot \frac{(X-\mu)^T \Sigma^{-1} (X-\mu)}{(X-\mu)^T S^{-1} (X-\mu)},$$

we see that $R \sim \chi^2_{m-p+1}$, independent of X. We also have

$$U = (X-\mu)^T \Sigma^{-1} (X-\mu) \sim \chi^2_p,$$

by Prop. 2. Therefore,

$$T_p^2(m) = \frac{mU}{R}$$

with R, U having independent χ^2 distributions. Finally

$$\frac{m-p+1}{mp} T_p^2(m) = \left(\frac{U}{p}\right) \Big/ \left(\frac{R}{m-p+1}\right) \sim F_{p,m-p+1},$$

as required.

1.4 The joint distribution of the sample mean and the sample covariance matrix

Suppose $X_1, ..., X_n$ are independent random vectors with the common distribution $MVN_p[\mu, \Sigma]$. Let $\bar{X} = (X_1 + ... + X_n)/n$.

Proposition 9.

$$\sum_i (X_i - \mu)(X_i - \mu)^T = n(\bar{X} - \mu)(\bar{X} - \mu)^T + \sum_i (X_i - \bar{X})(X_i - \bar{X})^T.$$

Proof. $\sum_i (X_i - \mu)(X_i - \mu)^T = \sum_i (X_i - \bar{X} + \bar{X} - \mu)(X_i - \bar{X} + \bar{X} - \mu)^T = \sum_i (X_i - \bar{X})(X_i - \bar{X})^T + 2\sum_i (\bar{X} - \mu)(X_i - \bar{X})^T + n(\bar{X} - \mu)(\bar{X} - \mu)^T$. However the middle term in the expansion is 0, giving the result.

Define the sample covariance matrix,

$$S = \frac{1}{n-1} \sum_i (X_i - \bar{X})(X_i - \bar{X})^T.$$

Proposition 10. $\bar{X}$ and S are independent, with

$$\sqrt{n}(\bar{X} - \mu) \sim MVN_p[0, \Sigma], \tag{1.8}$$

$$(n-1)S \sim W_p[\Sigma, n-1]. \tag{1.9}$$

Proof. $\mathrm{E}\{n(\bar{X} - \mu)(\bar{X} - \mu)^T\} = \frac{1}{n}\mathrm{E}\ \{\sum_i (X_i - \mu)(X_i - \mu)^T\} = \Sigma$, so (1.8) follows directly from Prop. 1. The main part of the proof is to show that $(n-1)S$ has a Wishart distribution *and* is independent of $\bar{X}$.

Let $D = (d_{ij})$ be a $n \times n$ orthogonal matrix ($D^T D = DD^T = I_n$, where I_n is the $n \times n$ identity matrix) such that $d_{i1} = 1/\sqrt{n}$ for $i = 1, ..., n$. Define $Y_i = \sum_j (X_j - \mu) d_{ji}$. We claim:

(i) $\mathrm{E}\{Y_i Y_k^T\} = \Sigma$ if $i = k$, 0 otherwise.

(ii) $\sum_i Y_i Y_i^T = \sum_i (X_i - \mu)(X_i - \mu)^T$.

(iii) $Y_1 Y_1^T = n(\bar{X} - \mu)(\bar{X} - \mu)^T$.

Note that (i) implies that $Y_1, ..., Y_n$ are independent, since for random variables which are jointly normally distributed, uncorrelated implies independent.

Proof of (i). $\mathrm{E}\{\sum_j \sum_\ell (X_j - \mu) d_{ji} (X_\ell - \mu)^T d_{\ell k}\} = \Sigma \cdot \sum_j d_{ji} d_{jk}$. But by orthogonality of D, $\sum_j d_{ji} d_{jk}$ is 1 if $i = k$, 0 otherwise, as required.

Proof of (ii). $\sum_i \sum_j \sum_k (X_j - \mu) d_{ji} (X_k - \mu)^T d_{ki} = \sum_j \sum_k (X_j - \mu)(X_k - \mu)^T \cdot$ $\cdot \{\sum_i d_{ji} d_{ki}\}$ which reduces to $\sum_j (X_j - \mu)(X_j - \mu)^T$, again using orthogonality of D.

Proof of (iii). Immediate by substitution.

Therefore, we may write

$$\begin{aligned}\sum_i (X_i - \mu)(X_i - \mu)^T &= Y_1 Y_1^T + \sum_{i=2}^{n} Y_i Y_i^T \\ &= n(\bar{X} - \mu)(\bar{X} - \mu)^T + \sum_{i=2}^{n} Y_i Y_i^T\end{aligned}$$

and the two summands are therefore *independent* with distributions $W_p[\Sigma, 1]$ and $W_p[\Sigma, n-1]$ respectively. Moreover, by Prop. 9, $\sum_{i=2}^n Y_i Y_i^T$ is the same as $(n-1)S$. With this, the proof of Prop. 10 is complete.

2. INFERENCE ABOUT THE MULTIVARIATE NORMAL DISTRIBUTION

2.1 Point estimates

Suppose $X_1, ..., X_n$ are independent random vectors with common distribution $MVN_p[\mu, \Sigma]$. In this chapter, we consider estimation and hypothesis testing problems associated with μ and Σ.

A natural starting place is to consider the maximum likelihood estimators for this problem. Ignoring the $(2\pi)^{-p/2}$ component of the probability density function, which plays no part in any inferential procedures, the log likelihood is given by

$$\begin{aligned}\ell(\mu,\Sigma) &= -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_i (X_i-\mu)^T\Sigma^{-1}(X_i-\mu) \\ &= -\frac{n}{2}\log|\Sigma| - \frac{n}{2}(\bar{X}-\mu)^T\Sigma^{-1}(\bar{X}-\mu) - \frac{1}{2}\sum_i (X_i-\bar{X})^T\Sigma^{-1}(X_i-\bar{X}). \quad (2.1)\end{aligned}$$

The middle term is ≤ 0, and $= 0$ if and only if $\mu = \bar{X}$. Therefore, the maximum likelihood estimator of μ is $\hat{\mu} = \bar{X}$.

The third term in (2.1) may be written in the form $-\frac{1}{2}\text{tr}\{\sum_i (X_i-\bar{X})^T\Sigma^{-1}(X_i-\bar{X})\} = -\frac{1}{2}\text{tr}\{\Sigma^{-1}\sum_i (X_i-\bar{X})(X_i-\bar{X})^T\} = -\frac{n-1}{2}\text{tr}\{\Sigma^{-1}S\}$, using the formula tr$(AB)$ =tr(BA) whenever AB and BA are both well-defined square matrices. However, it is more convenient for our present purposes to write this as $-\frac{n}{2}\text{tr}\{\Sigma^{-1}S_0\}$, where

$$S_0 = \frac{\sum_{i=1}^n (X_i-\bar{X})(X_i-\bar{X})^T}{n},$$

in other words, using n rather than $n-1$ in the denominator.

With these notational conventions, we now see that the maximum likelihood estimator of Σ is the positive-definite symmetric matrix which minimizes

$$\log|\Sigma| + \text{tr}(\Sigma^{-1}S_0). \quad (2.2)$$

We claim this is achieved by setting $\Sigma = S_0$.

To see this, write $\Sigma^{-1}S_0 = A$ and use to formula $|AB| = |A|\cdot|B|$ to write (2.2) as

$$\log|S_0| - \log|A| + \text{tr}(A).$$

But if the eigenvalues of A are $\lambda_1, ..., \lambda_p$, we have $|A| = \prod_i \lambda_i$ and $\text{tr}(A) = \sum_i \lambda_i$ so the problem becomes to minimize

$$\sum_{i=1}^p (\lambda_i - \log\lambda_i). \quad (2.3)$$

However it is readily checked that the expression $f(\lambda) = \lambda - \log \lambda$ is minimized over $0 < \lambda < \infty$ by $\lambda = 1$. Hence the solution to (2.3) is achieved when $\lambda_1 = ... = \lambda_p = 1$, or in other words when A is the identity matrix, which is what we set out to establish. Hence, S_0 is the maximum likelihood estimator of Σ.

We have shown that $\bar{X}$ and S_0 are the maximum likelihood estimators of μ and Σ respectively, and this means that they are asymptotically efficient estimators, but they do not necessarily have optimal small-sample properties. In the case of Σ, we have shown that S_0 is the maximum likelihood estimator but we know from Prop. 10 that S rather than S_0 is an unbiased estimator, and for that reason most statisticians in practice use S rather than S_0 in defining the sample covariance matrix. Nevertheless, the likelihood function plays an important role in defining estimators and tests for more complicated multivariate analysis problems, so it is important to understand how to calculate the maximum likelihood estimator for this problem.

This discussion is of course directly analogous to the corresponding discussion in univariate problems, so it should come as no surprise that there is a conflict between the MLE and the best unbaised estimator. What is more surprising is that in multivariate problems there is also a debate over the appropriate estimator for μ. In a series of papers in the 1950s, Charles Stein showed that the MLE is *inadmissible* with respect to mean squared error, whenever $p \geq 3$. In other words, there are other estimators which achieve a smaller expected mean squared error, whatever the true value of μ. This discovery led to a rich field of research on improved estimators in very wide classes of multivariate problems, but these ideas are not very easy to apply in practice, and in any case none of this theory impinges on the *asymptotic* properties of maximum likelihood estimators — MLEs are asymptotically efficient for virtually all of the problems we shall consider. Therefore, the discussion in these notes will continue to focus around maximum likelihood techniques. The reader interested in finding out more about the Stein effect and inadmissibility should read the celebrated paper by James and Stein (1961), or for a modern account, Berger (1985).

2.2 Testing hypotheses via likelihood ratios

There are two very widely used approaches to hypothesis testing, (a) likelihood ratio tests, (b) the union-intersection principle.

We first consider likelihood ratio tests. The general principle is as follows: suppose H_0 is the null hypothesis and H_1 is the alternative hypothesis, where the hypotheses are "nested" in the sense that H_0 is contained within H_1. More specifically, suppose H_0 is obtained from H_1 by placing ν independent parameter constraints on H_1, so that ν is the difference in dimensions between the two models. Let L_0 and L_1 be the maximized likelihood under H_0 and H_1 respectively; we should always have $L_1 \geq L_0$ because of the nesting. Let $W = 2\log(L_1/L_0)$. Then W is the likelihood ratio statistic, also known as Wilks' statistic. In some cases, we can figure out the exact distribution of W when H_0 is true. In other cases, it is usual to use the asymptotic result $W \sim \chi^2_\nu$ under H_0, which is true as the sample size $n \to \infty$ in any regular parametric problem.

Having outlined the general principle, we now consider three specific examples:

Case I. Testing $H_0 : \ \mu = \mu_0$ against $H_1 : \ \mu \neq \mu_0$, when μ_0 is given and the covariance matrix Σ is assumed known.

In this case the log likelihood (modulo constants) is of form

$$-\frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum(X_i - \mu)^T\Sigma^{-1}(X_i - \mu). \tag{2.4}$$

Under H_0, we simply substitute $\mu = \mu_0$ in (2.4). Under H_1, the MLE for μ is $\bar{X}$, so we substitute this value in (2.4). We then have that

$$\begin{aligned} W &= (X_i - \mu_0)^T\Sigma^{-1}(X_i - \mu_0) - (X_i - \bar{X})^T\Sigma^{-1}(X_i - \bar{X}) \\ &= n(\bar{X} - \mu_0)^T\Sigma^{-1}(\bar{X} - \mu_0). \end{aligned} \tag{2.5}$$

The asymptotic theory says that as $n \to \infty$, the distribution of W approaches χ^2_p. In fact, Prop. 2 shows that this is the exact distribution which holds for all n.

Case II. Testing $H_0 : \ \mu = \mu_0$ against $H_1 : \ \mu \neq \mu_0$, when μ_0 is given and the covariance matrix Σ is assumed unknown.

In this case, maximum likelihood estimators for both μ and Σ must be substituted in (2.4). Under H_1, the MLEs are $\hat{\mu} = \bar{X}$ and $\hat{\Sigma} = S_0$ as in section 2.1. Under H_0, $\mu = \mu_0$ and the MLE of Σ is $\sum(X_i - \mu_0)(X_i - \mu_0)^T/n$ by the same argument as in section 2.1. However, Prop. 9 shows that this may be written in the form $S_0 + dd^T$ where $d = \sqrt{n}(\bar{X} - \mu_0)$. Moreover, the second term in (2.4) under H_1 is $-\frac{1}{2}\sum_i(X_i - \bar{X})^T S_0^{-1}(X_i - \bar{X}) = -\frac{1}{2}\mathrm{tr}\{S_0^{-1}\sum_i(X_i - \bar{X})(X_i - \bar{X})^T\} = -\frac{n}{2}\mathrm{tr}\{S_0^{-1}S_0\} = -\frac{np}{2}$, and the identical calculation holds under H_0. Therefore, the second term in (2.4) cancels in calculating W, leaving us with

$$\begin{aligned} W &= n\log|S_0 + dd^T| - n\log|S_0| \\ &= n\log(1 + d^T S_0^{-1} d), \end{aligned} \tag{2.6}$$

where in going from the first line in (2.6) to the second, we have used the determinantal identity $|I_p + BC| = |I_n + CB|$, which holds whenever B is a $p \times n$ matrix and C is a $n \times p$ matrix. In this case we have applied this identity with $p = 1$, $B = d^T S_0^{-1/2}$, $C = S_0^{-1/2}d$, $S_0^{-1/2}$ being the positive-definite symmetric square root of S_0^{-1}, which exists because S_0 is itself symmetric and positive definite with probability 1.

The conclusion of (2.6) is that the likelihood ratio test depends on the value of $d^T S_0^{-1} d$. However, except for the change of normalizing constant from $n - 1$ to n, this is the same as

$$T^2 = n(\bar{X} - \mu_0)S^{-1}(\bar{X} - \mu_0),$$

Hotelling's T^2 statistic which was defined in Chapter 1. Therefore, the conclusion of the likelihood ratio argument is that we should reject H_0 if T^2 is greater than some critical value c. Prop. 8 combined with Prop. 10 shows that under H_0,

$$\frac{n-p}{(n-1)p}T^2 \sim F_{p,n-p}, \tag{2.7}$$

so (2.7) may be used in determining the critical value c corresponding to a given size of test α.

Remark 3. The final result may also be written in the form

$$\frac{n-p}{p}(\bar{X}-\mu_0)^T S_0^{-1}(\bar{X}-\mu_0) \sim F_{p,n-p} \quad \text{under } H_0,$$

which is preferred by some authors, e.g. Mardia, Kent and Bibby (1979), p. 126.

Case III. Testing $H_0 : \Sigma = \Sigma_0$ against $H_1 : \Sigma \neq \Sigma_0$, when Σ_0 is given and μ is unknown. (The case where μ is known is very similar and not worth treating as a separate case.)

In this case, if we substitute the MLE $\bar{X}$ for μ, we get the profile log likelihood for Σ (ignoring constants)

$$\ell(\hat{\mu},\Sigma) = -\frac{n}{2}\log|\Sigma| - \frac{n}{2}\mathrm{tr}(\Sigma^{-1}S_0). \tag{2.8}$$

Under H_0, we simply substitute $\Sigma = \Sigma_0$ into (2.8). Under H_1, we estimate $\hat{\Sigma} = S_0$ as shown in section 2.1, leading to

$$\ell(\hat{\mu},\hat{\Sigma}) = -\frac{n}{2}\log|S_0| - \frac{n}{2}\mathrm{tr}(S_0^{-1}S_0) = -\frac{n}{2}\log|S_0| - \frac{np}{2}.$$

Therefore, the likelihood ratio statistic in this case reduces to

$$W = -n\log|\Sigma_0^{-1}S_0| - np + n\ \mathrm{tr}(\Sigma_0^{-1}S_0).$$

However, if $\lambda_1, ..., \lambda_p$ are the eigenvalues of $\Sigma_0^{-1}S_0$, then by the same reasoning as in section 2.1, we get

$$W = n\sum_{i=1}^{n}(\lambda_i - 1 - \log\lambda_i). \tag{2.9}$$

Equation (2.9) shows that the exact distribution of the likelihood ratio statistic may in principle be calculated as an exercise in the distribution of the eigenvalues of $\Sigma_0^{-1}S_0$ where $nS_0 \sim W_p[\Sigma_0, n-1]$. The distribution does not depend on Σ_0, because by matrix similarity, the eigenvalues of $n\Sigma_0^{-1}S_0$ are the same as those of $n\Sigma_0^{-1/2}S_0\Sigma_0^{-1/2}$, whose distribution is $W_p[I_p, n-1]$ by Corollary 2 of chapter 1. In practice, however, the distribution is rather complicated so a practical alternative is to use the asymptotic theory, approximating the null distribution of W by χ^2_ν, where $\nu = p(p+1)/2$ is the number of free parameters in Σ.

2.3 The union-intersection principle

The basic idea is illustrated by Case I from section 2.2, i.e. test $H_0: \ \mu = \mu_0$ assuming Σ known.

For any $a \in \mathcal{R}^p$ such that $a^T a = 1$, we can consider the directional hypothesis $H_0(a)$: $a^T \mu = a^T \mu_0$. The real hypothesis of interest, H_0 is the *intersection* of all the univariate hypotheses $H_0(a)$.

Suppose z_a^2 is some (squared) statistic for $H_0(a)$, such that we reject $H_0(a)$ whenever $z_a^2 > c$ for some $c > 0$. It is usual, though not essential, to try to define z_a^2 in such a way that the null distribution is the same for every a. The union-intersection principle is that we reject H_0 if $z_a^2 > c$ for at least one value of a. Thus the rejection region for H_0 is the *union* of rejection regions for the individual $H_0(a)$ hypotheses.

Of course the constant c must be adjusted to account for the simultaneous testing aspect of this procedure, and this means that the effective test statistic is $\max_a z_a^2$, with the rejection constant c at level α being chosen so that $\Pr\{\max_a z_a^2 > c\} = \alpha$ when H_0 is true.

One feature of the union-intersection test which makes it different from the likelihood ratio test is that in cases where the procedure results in rejection of H_0, we can identify the set of all a such that $z_a^2 > c$ as the set of directions in which the null hypothesis is rejected. This may be valuable for deciding what to do next, e.g. what alternative models to fit after the original hypothesis has been rejected.

With these preliminaries, let us now consider what actually happens in Case I. For fixed a, the natural procedure is to define $y_i = a^T(X_i - \mu_0)$ which, under the null hypothesis, has a normal distribution with mean 0 and variance $a^T \Sigma a$. Therefore, the natural squared test statistic is

$$z_a^2 = \frac{n\bar{y}^2}{a^T \Sigma a} = \frac{n||a^T(\bar{X} - \mu_0)||^2}{a^T \Sigma a}$$

with a null χ_1^2 distribution. However, the Cauchy-Schwartz inequality

$$||a^T(\bar{X} - \mu_0)||^2 = ||a^T \Sigma^{1/2} \cdot \Sigma^{-1/2}(\bar{X} - \mu_0)||^2 \le a^T \Sigma a \cdot (\bar{X} - \mu_0)^T \Sigma^{-1} (\bar{X} - \mu_0)$$

shows that

$$\max_{a:\ a^T a = 1} z_a^2 = n(\bar{X} - \mu_0)^T \Sigma^{-1} (\bar{X} - \mu_0). \tag{2.9}$$

Thus we end up, by a different route, with exactly the same test statistic as in (2.5), and as there, the null distribution is exactly χ_p^2.

Now let us consider the union-intersection principle applied to Case II from Section 2.2, testing $H_0: \ \mu = \mu_0$ with Σ unknown. In the case the univariate hypothesis $H_0(a)$: $a^T\mu = a^T\mu_0$ is usually handled through a student's t statistic, in squared form

$$z_a^2 = \frac{n||a^T(\bar{X} - \mu_0)||^2}{a^TSa}$$

with S the usual sample covariance matrix. By the same reasoning as above,

$$\max_{a:\ a^Ta=1} z_a^2 = n(\bar{X} - \mu_0)^T S^{-1}(\bar{X} - \mu_0),$$

which is Hotelling's T^2 statistic. In this case, therefore, we are again led to the same test statistic as for the likelihood ratio procedure.

Now consider Case III: test $H_0 \ : \ \Sigma = \Sigma_0$ with μ unknown. In this case, we get something different from the likelihood ratio procedure. The natural univariate hypothesis is $\text{Var}\{y_i\} = a^T\Sigma_0 a$ which suggests the test statistic

$$z_a^2 = (n-1)\frac{a^TSA}{a^T\Sigma_0 a},$$

whose null distribution for fixed a is χ_1^2. For a two-sided alternative hypothesis, we reject whenever z_a^2 is too small or too large. The natural union-intersection analog, then, is to define

$$z^{(1)} = \min_a z_a^2, \qquad z^{(2)} = \max_a z_a^2,$$

rejecting H_0 when $z^{(1)} \le c_1$ or $z^{(2)} \ge c_2$ for suitable c_1, c_2. If we rewrite $a = \Sigma_0^{-1/2}b$ and hence $a^TSa/a^T\Sigma_0 a = b^T\Sigma_0^{-1/2}S\Sigma_0^{-1/2}/b^Tb$, we see that the maximum and minimum of z_a^2 are attained when b is one of the eigenvectors of $\Sigma_0^{-1/2}S\Sigma_0^{-1/2}$, and in that case, equal the corresponding eigenvalues. However, the two matrices $\Sigma_0^{-1/2}S\Sigma_0^{-1/2}$ and $\Sigma_0^{-1}S$ are similar and therefore have the same eigenvalues. Therefore, the union-intersection test is to reject H_0 when

$$\min\{\lambda_1, ..., \lambda_p\} \le c_1 \text{ or } \max\{\lambda_1, ..., \lambda_p\} \ge c_2,$$

where $\lambda_1, ..., \lambda_p$ are the eigenvalues of $\Sigma_0^{-1}S$. This is different from the likelihood ratio test but has one thing in common, namely, that it depends on the eigenvalues of $\Sigma_0^{-1}S$. As with the LRT, however, the exact distribution of the test statistic is complicated.

2.4 One-way MANOVA

The principles of the LRT and UIT may be applied to a whole range of multivariate testing problems; Mardia *et al.* (1979) have many examples. To give one example of the kind of problem they can be applied to, we consider here the problem of testing for equality in several multivariate normal means with common unknown covariance matrix. This is the one-way multivariate analysis of variance problem, usually abbreviated to *MANOVA*.

Suppose we have K groups of observations and $X_{ki} \sim MVN_p[\mu_k, \Sigma]$ in the k'th group. Here X_{ki} is the i'th observation from the k'th group. We assume there are n_k observations in the k'th group and $n = n_1 + ... + n_K$ observations altogether. The most usual formulation of a null hypothesis is $H_0: \ \mu_1 = ... = \mu_K$ against the alternative H_1 that $\mu_1, ..., \mu_K$ are not all equal.

Let $\bar{X}_{k\cdot} = \sum_i X_{ki}/n_k$ denote the sample mean of the k'th group and $\bar{X}_{\cdot\cdot} = \sum_k \sum_i X_{ki}/n$ the overall mean of the observations. As with the univariate ANOVA problem, we have a decomposition

$$\begin{aligned}&\sum_k \sum_i (X_{ki} - \bar{X}_{\cdot\cdot})(X_{ki} - \bar{X}_{\cdot\cdot})^T \\ &= \sum_k n_k (\bar{X}_{k\cdot} - \bar{X}_{\cdot\cdot})(\bar{X}_{k\cdot} - \bar{X}_{\cdot\cdot})^T + \sum_k \sum_i (X_{ki} - \bar{X}_{k\cdot})(X_{ki} - \bar{X}_{k\cdot})^T,\end{aligned} \tag{2.10}$$

whose proof we omit ((2.10) generalizes Prop. 9 from Chapter 1).

We derive a test following the likelihood ratio principle. The log likelihood may be written

$$\ell = -\frac{n}{2} \log |\Sigma| - \frac{1}{2} \sum_k \sum_i (X_{ki} - \mu_k)^T \Sigma^{-1} (X_{ki} - \mu_k). \tag{2.11}$$

Under H_0, the common $\mu_1, ..., \mu_K$ are estimated by $\bar{X}_{\cdot\cdot}$ and the covariance matrix by $S_0 = \{\sum_k \sum_i (X_{ki} - \bar{X}_{\cdot\cdot})(X_{ki} - \bar{X}_{\cdot\cdot})^T\}/n$, as in the single-sample problem. Under H_1, it is fairly easy to check that μ_k is estimated by $\bar{X}_{k\cdot}$ and Σ by the pooled sample covariance matrix, $S_1 = \{\sum_k \sum_i (X_{ki} - \bar{X}_{k\cdot})(X_{ki} - \bar{X}_{k\cdot})^T\}/n$. Under both H_0 and H_1, the second term in (2.11) reduces to $-np/2$, so the likelihood ratio statistic is

$$W = n \log \frac{|S_0|}{|S_1|}.$$

However, the ANOVA decomposition (2.10) leads to

$$S_0 = S_1 + B,$$

where

$$B = \sum_k \frac{n_k}{n} (\bar{X}_{k\cdot} - \bar{X}_{\cdot\cdot})(\bar{X}_{k\cdot} - \bar{X}_{\cdot\cdot})^T.$$

Modulo constants, the LRT statistic is $|S_1 + B|/|S_1| = |1 + S_1^{-1}B|$. If $\lambda_1, ..., \lambda_p$ are the eigenvalues of $S_1^{-1}B$, the test statistic is

$$U = \prod_{i=1}^{p} (1 + \lambda_i). \tag{2.12}$$

(2.12) is equivalent to the Wilks statistic for this problem. There are, however, several other test statistics which may be defined in terms of the eigenvalues of $S_1^{-1}B$. The union-intersection procedure in this case leads to a test based on $\max\{\lambda_i\}$, also known as Roy's statistic. Other widely used test statistics are

$\sum \lambda_i$	(Lawley-Hotelling),
$\sum \lambda_i/(1+\lambda_i)$	(Pillai),
$\prod \lambda_i/(1+\lambda_i)$	(Roy-Gnanadesikan-Srivastava).

Power comparisons among these different tests do not point towards a single best test. Splus allows a choice among the Wilks, Pillai, Lawley-Hotelling and Roy procedures.

2.5 Two-way MANOVA

We consider the most basic form of two-way ANOVA procedure, where there are t treatments and c blocks and one multivariate observations for each treatment-block combination. The most usual model in this situation is

$$X_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, \quad 1 \le i \le t,\ 1 \le j \le c, \tag{2.13}$$

in which e_{ij} are independent $MVN_p[0,\Sigma]$ vectors and we impose the constraints $\sum_i \alpha_i = \sum_j \beta_j = 0$.

A standard problem in this field is to test the null hypothesis of no treatment effect in a situation which allows for the existence of a block effect. This therefore suggests the hypothesis

$$H_0:\ \alpha_1 = ... = \alpha_t = 0,$$

against the alternative H_1 that the α_i are not all equal. A basic result, exactly analogous with one-dimensional ANOVA calculations, is the identity

$$\begin{aligned}
&\sum\sum(X_{ij}-\bar{X}_{\cdot\cdot})(X_{ij}-\bar{X}_{\cdot\cdot})^T\\
&=\sum\sum\{(\bar{X}_{i\cdot}-\bar{X}_{\cdot\cdot})+(\bar{X}_{\cdot j}-\bar{X}_{\cdot\cdot})+(X_{ij}-\bar{X}_{i\cdot}-\bar{X}_{\cdot j}+\bar{X}_{\cdot\cdot})\}\cdot\\
&\quad\cdot\{(\bar{X}_{i\cdot}-\bar{X}_{\cdot\cdot})+(\bar{X}_{\cdot j}-\bar{X}_{\cdot\cdot})+(X_{ij}-\bar{X}_{i\cdot}-\bar{X}_{\cdot j}+\bar{X}_{\cdot\cdot})\}^T\\
&=c\sum_i(\bar{X}_{i\cdot}-\bar{X}_{\cdot\cdot})(\bar{X}_{i\cdot}-\bar{X}_{\cdot\cdot})^T+t\sum_j(\bar{X}_{\cdot j}-\bar{X}_{\cdot\cdot})(\bar{X}_{\cdot j}-\bar{X}_{\cdot\cdot})^T\\
&\quad+\sum_i\sum_j(X_{ij}-\bar{X}_{i\cdot}-\bar{X}_{\cdot j}+\bar{X}_{\cdot\cdot})(X_{ij}-\bar{X}_{i\cdot}-\bar{X}_{\cdot j}+\bar{X}_{\cdot\cdot})^T
\end{aligned} \tag{2.14}$$

We may write (2.14) in the form

$$T = H + B + R,$$

where T is the total sum of squares and products matrix (SSPM), H is the treatment SSPM, B is the blocks SSPM, and R is the residual SSPM. The corresponding decomposition of degrees of freedom is

$$(n-1) = (t-1) + (b-1) + (n-t-b+1).$$

Hence we have the MANOVA table:

Source	**d.f.**	**SSPM**
Treatments	$t-1$	H
Blocks	$c-1$	B
Residual	$n-t-c+1$	R
Total	$n-1$	T

The likelihood ratio test statistic in this situation reduces to

$$\Lambda = \frac{|R|}{|H+R|} = \prod_i \left(\frac{1}{1+\lambda_i} \right),$$

where $\lambda_1, ..., \lambda_p$ are the eigenvalues of $R^{-1}H$. Note that in the notation used earlier, the likelihood ratio statistic is $W = -n \log \Lambda$, but in practice, the Λ statistic (also called Wilks' Lambda) is more widely used.

However, as with one-way MANOVA, there are several alternative statistics. Each of the alternatives mentioned at the end of section 2.4 applies here as well, except that they are defined in terms of the eigenvalues of $R^{-1}H$.

Distributional approximations

Suppose H is the treatment SSPM and R the residual SSPM, with degrees of freedom h and r respectively.

Direct application of the asymptotic theory for likelihood ratio tests would imply that $W = -n \log \Lambda$ has an approximate χ^2_{hp} distribution.

The Bartlett correction is a general method of improving the convergence of a likelihood ratio statistic to its limiting χ^2 distribution. In the present context it reduces to

$$-\left(r - \frac{p-h+1}{2} \right) \log \Lambda \sim \chi^2_{hp}. \tag{2.15}$$

There is a further refinement due to C.R. Rao:

$$\frac{1-\Lambda^{1/h}}{\Lambda^{1/h}} \cdot \frac{ab-c}{ph} \sim F_{ph,ab-c}. \tag{2.16}$$

Here

$$a = r - \frac{p-h+1}{2}, \qquad b = \sqrt{\frac{p^2h^2-4}{p^2+h^2-5}}, \qquad c = \frac{ph-2}{2}.$$

In some cases, there are exact results:

$$\begin{aligned} &h = 1,\ \text{any } p, &&\frac{1-\Lambda}{\Lambda}\cdot\frac{r-p+1}{p} \sim F_{p,r-p+1},\\ &h = 2,\ \text{any } p, &&\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\cdot\frac{r-p+1}{p} \sim F_{2p,2(r-p+1)},\\ &p = 2,\ \text{any } h, &&\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\cdot\frac{r-1}{h} \sim F_{2h,2(r-1)}. \end{aligned} \tag{2.17}$$

2.6 Multivariate regression

The standard multivariate regression model is

$$Y = XB + U,$$

where Y is a $n \times p$ data matrix (in other words, each of the n rows of Y represents an independent MVN vector), X is a $n \times q$ matrix of known regressors, B is a $q \times p$ matrix of unknown coefficients, and

$$U = \begin{bmatrix} U_1^T \\ U_2^T \\ \vdots \\ U_n^T \end{bmatrix}, \qquad U_i \sim MVN_p[0,\Sigma] \ \text{(independent)}.$$

The log likelihood for this problem is

$$\ell(B,\Sigma) = -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\mathrm{tr}\{(Y-XB)\Sigma^{-1}(Y-XB)^T\}.$$

Define

$$P = I - X(X^TX)^{-1}X^T.$$

The maximum likelihood estimators are

$$\begin{aligned} \hat{B} &= (X^TX)^{-1}X^TY,\\ \hat{\Sigma} &= \frac{1}{n}Y^TPY. \end{aligned}$$

Also let $\hat{U} = Y - X\hat{B} = PY$, so that $\hat{U}\hat{U}^T = n\hat{\Sigma}$.

Distributional results:

(a) $\hat{B}$ is unbiased for B.

(b) $\mathrm{E}\{\hat{U}\} = 0$.

(c) $\hat{B}$, $\hat{U}$ are each multivariate normally distributed.

(d) $\hat{B}$ is independent of $\hat{U}$ and hence also of $\hat{\Sigma}$.

(e) $\mathrm{Cov}\{\hat{b}_{ij}, \hat{b}_{k\ell}\} = \sigma_{j\ell} g_{ik}$, where $\hat{b}_{ij}$, $\hat{b}_{k\ell}$ are generic entries of $\hat{B}$, $\sigma_{j\ell}$ is the (j, ℓ) entry of Σ and g_{ik} is the (i, k) entry of the matrix $G = (X^T X)^{-1}$.

(f) $n\hat{\Sigma} \sim W_p[\Sigma, n - q]$.

Note that, as a consequence of (f), $\hat{\Sigma}$ is biased, but $n\hat{\Sigma}/(n - q)$ is unbiased.

The proofs of (a)–(f) are omitted.

2.7 Example

Fisher's iris data is a famous data due to R.A. Fisher (1936) and subsequently discussed by many other authors, including Mardia, Kent and Bibby (1979). The data consist of the measurement of four quantitites on each of 50 iris plants of each of three types. The data from the first four plants of each type are shown in Table 2.1.

	Iris setosa				*Iris versicolour*				*Iris virginica*		
X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3	6.3	2.9	5.6	1.8
...				...				...			

Table 2.1. Beginning of Fisher's iris data; measurements are sepal length (X_1), sepal width (X_2), petal length (X_3), petal width (X_4) on speciments of each of three types of iris.

Probably the simplest way to handle this in SPlus is to input the data in the form of a 150×4 matrix `iris`, the first fifty rows corresponding to type A (iris setosa), the next fifty to type B (iris versicolour) and the remainder to type C (iris virginica). The commands

```
type<-factor(rep(LETTERS[1:3],c(50,50,50)))

iris.df<-data.frame(iris,type)
```

form a *data frame* called `iris.df` which consists of the four columns of `iris` together with a label `type` which is the letter A for the first 50 rows, B for the next 50, and C for the rest.

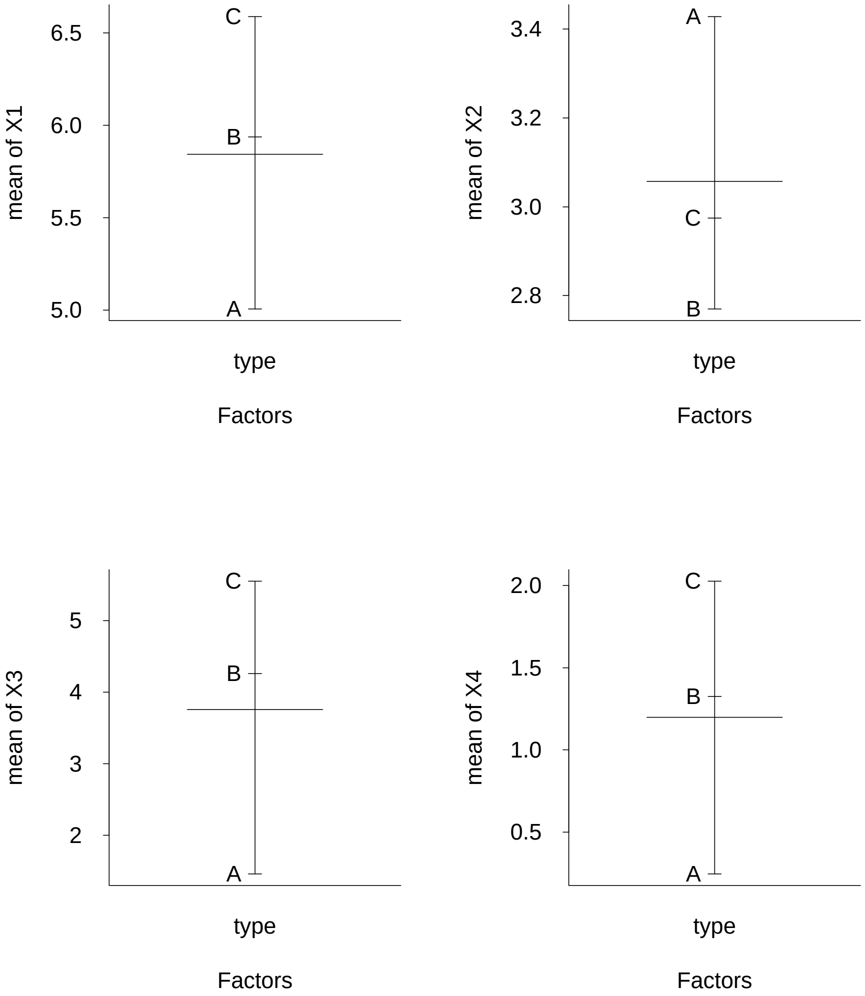

Fig. 2.1. Result of `plot.design` applied to iris data.

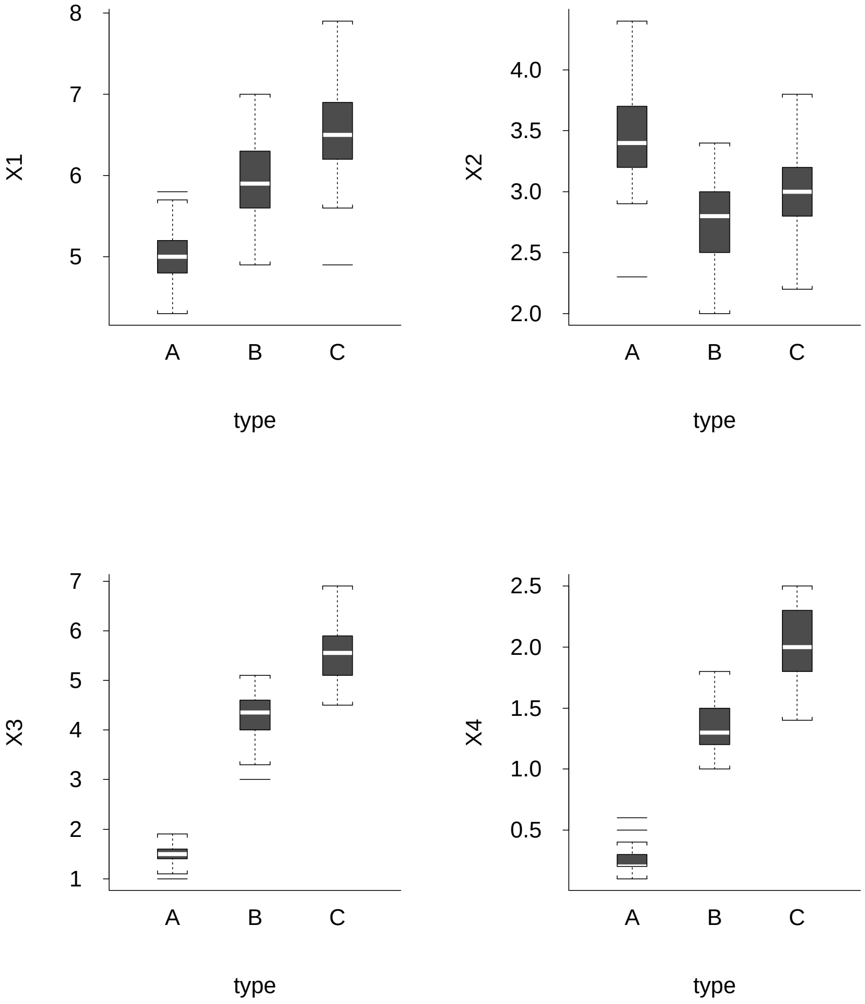

Fig. 2.2. Result of `plot.factor` applied to iris data.

The command

```
plot.design(iris.df)
```

produces a plot of the means for each of the four variables, labelled according to each of the three types. The command

```
plot.factor(iris.df)
```

creates a more informative boxplot for each variable, again classified by type. The results of these plots are shown in Figures 2.1 and 2.2. The plots make it obvious that there are large differences among the three types.

Given this information, it is hardly necessary to test whether the three types are significantly different, but to illustrate the use of the MANOVA tests, we shall do that. The Splus commands

```
m1<-manova(iris type,iris.df)
m2<-summary(m1)
for(i in c("p","w","h","r"))print(m2,test=i)
```

perform the MANOVA calculations and print out the results of the four tests, respectively Pillai, Wilks, Hotelling-Lawley and Roy.

```
              Df     Pillai Trace        approx. F  num df   den df   P-value
 type         2      1.1919              53.4665    8        290      0
 Residuals  147
              Df     Wilks Lambda        approx. F  num df   den df   P-value
 type         2      0.0234              199.1453   8        288      0
 Residuals  147
              Df     Hotelling-Lawley    approx. F  num df   den df   P-value
 type         2      32.4773             580.5321   8        286      0
 Residuals  147
              Df     Roy Largest         approx. F  num df   den df   P-value
 type         2      32.192              1166.957   4        145      0
 Residuals  147
```

As an illustration of how these calculation are performed, consider the Wilks test statistic $\Lambda = .0234$. In this case $h = 2$ and $r = 147$, so (2.17) shows that $F = (1 - \sqrt{.0234}) \cdot 144/(\sqrt{.0234} \cdot 4) = 199.3$ (199.1 after correction for rounding error) and the degrees of freedom of the F statistic are $2p = 8$ and $2(r - p + 1) = 288$. The result is of course massively significant, as are the other three test statistics.

3. PRINCIPAL COMPONENTS

3.1 Introduction

The method of principal components is a technique for extracting linear combinations from multivariate data which capture most of the variability in the data. It is used in numerous different ways. One interpretation is that it is largely a descriptive technique — given a large array of high-dimensional data which we do not know what to do with, a principal components analysis (PCA) will help us identify key components which can then be subjected to more detailed examination. Another application is to the formation of indices, e.g. given annual statistics of crimes committed in a number of different categories, how can we best combine the different numbers into an overall index of criminal behavior? A third interpretation is that PCA is a dimension reduction technique to be applied prior to some other form of analysis. For example, one way to reduce the dimensionality of a multiple regression problem is to perform an initial PCA on the regressor variables, followed by an ordinary multiple regression on some of the leading components. This is called *principal components regression.*

The usual procedure is as follows. Suppose X is a p-dimensional vector with covariance matrix Σ. The first principal component is a linear combination g^TX, for some vector g which satisfies $g^Tg = 1$, which is chosen to maximize the variance among all such linear combinations. In other words, we find g to maximize $g^T\Sigma g$ subject to $g^Tg = 1$. If the solution is $g = g_1$, then g_1^TX is called the *first principal component* of X.

If we want to go beyond the first PC, we then repeat the optimization in the space orthogonal to g_1: find g_2 to maximize $g_2^T\Sigma g_2$ such that $g_2^Tg_2 = 1$, $g_2^Tg_1 = 0$. Then g_2^TX is the *second principal component.*

The process continues iteratively: suppose $g_1, ..., g_{k-1}$ are given, for some $k \leq p$, then find g_k to maximize $g_k^T\Sigma g_k$ subject to $g_k^Tg_k = 1$, $g_k^Tg_j = 0$ for $j = 1, ..., k-1$. Then g_k^TX is the *k'th principal component.* In principle we could go on to find all p PCs, though in practice it is usual to stop after selecting enough PCs to capture most of the variability in the data — how many is one of the questions we shall be discussing (section 3.3).

Now we show how to calculate $g_1, g_2, ...$ Let G be an orthogonal matrix such that $G^T\Sigma G = D$, where D is a diagonal matrix with diagonal entries ordered so that $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$. We can always find such a representation, because Σ is symmetric and non-negative definite. Let g_k be the k'th column of G. So g_k is a norm-one eigenvector of Σ with eigenvalue λ_k.

Claim: These g_k's are the solution of the optimization problem described above. Moreover, the principal components $g_1^TX, g_2^TX, ..., g_p^TX$ are uncorrelated, and the sum of their variances is the sum of the variances of the individual components of X.

Proof of claim. The eigenvectors $\{g_k,\ 1 \le k \le p\}$ form a complete orthonormal basis in $\mathcal{R}^p$, so for any $g \in \mathcal{R}^p$, there exist constants $c_1, ..., c_p$ such that $g = \sum_k c_k g_k$. Then

$$\begin{aligned} g^T g &= \sum_{k=1}^{p} \sum_{\ell=1}^{p} c_k c_\ell g_k^T g_\ell = \sum_{k=1}^{p} c_k^2, \\ g^T \Sigma g &= \sum_{k=1}^{p} \sum_{\ell=1}^{p} c_k c_\ell g_k^T \Sigma g_\ell \\ &= \sum_{k=1}^{p} \sum_{\ell=1}^{p} c_k c_\ell g_k^T g_\ell \lambda_\ell = \sum_{k=1}^{p} \lambda_k c_k^2. \end{aligned}$$

Since the $\{\lambda_k\}$ are ordered, $g^T \Sigma g \le \lambda_1 \sum_k c_k^2 = \lambda_1$, with equality if $c_2 = ... = c_p = 0$. This proves that g_1 has the property of maximizing $g^T \Sigma g$ subject to $g^T g = 1$. It is the unique solution, up to changes of sign, if $\lambda_1 > \lambda_2$, but we have not excluded the possibility that $\lambda_1 = \lambda_2$ in which case the solution is not unique.

To get the second PC, we restrict attention to $g = \sum_k c_k g_k$ which are orthogonal to g_1, in other words, for which $c_1 = 0$. But then, an extension of the same reasoning shows that $g^T \Sigma g \le \lambda_2 \sum_k c_k^2 = \lambda_2$, with equality if $c_3 = ... = c_p = 0$. This proves that g_2 solves the equation for the second PC, and is unique if $\lambda_1 > \lambda_2 > \lambda_3$. We proceed in similar fashion to derive the third, fourth, and subsequent PCs.

The PCs are orthogonal, because for $k \ne \ell$, $g_k^T \Sigma g_\ell = \lambda_\ell g_k^T g_\ell = 0$ by orthogonality of the g_k's, and the k'th PC has variance $g_k^T \Sigma g_k = \lambda_k g_k^T g_k = \lambda_k$. Finally, the sum of the variances of the PCs is $\sum_k \lambda_k = \mathrm{tr}(D) = \mathrm{tr}(\Sigma)$, which is the sum of the variances of the individual components of X. With this, the proof of the claim is completed.

The above has been presented as if it all applied to the population covariance matrix X. In practice, we usually don't know Σ, and have to estimate it by the sample covariance matrix S based on a sample of n values of X. The preceding operations are then performed on S instead of Σ to produce the sample PCs. One side comment here is that for a continuous probability distribution with nondegenerate Σ, the sample covariance matrix S has distinct eigenvalues with probability 1, so the sample PCs are uniquely defined even though the population PCs may not be. In the case of multivariate normal data, there is a rich sampling theory of how well the sample g_k's and λ_k's approximate the corresponding population values (see, e.g., Mardia, Kent and Bibby (1979), Theorem 8.3.3, page 230), but it is more usual in practice to treat PCA as largely a data-descriptive technique without giving particular attention to sampling issues. In SPlus, the sample covariance matrix is the one created with divisor n rather than $n-1$ (in the notation of chapter 2, S_0 rather than S).

3.2. PCA based on the correlation matrix.

The description in section 3.1 assumes we are performing PCA on the population covariance matrix Σ or its sample counterpart S. One difficulty associated with this is that the problem is not scale invariant — if the original data were lengths measured in inches and weights measured in pounds, and if we then changed the scales of measurement to centimeters and kilograms, the PCs and the corresponding λ_k's would change. A way to avoid this difficulty is to rescale the problem prior to computing the PCs, so that each component of X had either population variance or sample variance equal to 1. This is, of course, equivalent to replacing Σ or S by the corresponding correlation matrix. Thus, an alternative way of proceeding is via the correlation matrix instead of the covariance matrix.

There are arguments both for and against doing this. The problem just mentioned — lack of invariance to changes in the units of measurement — is one argument very commonly made in favor of using the correlation matrix. On the other hand, the PCs may be more meaningful if expressed on the scale of the original data. If the variables X consisted, for example, of turnover in dollars by the different sections of a large company, the effect of using the correlation matrix would be to give sections with a small turnover as much weight as those with very large turnover. Given that all the variables are here measured on the same scale, there is no need to standardize prior to performing the PCA. If the company were an international company with turnover measured in local currencies, it would make sense to standardize all the amounts by converting them to a common currency, but this is easily done, and a quite different operation from standardizing all the turnovers to have variance 1. Which form of PCA we should use — covariance-based or correlation-based — depends very much on the individual application.

3.3. Deciding how many PCs to include.

To reduce the dimensionality of the problem, we would like to restrict attention to the first k PCs, where k is much less than p, but to avoid losing too much of the variability in the original data, we would also like to choose this so that the proportion of variance explained by the first k PCs, which may be expressed as

$$\psi_k = \frac{\lambda_1 + ... + \lambda_k}{\lambda_1 + ... + \lambda_p}, \tag{3.1}$$

is close to 1. The question is, how should we choose k to balance these two criteria?

Three methods are widely proposed.

1. The "screeplot": plot the ordered λ_k against k and decide visually when the plot has flattened out. The name comes from an analogy with rocks on a mountain — the initial part of the plot, in which λ_k is decreasing rapidly with k, is like the side of a mountain, while the flat portion, in which each λ_k is only slightly smaller than its predecessor λ_{k-1}, is like the rough scree at the bottom. The task of the data analyst is to decide when the "scree" begins.

2. Choose k so that $\psi_k \geq c$, for some arbitrary cutoff c. For some reason, everybody uses $c = 0.9$ when applying this rule. Presumably this is no less arbitrary than the convention that all tests of signficance should be based on $\alpha = .05$.

3. Kaiser's rule: exclude all PCs with eigenvalues less than the overall average of the eigenvalues (which, in the case of a correlation-based PCA, is always 1). This rule also seems to be arbitrary, e.g. we could with no less logic set the cutoff at twice, or half, or 0.9354 times, the mean eigenvalue. (In fact it seems to be widely believed that Kaiser's rule leads to the inclusion of too few PCs, whereas the screeplot often tempts one to include too many. This would seem to be an argument in favor of using a smaller multiplying factor than 1 in Kaiser's rule.)

4. Formal tests of significance (discussed in more detail by Mardia, Kent and Bibby (1979)). If the population eigenvalues are $\lambda_1, ..., \lambda_p$ and the corresponding sample eigenvalues are $\hat{\lambda}_1, ..., \hat{\lambda}_p$, one's first thought is to test whether $\lambda_{k+1} = ... = \lambda_p = 0$, but this does not actually make sense, because if the null hypothesis were satisfied, then the population distribution is contained entirely within a k-dimensional subspace and therefore the same would be true of any sample (in other words, under the null hypothesis, we would have $\hat{\lambda}_{k+1} = ... = \hat{\lambda}_p = 0$ with probability 1). So instead, one possibility is to test $H_0 : \lambda_{k+1} = ... = \lambda_p$ (without requiring that the common value being 0), presumably on the grounds that if the sample eigenvalues are indistinguishable from some common number there is no significance in the individual values. A test for this hypothesis is to form the algebraic and geometric means

$$a_0 = \frac{\hat{\lambda}_{k+1} + ... + \hat{\lambda}_p}{p-k} \qquad g_0 = (\hat{\lambda}_{k+1} ... \hat{\lambda}_p)^{1/(p-k)},$$

and then construct

$$-2 \log \lambda = n(p-k) \log \frac{a_0}{g_0}.$$

Approximate distributional results under H_0:

(a) The usual χ^2 approximation to a likelihood ratio test,

$$n(p-k) \log \frac{a_0}{g_0} \sim \chi^2_\nu, \quad \nu = \frac{(p-k+2)(p-k-1)}{2}, \tag{3.2}$$

(b) The Bartlett correction: replace n in (3.2) with

$$n' = n - \frac{2p+11}{6}.$$

These results assume multivariate normality, and they are only valid as stated for the covariance-based version of PCA, not the correlation-based version. In practice, many data analysts who view PCA as primarily a descriptive technique do not want to make assumptions of multivariate normality, and the asymptotic nature of the above results is cited as further reason for distrusting them. Therefore, in practice, the simple data-based methods for choosing k are used much more than the formal test just described.

3.4. Principal components in regression

Consider a regression model of the form

$$y_i = \sum_{j=1}^{p} \beta_j x_i^{(j)} + \epsilon_i, \tag{3.3}$$

in which $\{\epsilon_i\}$ satisfy the usual assumptions (for example, uncorrelated, mean 0, common variance) but there are a large number p of possible regressors $\{x_i^{(j)}\}$. The idea of PC regression is to use PCA to reduce the number of regressors prior to fitting a model of the form (3.3).

A further reason for doing this is that since the PCs are orthogonal, the X matrix in the transformed regression problem will be orthogonal, thus avoiding all the problems which often arise in regression analysis due to multicollinearity. Indeed, PC regression is sometimes cited as an alternative to ridge regression, which has also been proposed as a way of dealing with multicollinearity in high-dimensional regression analysis, but which goes about the problem in a quite different way.

The main problem posed by this approach is, once again, the selection of which PCs to include. The methods proposed in section 3.3 can of course be applied, but there are additional possibilities based on the correlation between the PCs and the dependent variable y_i. We can, for example,

(a) Order the PCs according to their sample variances, choosing some k such that we ignore all PCs after the k'th — this is the procedure of section 3.3,

(b) Order the PCs according to their correlations with y, again choosing a cutoff k,

(c) A compromise between (a) and (b), in which we order $\lambda_1 \geq ... \geq \lambda_p$ as usual and then test in reverse order for the significance of $g_p^T X^{(j)}$, $g_{p-1}^T X^{(j)}, ...$, stopping a soon as one is significant (Jolliffe's rule).

(d) Another strategy entirely is to use the y_i's in defining the components, for example, defining $t_i^{(1)} = \sum_j x_i^{(j)} c_j$ with weights $c_j^{(1)}$ such that $\sum {c_j^{(1)}}^2 = 1$, to maximize the sample correlation between $\{y_i\}$ and $\{t_i^{(1)}\}$, then choosing $t_i^{(2)} = \sum_j x_i^{(2)} c_j$ with $\sum {c_j^{(2)}}^2 = 1$ to maximize the correlation with $\{y_i\}$ among all linear combinations orthogonal to $\{t_i^{(1)}\}$, and so on, followed by ordinary least squares regression of y_i on $t_i^{(1)}, t_i^{(2)}, ...$ This is, however, really a quite different method, known as *partial least squares regression*, but also one which has been studied widely in recent years.

3.5 Implementation in SPlus

SPlus provides a good implementation of PCA. As an illustration, we shall discuss it for a table of data taken from Mardia *et al.* (1979), partly reproduced in Table 3.1. The table represents scores of 88 students on five Mathematics examinations, arranged in rough order or merit. The first four and the last four students' scores are shown.

Vectors	Mechanics	Algebra	Analysis	Statistics
77	82	67	67	81
63	78	80	70	81
75	73	71	66	81
55	72	63	70	68
⋮	⋮	⋮	⋮	⋮
5	30	44	36	18
12	30	32	35	21
5	26	15	20	20
0	40	21	9	14

Table 3.1. Scores of students on five Mathematics examinations.

A sample SPlus program is as follows:

```
exams<-matrix(scan(file='exams.dat'),byrow=T,ncol=5)
nr<-length(exams[,1])
dimnames(exams)<-list(1:nr,c("Vectors","Mechanics","Algebra",
"Analysis","Statistics"))
exams.prc<-princomp(exams,cor=F)
print(summary(exams.prc,loadings=T))
screeplot(exams.prc)
screeplot(exams.prc,style='lines')
plot(loadings(exams.prc))
exams.eigen<-exams.prc$sdev^2
print(sqrt(mean(exams.eigen)))
pr<-predict(exams.prc)
plot(1:88,pr[,1],xlab='Student',ylab='Component 1')
plot(1:88,pr[,2],xlab='Student',ylab='Component 2')
plot(1:88,pr[,3],xlab='Student',ylab='Component 3')
plot(1:88,pr[,4],xlab='Student',ylab='Component 4')
exam2<-matrix(scan(file='exam2.dat'),byrow=T,ncol=5)
print(predict(exams.prc,newdata=exam2))
biplot(exams.prc)
```

Line 1 inputs the data, and lines 2–4 define the variable names. (*Note*: In reality lines 3 and 4 are typed as one line, but are spread over two here to fit on the page.) Line 5 is then the main command to set up the PCA and store the results in `exams.prc`. The option `cor=F` (the default) means that the PCA is based on the covariance matrix; `cor=T` would choose the correlation matrix. Line 6 prints a summary which at minimum shows the successive values of the components standard deviations (i.e. $\sqrt{\lambda_k}$, the proportion of variance accounted for by the k'th component), and the cumulative proportions (the ψ_k values of (3.1)). The option `loadings=T` also prints the factor loadings, i.e. the components of the g_k vectors which therefore show what each of the five original variables contributed to each of the PCs. Lines 7 and 8 demonstrate a screeplot, both in its default form which represents the screeplot as a bar diagram, and with `style='lines'` which represents it as an ordinary scatterplot joined by lines. Line 9 forms a bar diagram plot of each of the factor loadings. Lines 10 and 11 compute and print the mean eigenvalue, used in Kaiser's criterion. Line 12 forms a set of "prediction scores", in other words, computing the value of each PC for each student in the sample. Lines 13–16 plot the first four prediction scores. Lines 17–18 show how to extend the analysis to include new subjects: with an additional data set of six students in the file `exam2.dat`, it shows the scores for the new students, computed using the PCs from the original data. Finally, line 19 illustrates another graphical device, the *biplot.*

Figures 3.1–3.4 show the screeplots, the component loadings plots, the plots of prediction scores and the biplots.

The screeplots show that the variances are small after the first two PCs, and the $\psi_k \geq .9$ criterion is almost satisfied for $k = 2$. The standard deviations of the five components are 26.06, 14.14, 10.13, 9.15, 5.64, with a root mean square value of 14.81, so Kaiser's criterion would tell us to ignore everything after the first PC, which does not seem the right interpretation in this instance. From Fig. 3.2, we can see that the first PC is close to an average of all five examinations, and the corresponding prediction plot (top graph in Fig. 3.3) reflects that the order of the students is virtually unchanged by using the first PC as a summary score. The second component shows negative weightings in vectors and mechanics, positive weightings in analysis and statistics, with a small positive weighting in algebra. This therefore seems to reflect a contrast between skills in applied mathematics and the more analytical skills needed for analysis and statistics. The loadings for components 3–5 do not have any such simple interpretation and may simply reflect the noise left over in the series after subtracitng the first two components. This is another reason for not considering PCs beyond the first two.

As an example of the interpretation of the second PC, the second plot of Fig. 3.3 shows that students numbered 23 and 28 scored particularly highly in component 2, reflecting that they performed very well on the analysis and statistics papers. In trying to rank the students overall, some would argue that very good performance on one part of the syllabus was worthy of more credit than a flat overall performance. Of course, a similar interpretation would apply to students who had outlying low scores in PC2, in this case, that they were very strong in applied mathematics.

The *biplot* (Fig. 3.4) is another useful device for visualizing the interaction between the first two PCs, obtained by plotting the first two PCs for each student as a scatterplot.

In this case the scatterplot does not show any particular structure, but it could be useful, for instance, for detecting a nonlinear relationship. The (linear) correlation between the first two components must be 0, this being part of the construction of PCs. Also shown on the biplot are the loadings of the first two PCs for each of the five original variables, which serves to reinforce the message given by Fig. 3.2, concerning the relationship among the five variables.

Sometimes one does not have the full data matrix available but it already summarized in the form of a covariance matrix. SPlus also provides for the computation of a PCA in this case. Suppose the sample covariance matrix has been computed and stored as the file `examcov.dat`. The following commands will compute the PCA, and print out the component standard deviations and loadings:

```
examcov<-matrix(scan(file='examcov.dat'),byrow=T,ncol=5)
cov.obj<-list(cov=examcov,center=c(0,0,0,0,0))
exams.prc<-princomp(covlist=cov.obj,cor=F)
print(summary(exams.prc,loadings=T))
```

Year	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959	1960	1961	1962	1963
Homicide	529	455	555	456	487	448	477	491	453	434	492	459	504	510
Wounding	5258	5619	5980	6187	6586	7076	8433	9774	10945	12707	14391	16197	16430	18655
Homosexual offences	4416	4876	5443	5680	6357	6644	6196	6327	5471	5732	5240	5605	4866	5435
Heterosexual offences	8178	9223	9026	10107	9279	9953	10505	11900	11823	13864	14304	14376	14788	14722
Breaking and entering	92839	95946	97941	88607	75888	74907	85768	105042	131132	133962	151378	164806	192302	219318
Robbery	1021	800	1002	980	812	823	965	1194	1692	1900	2014	2349	2517	2483
Larceny	301078	355407	341512	308578	285199	295035	323561	360985	409388	445888	489258	531430	588566	635627
Fraud	25333	27216	27051	27763	26267	22966	23029	26235	29415	34061	36049	39651	44138	45923
Receiving stolen goods	7586	9716	9188	7786	6468	7016	7215	8619	10002	10254	11696	13777	15783	17777
Injury to property	4518	4993	5003	5309	5251	2184	2559	2965	3607	4083	4802	5606	6256	6935
Forgery	3790	3378	4173	4649	4903	4086	4040	4689	5376	5598	6590	6924	7816	8634
Blackmail	118	74	120	108	104	92	119	121	164	160	241	205	250	257
Assault	20844	19963	19056	17772	17379	17329	16677	17539	17344	18047	18801	18525	16449	15918
Malicious damage	9477	10359	9108	9278	9176	9460	10997	12817	14289	14118	15866	16399	16852	17003
Revenue laws	24616	21122	23339	19919	20585	19197	19064	19432	24543	26853	31266	29922	34915	40434
Alcohol laws	49007	55229	55635	55688	57011	57118	63289	71014	69864	69751	74336	81753	89709	89149
Indecent exposure	2786	2739	2598	2639	2587	2607	2311	2310	2371	2544	2719	2820	2614	2777
Motor theft	3126	4595	4145	4551	4343	4836	5932	7148	9772	11211	12519	13050	14141	22896

Table 3.2. Recorded offences in Great Britain from 1950–1963.

As a second example of PCA, consider the data in Table 3.2, taken from Chatfield and Collins (1980). This consists of numbers recorded for 18 different types of crime over 14 years in Britain. The interest here may be in such questions as how to develop an overall index of criminal behaviour, though Chatfield and Collins cast doubt on whether

PCA is really useful for answering this kind of question in the present context. The wide disparity in numbers of events in different crime categories more or less forces us to use the correlation-based form of PCA, though in itself this is questionable because it exaggerates the importance of some crimes with comparatively small numbers.

The SPlus commands for this data set are similar to those for the examination scores data set. Having created a 14×18 data matrix `crime`, the PCA was performed with the command

```
crime.prc<-princomp(crime,cor=T)
```

The only real difference from the sequence of commands used for the examinations data set is the one used to compute the loadings plot (Fig. 3.6), which has to be more customized because of the greater number of variables. The command

```
plot(loadings(crime.prc),variables=1:4,nbars=18)
```

creates a plot for all 18 variables for the first four PCs. The four plots resulting, corresponding to Figs. 3.1–3.4 for the examinations data, are shown in Figs. 3.5–3.8 for the crimes data.

In this case the standard deviations of the leading PCs are 3.59, 1.65, 0.98, 0.83, 0.57,... Since this PC is based on a correlation matrix, Kaiser's criterion leads automatically to 1 as the cutoff value, which would mean keeping only the first two PCs. The $\psi_k > .9$ criterion is satisfied for $k = 3$, while visual inspection of the screeplot in Fig. 3.5 might lead one to adopt an even more conservative criterion, based on k about 6 or 7. The meaning of the individual PCs is hard to interpret. Component 1 seems to be close to an average of all the variables, but with some variables (notably, assualt and homosexual offences) receiving negative weights, which presumably means that the trend in these crimes was in the opposite direction to the others over the years in question. Component 2 seems also to reflect the contrast between homosexual offences and some of the others; component 3 is dominated by the homicide variable. It is hard to see whether these have any real meaning beyond implying that the overall trend be broken down into two components as suggested by the first two plots in Fig. 3.7. Finally, Fig. 3.8 suggests that as time has evolved, the pattern of crime has traced an interesting nonlinear curve through the space defined by the first two PCs, but again, what this means is far from clear.

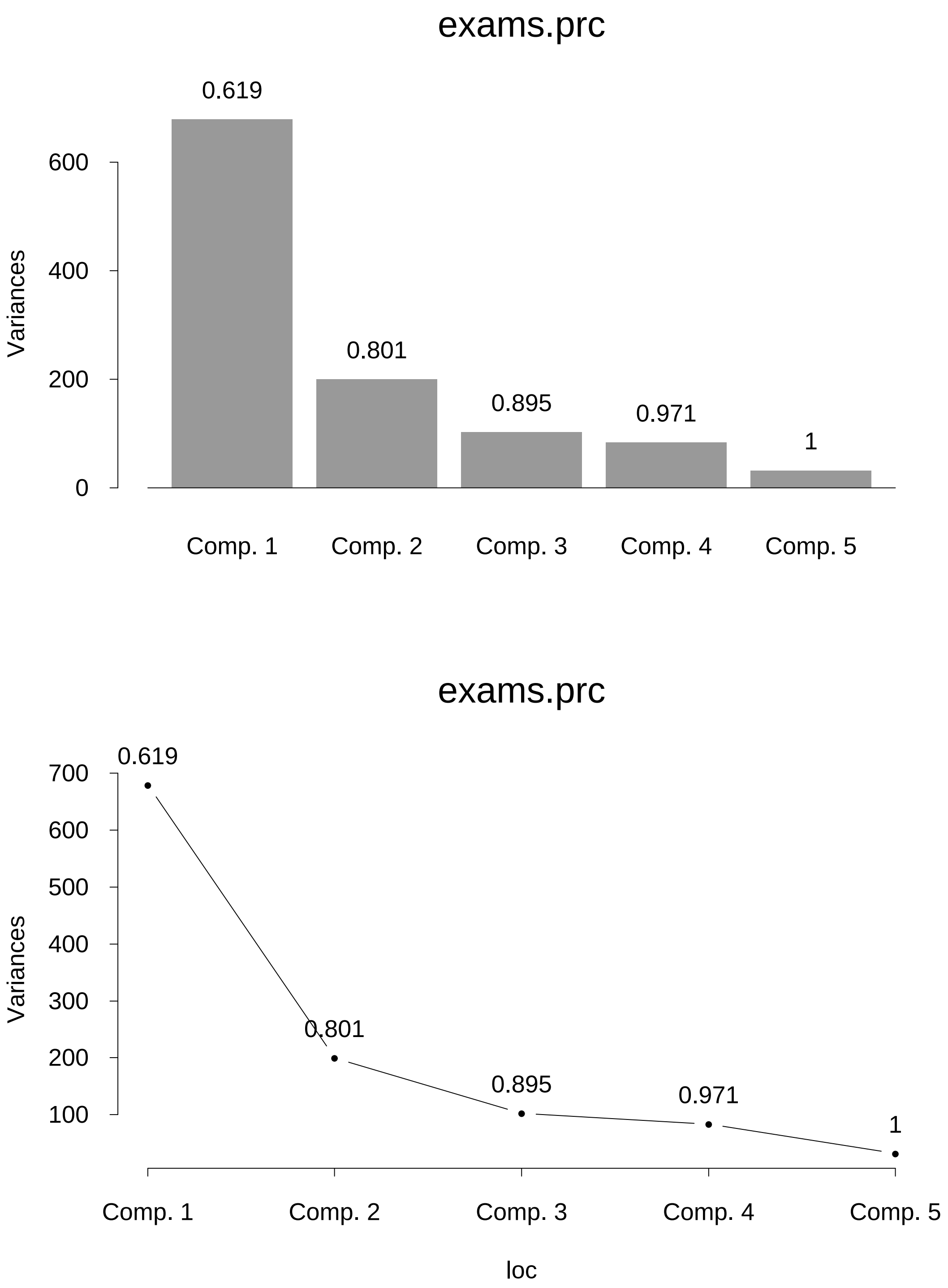

Fig. 3.1. Screeplots for the Exams data.

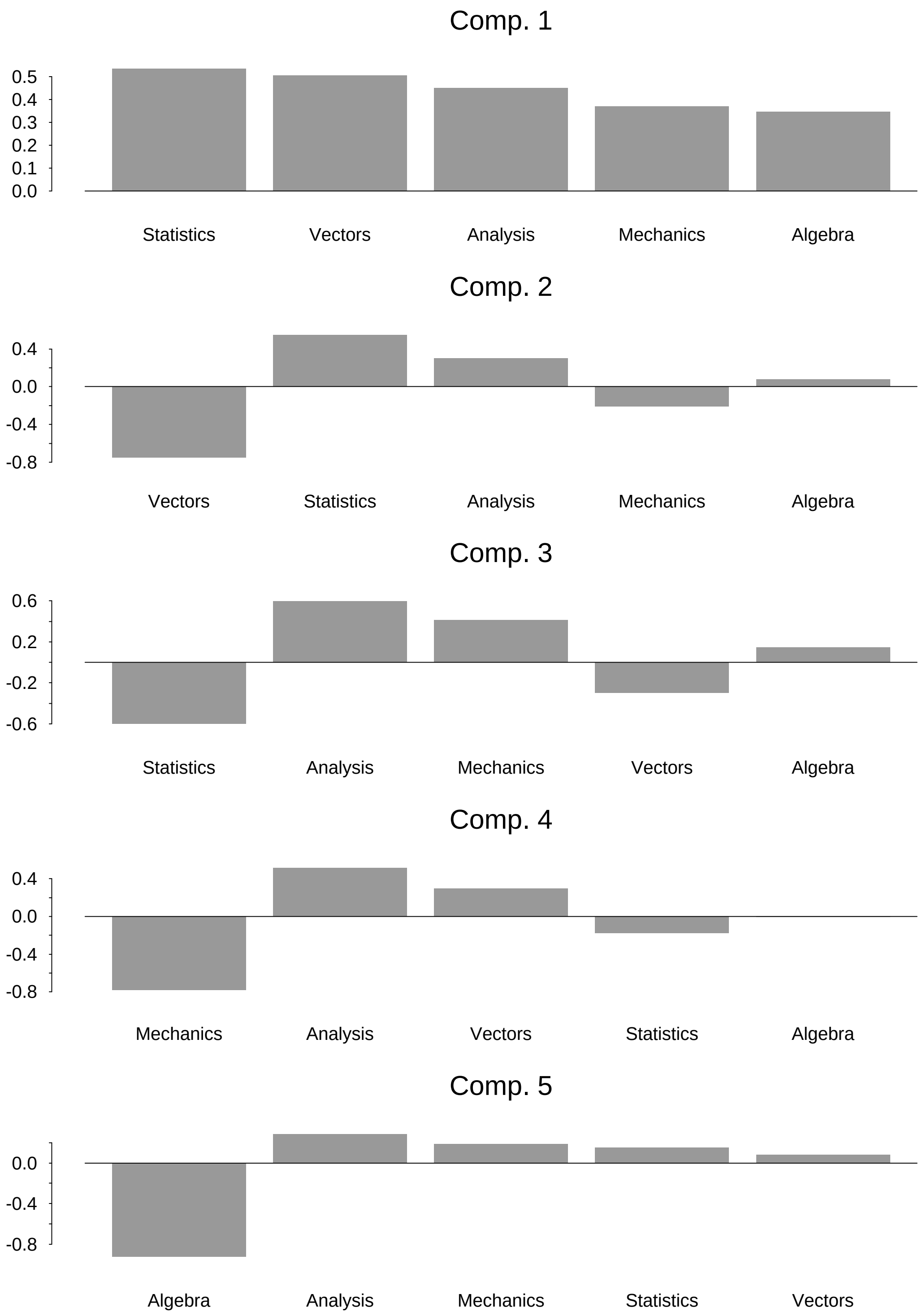

Fig. 3.2. PC loadings plot for the Exams data.

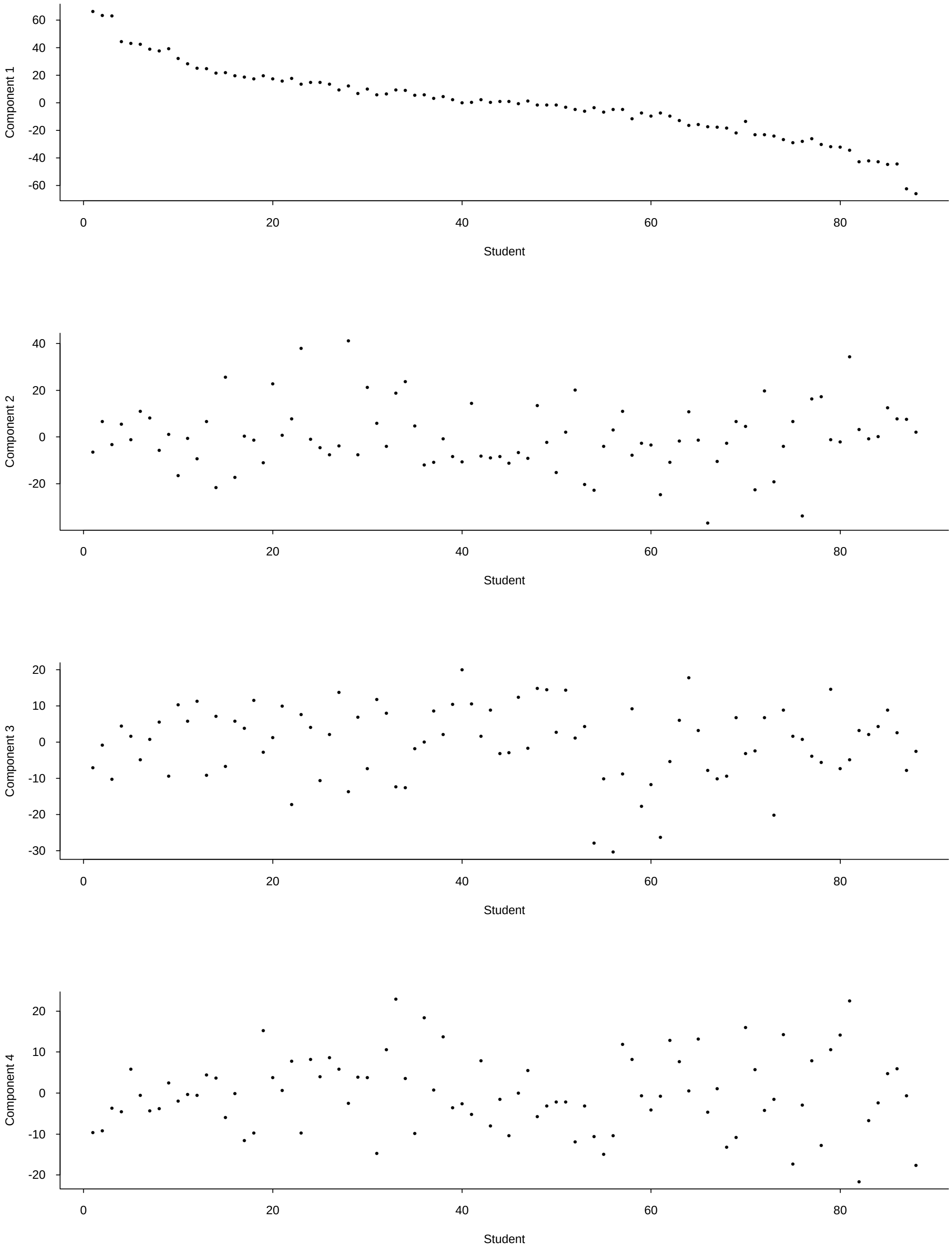

Fig. 3.3. Prediction plot for the Exams data.

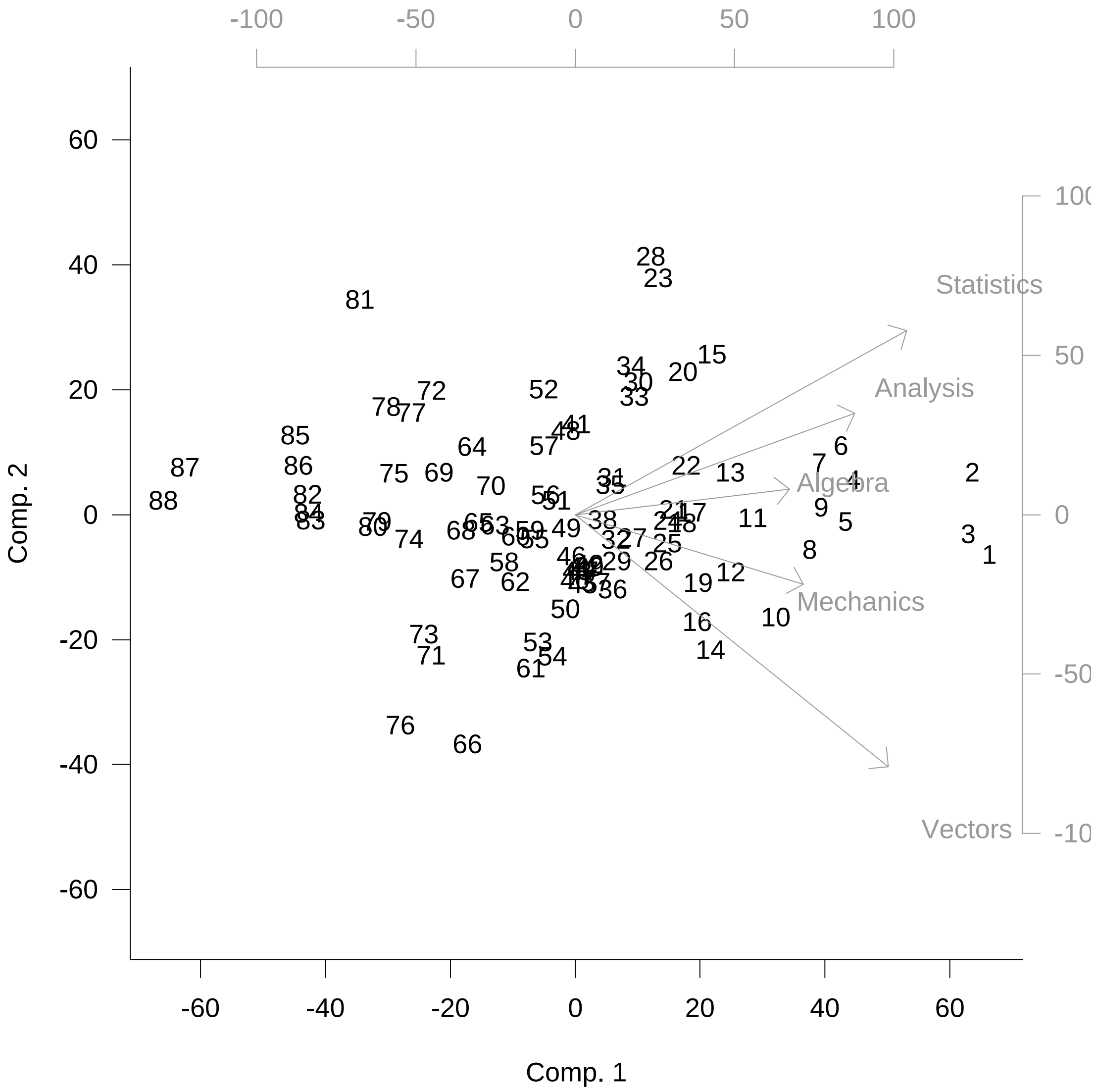

Fig. 3.4. Biplot for the Exams data.

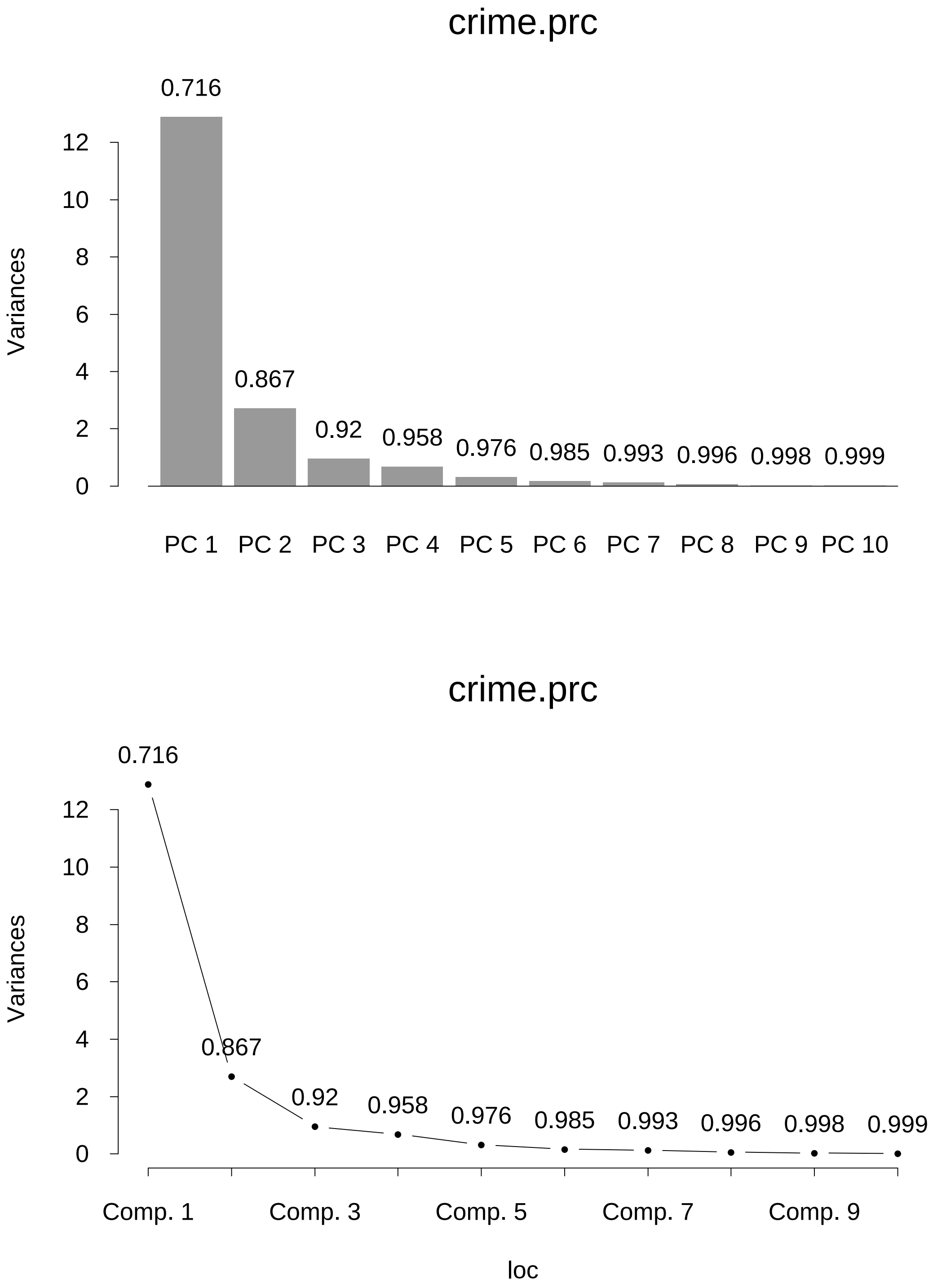

Fig. 3.5. Screeplots for the Crimes data.

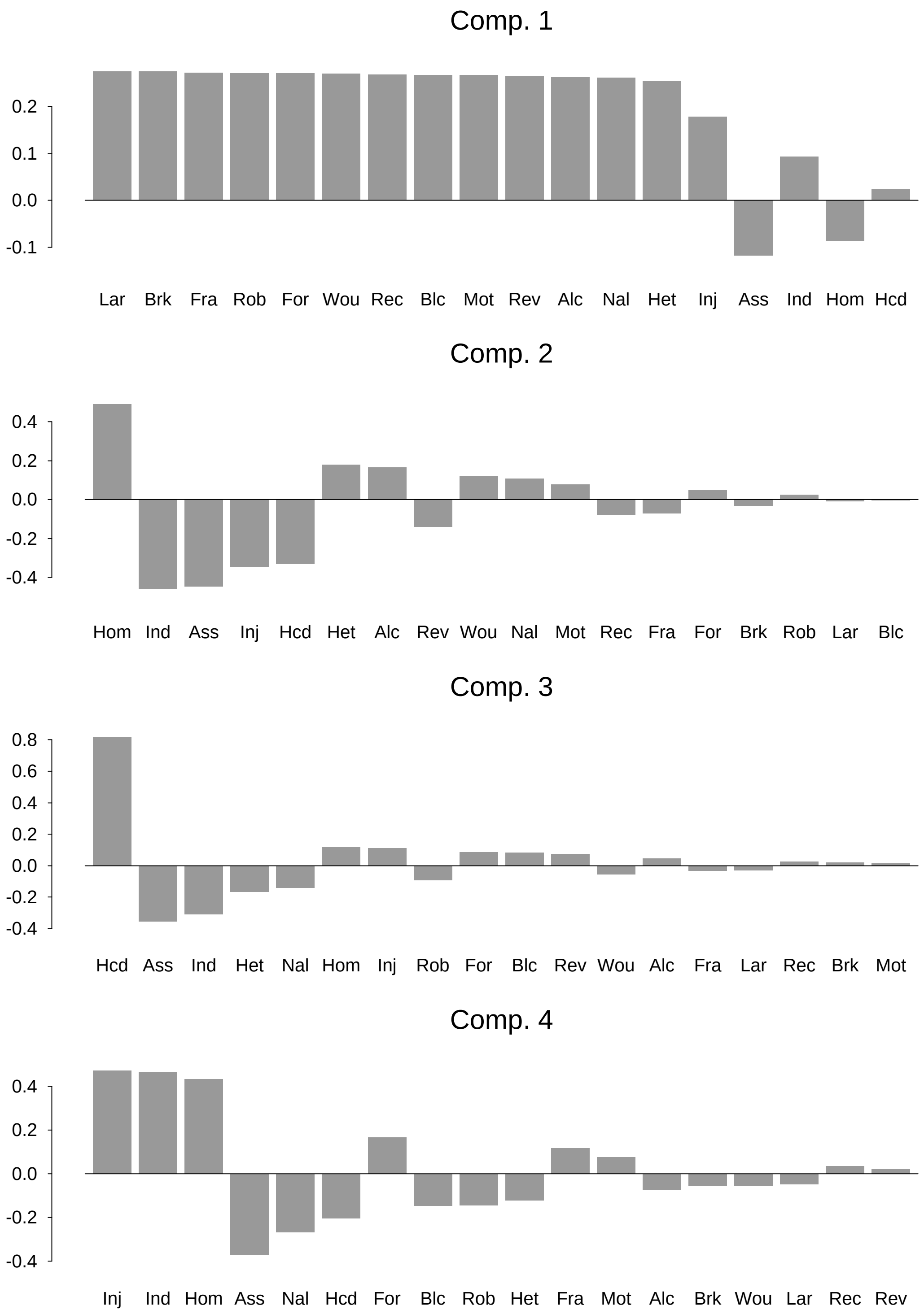

Fig. 3.6. PC loadings plot for the Crimes data.

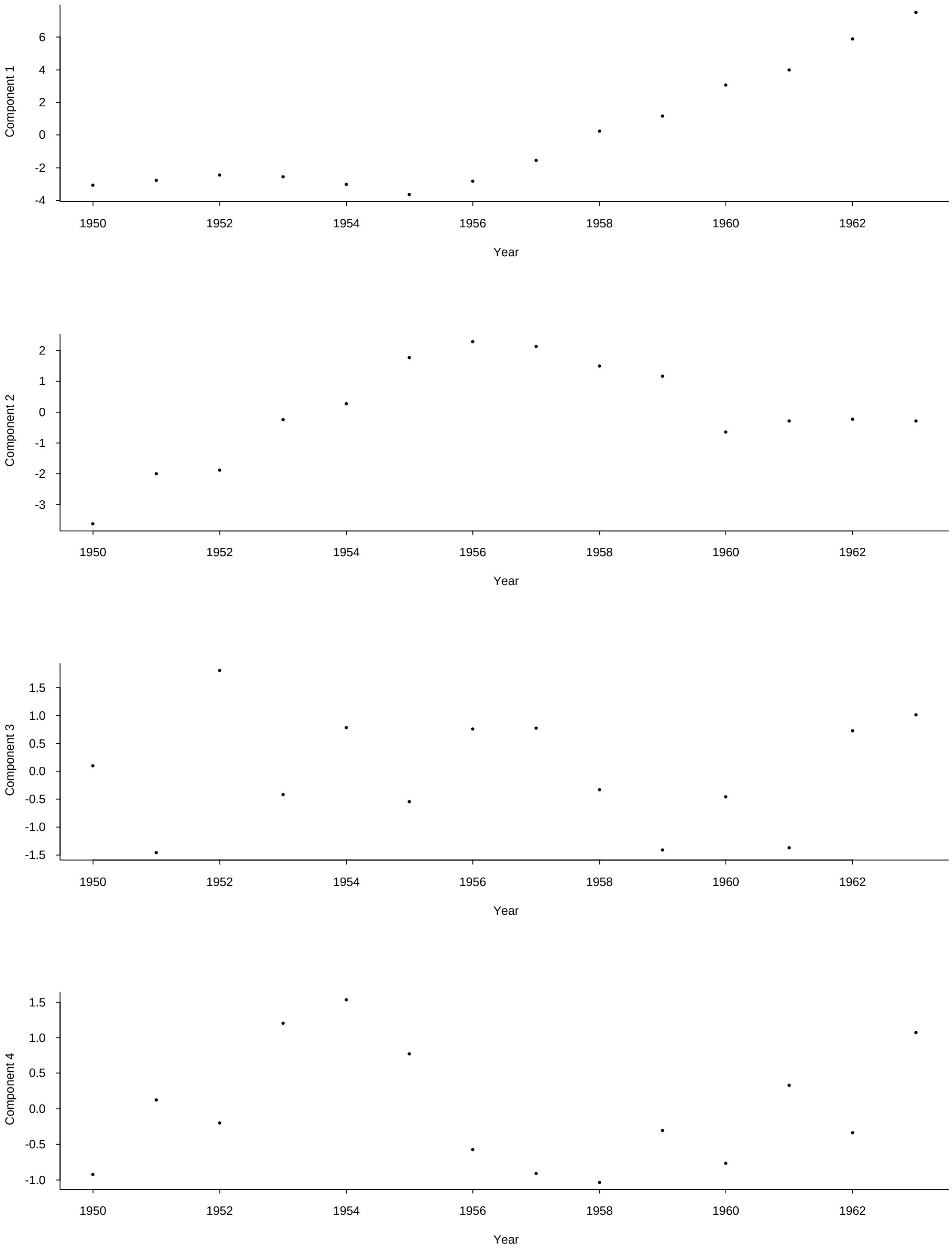

Fig. 3.7. Prediction plot for the Crimes data.

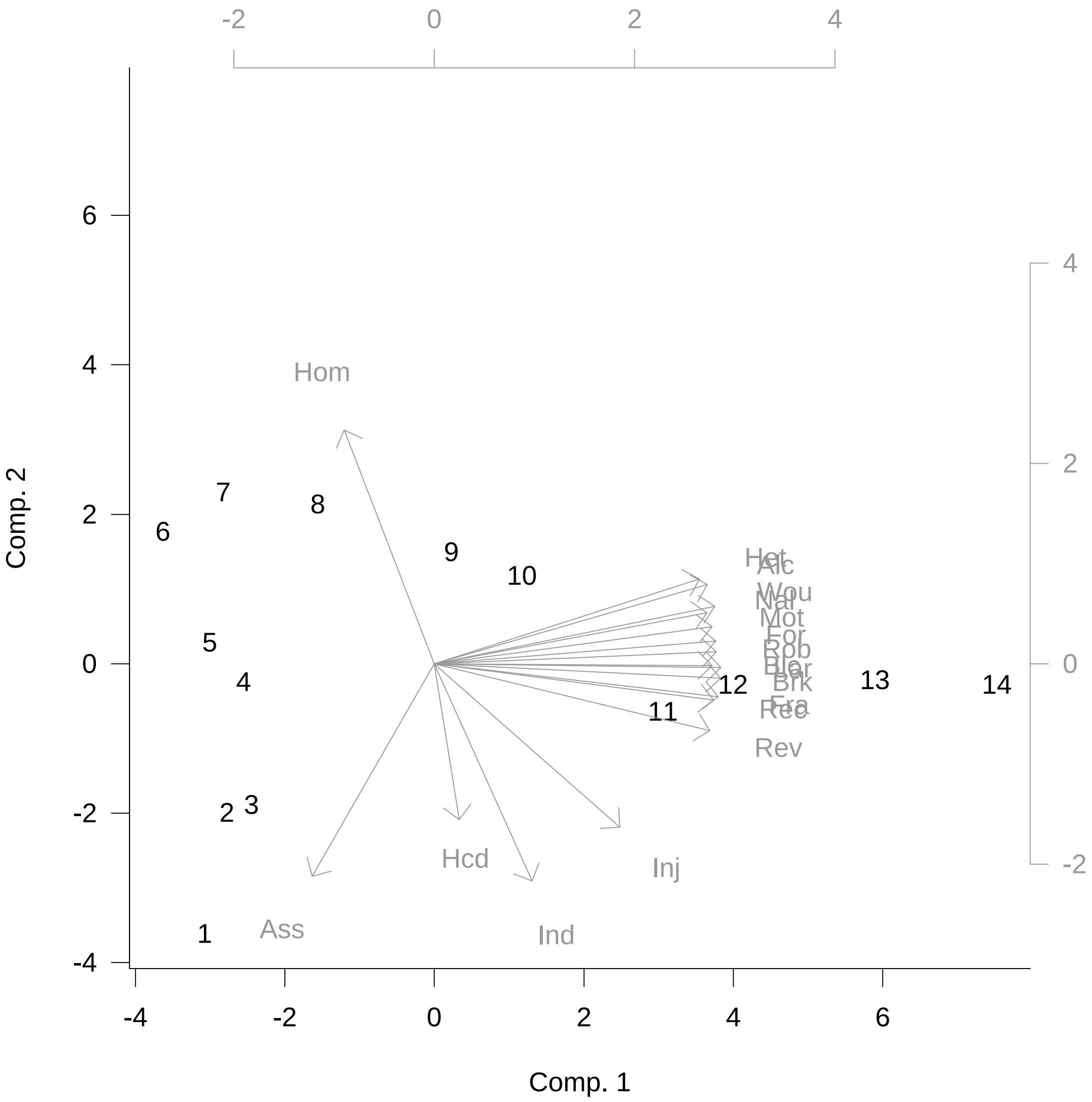

Fig. 3.8. Biplot for the Crimes data.

4. FACTOR ANALYSIS

4.1 Introduction

Factor analysis is an alternative to principal components analysis with which it is often confused, though the underlying principles behind the two methods are completely different. Like PCA, the objective is to "explain" the data in terms of a relatively small number of components, now known as *factors*. Unlike PCA, which is basically a model-free descriptive technique, factor analysis is model-based and its credibility depends a lot both on the model itself, and the extent to which a reasonable interpretation can be placed on the factors which are identified. The subject has a somewhat checkered history, having originally been developed by social scientists and in particular by psychologists interested in the measurement of intelligence, and this fact together with the tendency to apply the model in all circumstances regardless of its appropriateness gave the whole subject something of a tarnished image among statisticians (see Chatfield and Collins (1980) for considerable expansion on this theme). However, the trend in statistics research generally over the last twenty years has been much more towards the development of models containing latent variables, together with modern methods of model fitting and model checking based on Bayesian statistics and Monte Carlo methods. The use of state-space models based on the Kalman filter in time series analysis is one example of this trend, and indeed the basic model for factor analysis ((4.1) or (4.2) below) is reminiscent of the "measurement" equation in Kalman filtering. In this chapter, largely following Mardia, Kent and Bibby (1979), we shall adopt a traditional approach to the subject — in particular, all the inferential theory is based on the multivariate normal distribution — but it is worth keeping in mind that there are many possibilities for generalizing the models using modern statistical methods.

In a simple one-factor model, letting x_i denote the i'th component of a p-dimensional vector of dependent observations on an individual, we write

$$x_i = \lambda_i f + u_i, \qquad 1 \le i \le p. \tag{4.1}$$

Here f (the factor) is a univariate random variable defined for each individual, $\lambda_1, ..., \lambda_p$ are constant weights known as *factor loadings*, and $u_1, ..., u_p$ are independent random variables which could be interpreted as measurement errors.

The generalization of (4.1) to k factors, including also a non-zero mean, is

$$x_i = \sum_{j=1}^{k} \lambda_{ij} f_j + u_i + \mu_i, \qquad 1 \le i \le p, \tag{4.2}$$

in which $f_1, ..., f_k$ are k factors which may, without loss of generality, be assumed to be uncorrelated random variables of mean 0 and variance 1, $u_1, ..., u_p$ are uncorrelated random errors of mean 0 and variances $\psi_1, ..., \psi_p$ (also, the u's are uncorrelated with the f's), and $\mu_1, ..., \mu_p$ are arbitrary constants.

We also write (4.2) in vector notation as

$$X = \Lambda f + u + \mu, \tag{4.3}$$

in which f $(k \times 1)$, u $(p \times 1)$ and μ $(p \times 1)$ are vectors and Λ is a $p \times k$ matrix of constants known as factor loadings. In (4.3), we assume f and u both have mean 0, and covariance matrices I_k (the $k \times k$ identity matrix) and Ψ respectively, and $\mathrm{Cov}(f, u)$=0. Here Ψ is a diagonal matrix with diagonal entries $\psi_1, ..., \psi_p$.

As a consequence of (4.3), we have

$$\mathrm{Cov}(X) = \Sigma = \Lambda\Lambda^T + \Psi, \tag{4.4}$$

so another way of stating the problem is to say that we are going to investigate when a covariance matrix may represented in the form (4.4). If we further assume that the data are multivariate normal, such models can be estimated and tested within the framework of likelihood theory, as we shall see.

Another way of interpreting (4.2) is to say that

$$\begin{aligned} \mathrm{Var}\{x_i - u_i\} &= \mathrm{Var}\{\sum_j \lambda_j f_j\} = \sum_j \lambda_{ij}^2 \\ &= h_i^2 \text{ say} \end{aligned} \tag{4.5}$$

so that the portion of the variance of x_i which is “explained” by the common factors is a constant h_i^2 known as a *communality.*

4.2 Uniqueness and invariance properties

One of the problems we noted with PCA was that the analysis was not invariant to scale changes. In FA, this is not a problem: the model is invariant to scale changes, modulo obvious corresponding scale changes to the factor loadings and the variances of the random errors.

To see this, suppose X has covariance matrix (4.4) and $Y = CX$ where C is a diagonal matrix with diagonal entries $c_1, ..., c_p$. Then

$$\mathrm{Cov}\{Y\} = C\Sigma C = (C\Lambda)(C\Lambda)^T + C\Psi C$$

which is also of the form (4.4) with Λ replaced by $C\Lambda$ and Ψ by the new diagonal matrix $C^2\Psi$. In other words, the factor loading λ_{ij} is replaced by $c_i\lambda_{ij}$ and the i'th component error variance ψ_i by $c^2\psi_i$. Neither the number (or interpretation) of the factors nor the fit (or lack of fit) of the model is changed by these rescalings, so for all practical purposes, the scale of the data does not matter. This is quite different from PCA.

However, the invariance of the model to certain transformations also has a negative feature, in that it means the model is not, as we have specified it so far, uniquely defined. Suppose (4.4) holds, and let G be any orthogonal matrix. Then

$$\Sigma = \Lambda G G^T \Lambda^T + \Psi = (\Lambda G)(\Lambda G)^T + \Psi,$$

in other words, we may replace Λ by ΛG without changing the model. Another way of stating this is that the model is unchanged by *rotating* the factors in k-dimensional space.

In view of this, we need to specify some additional constraints to make the model uniquely defined. Three possibilities are

(i) Fix $\Lambda^T \Psi^{-1} \Lambda$ to be a diagonal matrix,

(ii) Fix $\Lambda^T D^{-1} \Lambda$ to be a diagonal matrix, where D is the diagonal matrix with diagonal entries $\sigma_{11}, ..., \sigma_{pp}$, the variances of the original random variables,

(iii) The *varimax* criterion, of which more later.

(i) and (ii) are largely mathematical constraints whose sole purpose is to make the problem uniquely defined — it can be shown that either such condition is achievable and that it does serve to define Λ uniquely — while (iii) is a condition more commonly used in practice. In particular, SPlus performs FA based on the varimax criterion.

One consequence of (i) or (ii) is that since a $k \times k$ matrix has $k(k-1)/2$ independent off-diagonal entries, either of these constraints has the effect of fixing that number of parameters. Thus, the number of free parameters in Λ is $pk - k(k-1)/2$, and the degrees of freedom restricted by the model are

$$s = \frac{p(p+1)}{2} - pk + \frac{k(k-1)}{2} = \frac{(p-k)^2 - (p+k)}{2}. \tag{4.6}$$

We usually assume $s > 0$; otherwise, the model is underdetermined, and in any case, there is not much point in employing a dimension reduction technique if the net effect is to increase the number of parameters which need to be estimated. The practical effect of $s > 0$ is that it restricts the value of k; for example, in $p = 5$ dimensions, we need $k \leq 2$; if $p = 10$, then $k \leq 5$ ($k = 6$ leads to $s = 0$ in this case).

4.3 Parameter estimation

There are two widely used methods for estimating the factor loadings and ψ_i's, the *principal factor* method and the method of maximum likelihood. Of the two, principal factor analysis is more intuitive and not tied to any particular distributional assumption, whereas maximum likelihood specifically assumes a multivariate normal distribution. On the other hand, when the multivariate normal assumption is reasonable, MLE may be expected to be a superior approach.

For the first part of this section, we describe principal factor analysis. Having noted earlier that the problem is invariant under scale changes, there is no loss of generality in assuming that the analysis is based on the correlation matrix rather than the covariance matrix. (In other words, we might as well use the correlation matrix because if we want to report the results for a different scaling we can quickly derive them from the results for the correlation matrix. This is a different situation from PCA, where analysis based on the correlation matrix is often used, but is a quite different analysis from the one based on the covariance matrix.) Let the sample correlation matrix be R with entries (r_{ij}).

First, we estimate the communalities h_i^2 in a possibly *ad hoc* way. Two possible estimators are

(a) $\hat{h}_i^2$ is the squared multiple correlation coefficient between x_i and the other $p-1$ variables,

(b) $\hat{h}_i^2 = \max_{j \neq i} |r_{ij}|$.

It follows that $1 - \hat{h}_i^2$ is an estimator of ψ_i, so we form the matrix $R - \hat{\Psi}$ by replacing each diagonal entry of R by the corresponding $\hat{h}_i^2$.

The second step of the analysis is to decompose

$$R - \hat{\Psi} = \sum_{i=1}^{p} a_i \gamma_{(i)} \gamma_{(i)}^T \tag{4.7}$$

where $\{a_i\}$ are the eigenvalues of $R - \hat{\Psi}$ and $\{\gamma_{(i)}\}$ are orthonormal eigenvectors. Without loss of generality we order $\{a_i\}$ so that $a_1 \geq a_2 \geq \ldots \geq a_p$.

Note that (4.7) is a principal components analysis performed on the matrix $R - \hat{\Psi}$. This explains the name of the method.

Given that R and $\hat{\Psi}$ are only estimates and not the true population covariance matrices, at this stage, there is no guarantee that $a_p \geq 0$. However, we are going to assume that $a_k > 0$, where k is the number of factors being fitted: if this condition is violated, we either need to use another estimator for h_i^2 or else abandon the k-factor model.

The third step of the analysis is to set $a_{k+1} = \ldots = a_p = 0$ in (4.7), to identify λ_i, the i'th column of the matrix Λ, with $\sqrt{a_i}\gamma_{(i)}$, and to re-estimate $\psi_1, \ldots, \psi_p$ as the diagonal entries of the matrix $\hat{\Psi} = R - \sum_{i=1}^{k} \lambda_i \lambda_i^T$.

Next, we describe the maximum likelihood method (MLE).

Assuming multivariate normality, the log likelihood for Σ (after maximizing with respect to μ) may be written as

$$\log L = -\frac{n}{2}\log|\Sigma| - \frac{n}{2}\mathrm{tr}(\Sigma^{-1}S)$$

where n is the sample size and S is the sample covariance matrix. (In the notation of chapter 2, for a literal maximum likelihood approach S should be S_0, the sample covariance matrix using n rather than $n-1$ in the denominator, though we shall ignore that distinction here.) We then define

$$F(\Lambda,\Psi) = \mathrm{tr}(\Sigma^{-1}S) - \log|\Sigma^{-1}S| - p, \quad \Sigma = \Lambda\Lambda^T + \Psi, \tag{4.8}$$

so that the maximum likelihood problem becomes to choose Λ and Ψ to minimize $F(\Lambda,\Psi)$ (the minimum possible value being 0, when $\Sigma = S$).

We can also write $F(\Lambda,\Psi)$ in the form $p(a-\log g-1)$, where a and g are the arithmetic and geometric means of the eigenvalues of $\Sigma^{-1}S$.

One feature that makes the MLE approach computationally feasible is that when Ψ is known, minimization of F with respect to Λ can be performed analytically. This is due to Joreskog and the algorithm is described in detail in Mardia, Kent and Bibby. Thus, the numerical optimization part of the procedure requires only an optimization in p dimensions, and this is feasible for most practical applications.

One advantage of the maximum likelihood approach is that we can also perform a test for the adequacy of the k-factor model (but still assuming multivariate normality). Consider the hypothesis H_0 that the k-factor model is correct, against the alternative that Σ is unrestricted. The log likelihood ratio statistic is

$$2\log\frac{L_1}{L_0} = nF(\hat{\Lambda},\hat{\Psi}), \tag{4.9}$$

where F is given by (4.8) and $\hat{\Lambda}$, $\hat{\Psi}$, are the MLEs. This has the usual asymptotic theory:

(i) As sample size $n \to \infty$, the distribution of (4.9) is asymptotically χ^2_s, where s is the number of degrees of freedom specified in (4.6).

(ii) The Bartlett correction: in (4.9), replace n by n', where in this case

$$n' = n - 1 - \frac{2p+5}{6} - \frac{2k}{3}.$$

This gives a better approximation to χ^2_s than (i). Mardia, Kent and Bibby (1979) suggest that method (ii) gives an adequate approximation if $n \geq p + 50$.

4.4 Other features

4.4.1 Fixing the rotation

In section 4.2 we mentioned the problem of fixing the rotation of the factors and described two artificial criteria. More widely used in practice is the Varimax principle, due to Kaiser (1958). This is an attempt to define the factors in such as way that, as far as possible, the loadings on each factor are concentrated on a small number of the original variables. More precisely, the method specifies $\Delta = \Lambda G = (\delta_{ij})$ as follows:

Let $d_{ij} = \delta_{ij}/h_i$, $\bar{d}_j = \sum_i d_{ij}^2/p$, then choose G and hence Δ to maximize

$$\phi = \sum\sum(d_{ij}^2 - \bar{d}_j)^2 = \sum\sum d_{ij}^4 - p\sum \bar{d}_j^2.$$

4.4.2 Estimating factor scores

Having estimated the factor loadings Λ and the residual variances Ψ, we still need a method for estimating the vector of factor scores f associated with the observations X on a specific individual, when the two are related by (4.3) (for simplicity, we assume $\mu = 0$ here).

Mardia, Kent and Bibby (1979) suggest two solutions to this problem. One of them effectively treats f as a fixed vector of constants and performs a generalized least squares analysis to produce the estimate

$$\hat{f} = (\Lambda^T\Psi^{-1}\Lambda)^{-1}\Lambda^T\Psi^{-1}X. \tag{4.10}$$

However, they also present an alternative "Bayesian" approach in which the prior distribution $f \sim MVN_k[0, I_k]$ is taken into account, to produce the estimate

$$f^* = (I_k + \Lambda^T\Psi^{-1}\Lambda)^{-1}\Lambda^T\Psi^{-1}X. \tag{4.11}$$

Elementary properties of these two estimators include

$$\begin{aligned} \mathrm{E}\{\hat{f} \mid f\} &= f, \\ \mathrm{E}\{f^* \mid f\} &= (I_k + \Lambda^T\Psi^{-1}\Lambda)^{-1}\Lambda^T\Psi^{-1}\Lambda f, \\ \mathrm{E}\{(\hat{f} - f)(\hat{f} - f)^T\} &= (\Lambda^T\Psi^{-1}\Lambda)^{-1}, \\ \mathrm{E}\{(f^* - f)(f^* - f)^T\} &= (I_k + \Lambda^T\Psi^{-1}\Lambda)^{-1}. \end{aligned} \tag{4.12}$$

Thus $\hat{f}$ is conditionally unbiased as an estimator of f, whereas f^* is biased. On the other hand, f^* has uniformly smaller mean squared error. This led Mardia, Kent and Bibby to conclude that there was no clear-cut choice between the two estimators.

Nevertheless, it seems to the present writer that f^* is the better motivated estimator. Equation (4.3) is reminiscent of the observation equation in the Kalman filter (where f is the unknown state of the system and X the observation) and in that case, it is standard to calculate the conditional mean of f given X. Applying the same logic here, we write the joint covariance matrix of f and X in the form

$$\begin{pmatrix} I_k & \Lambda^T \\ \Lambda & \Lambda\Lambda^T + \Sigma \end{pmatrix},$$

and noting that the means of X and f are both 0, we calculate the conditional mean (Prop. 3 from chapter 1) as

$$\mathrm{E}\{f \mid X\} = \Lambda^T(\Lambda\Lambda^T + \Sigma)^{-1}X, \tag{4.13}$$

with conditional variance $I_k - \Lambda^T(\Lambda\Lambda^T + \Sigma)^{-1}\Lambda$.

However, we also note the formula (Mardia, Kent and Bibby 1979, p. 459)

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}, \tag{4.14}$$

applicable when, say, A is $n \times n$, B is $n \times q$, C is $q \times q$ and D is $q \times n$. When n is much bigger than q and when A is easily invertible (especially, this applies if A is diagonal) then (4.14) greatly simplifies the calculation of the inverse. The proof of (4.14) follows by multiplying both sides by $A + BCD$ and then performing elementary manipulations on the right hand side.

Applying (4.14) with $A = \Sigma,\ B = \Lambda,\ C = I_k,\ D = \Lambda^T$, we have

$$(\Lambda\Lambda^T + \Sigma)^{-1} = \Sigma^{-1} - \Sigma^{-1}\Lambda(I + \Lambda^T\Sigma^{-1}\Lambda)^{-1}\Lambda^T\Sigma^{-1},$$

and after a few further manipulations it is easily checked that (4.13) agrees with (4.11), and that the conditional variance is the same as that given in (4.12). This confirms that f^* is the conditional mean of f given X in the multivariate normal distribution, and this seems to speak in its favor as practical estimator.

4.4.3 Relations between FA and PCA

The introduction to this chapter emphasized the contrasts between factor analysis and principal components analysis, but it needs to be pointed out that there are also many common features to the two methods. Given a random vector X, PCA creates an orthogonal matrix Γ such that $Y = \Gamma^T X$ is a vector of uncorrelated observations. If the coordinates of Y are ranked in descending order of variance, we may decide to retain only the first k coordinates of Y. Writing $Y^T = (\,Y_1^T \quad Y_2^T\,)$ where Y_1 and Y_2 are vectors of dimensions k and $p - k$ respectively, and performing a corresponding partition of $\Gamma = (\,\Gamma_1 \quad \Gamma_2\,)$, we have

$$X = \Gamma Y = \Gamma_1 Y_1 + \Gamma_2 Y_2$$

which might be thought of as a partition of X into a k-dimensional "signal" and a "noise" component.

On the other hand, in factor analysis, we have the decomposition

$$X = \Lambda f + u,$$

also representing a k-dimensional "signal" plus "noise", but with a rather different interpretation of the noise. Specifically, in factor analysis the decomposition is determined by a definite model, whereas in principal components analysis the decomposition is entirely data-based.

A further point is that the principal factors method of estimating a FA model works by calculating a PCA on the sample correlation matrix, after subtracting initial estimates of the noise variances $\{\psi_i\}$. When the noise variances are small, the two methods may be expected to produce almost the same answer.

Thus in spite of the differences in both the assumed model and the method of calculation, it can be seen that both methods have the same ultimate objective of approximating a p-dimensional vector by one concentrated in a k-dimensional subspace, and the choice between them may well depend on to what extent one is willing to make the additional modeling assumptions which FA requires.

4.5 Implementation in SPlus

As an example of the implementation of factor analysis in SPlus, here is a sample program to perform FA on the examinations data set discussed in Chapter 3:

```
motiv()
# Perform FA on exams data
exams<-matrix(scan(file='exams.dat'),byrow=T,ncol=5)
nr<-length(exams[,1])
dimnames(exams)<-list(1:nr,c("vec","mech","alg","anal","stat"))
exams.fa<-factanal(exams,factors=1)
print(exams.fa)
print(summary(exams.fa))
exams.fa<-factanal(exams,factors=1,method="mle")
print(exams.fa)
print(summary(exams.fa))
plot(1:88,exams.fa$scores[,1])
exams.fa<-factanal(exams,factors=2)
print(exams.fa)
print(summary(exams.fa))
exams.fa<-factanal(exams,factors=2,method="mle")
print(exams.fa)
print(summary(exams.fa))
biplot(exams.fa)
plot(1:88,exams.fa$scores[,1])
plot(1:88,exams.fa$scores[,2])
plot(loadings(exams.fa))
```

This program performs factor analysis with either 1 or 2 factors, by both the principal factors method (the default) and by MLE. As an example of the output, here is part of what is produced by the last two print statements (`exams.fa` and `summary(exams.fa)` under the two-factor model with MLE fit):

```
Sums of squares of loadings:

           Factor1       Factor2
           1.790119      1.353543

The number of variables is 5 and the number of observations is 88

Test of the hypothesis that 2 factors are sufficient
versus the alternative that more are required:
The chi square statistic is 0.07 on 1 degree of freedom.
The p-value is 0.785

Importance of factors:

                        Factor1     Factor2
SS loadings             1.7901191   1.3535427
Proportion Var          0.3580238   0.2707085
Cumulative Var          0.3580238   0.6287323

The degrees of freedom for the model is 1.
Uniquenesses:

      vec         mech        alg         anal        stat
      0.465897    0.4190561   0.1885692   0.3517931   0.4310229

Loadings:

             Factor1  Factor2
     vec     0.270    0.679
     mech    0.360    0.672
     alg     0.743    0.509
     anal    0.740    0.317
     stat    0.698    0.286
```

The results under the principal factor method are very similar — for example, in the two-factor model the uniquenesses are .471, .412, .198, .349 and .423 and the loadings are (.271, .354, .734, .742, .704) for factor 1, (.675, .680, .513, .319, .285) for factor 2.

Note that the results just quoted include a test that the two-factor model is sufficient, reporting a p-value of .785 (in other words, two factors are sufficient). In fact, a corresponding test for a one-factor model produces the p-value of .124, suggesting that this is also sufficient. The test results are produced only under the MLE method.

Finally, Figs. 4.1–4.3 show the output of the plots produced within the above program.

The interpretation of these results is similar to the interpretation of the PCAs in chapter 3, though not identical. Even for a one-factor model, the factor scores (Fig. 4.1(a)) look rather different from the leading principal component (top plot of Fig. 3.3) — the latter almost entirely preserve the original order of the students whereas the factor scores create a somewhat different ordering. A biplot of the two-factor model (Fig. (4.1(b)) again shows a clustering of the subjects with "vec" and "mech" having very similar factor scores, and similarly for "stat" and "anal". In the factor loadings (Fig. 4.3), it can be seen that factor 1 concentrates its weights on "alg", "anal" and "stat", while factor 2 gives biggest weight to "vec" and "mech". Thus the factors appear to be giving weight to contrasting skills. One clear difference between PCA and FA in this context is that, whereas the PCs come out in order of decreasing variance, the factors are equally weighted (factors 1 and 2 in Figs. 4.2 and 4.3 are completely interchangeable). This partly explains why the first PC is almost the mean of the five scores and the second PC a contrast between the two groups of subjects, whereas in FA, the two factors are rotated in such a way that each factor appears to be giving primary weight to one of the two groups of subjects.

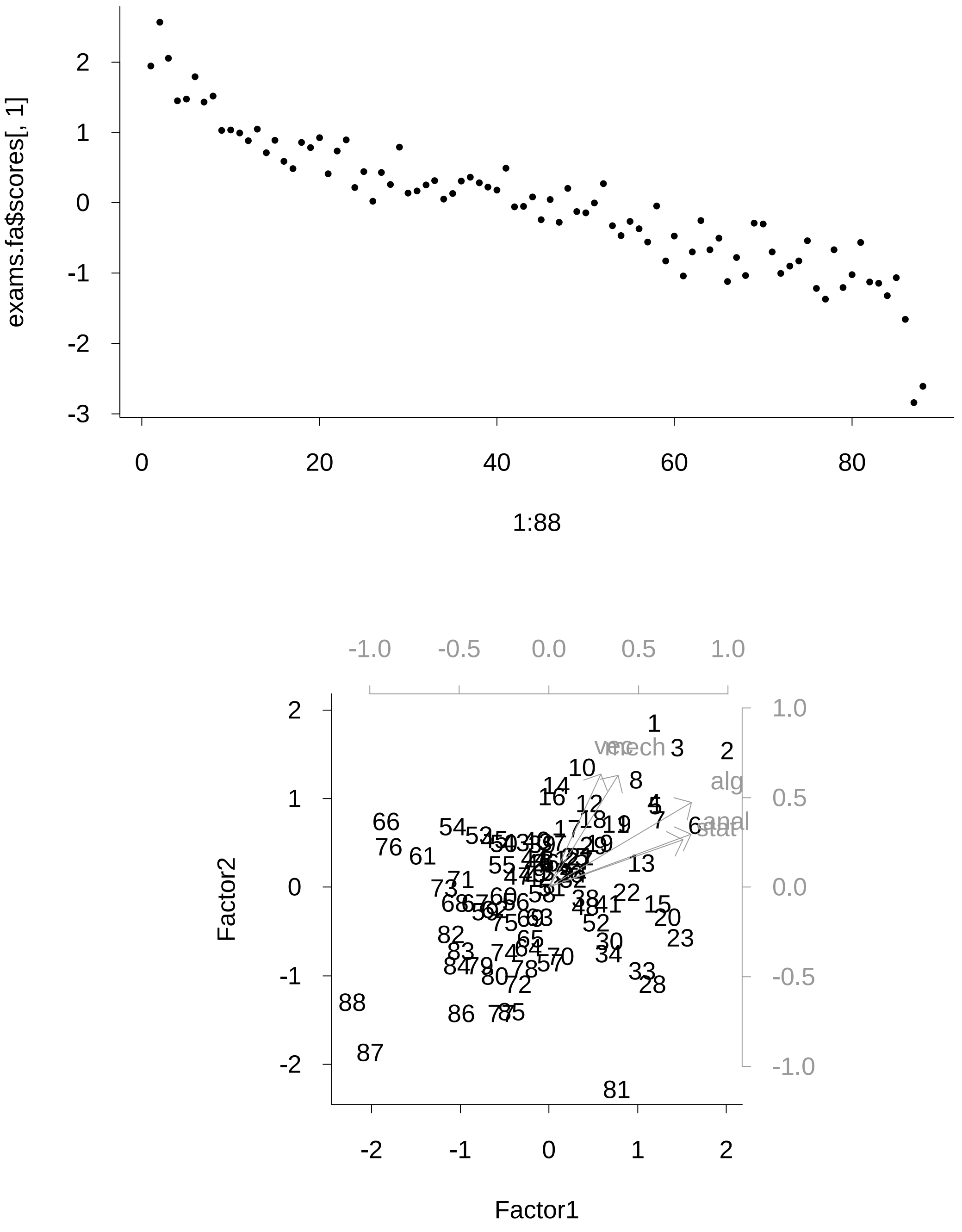

Fig. 4.1. Top plot (a): Plot of factor scores for one-factor model, MLE method. Bottom plot (b): Biplot for the two-factor model, MLE method.

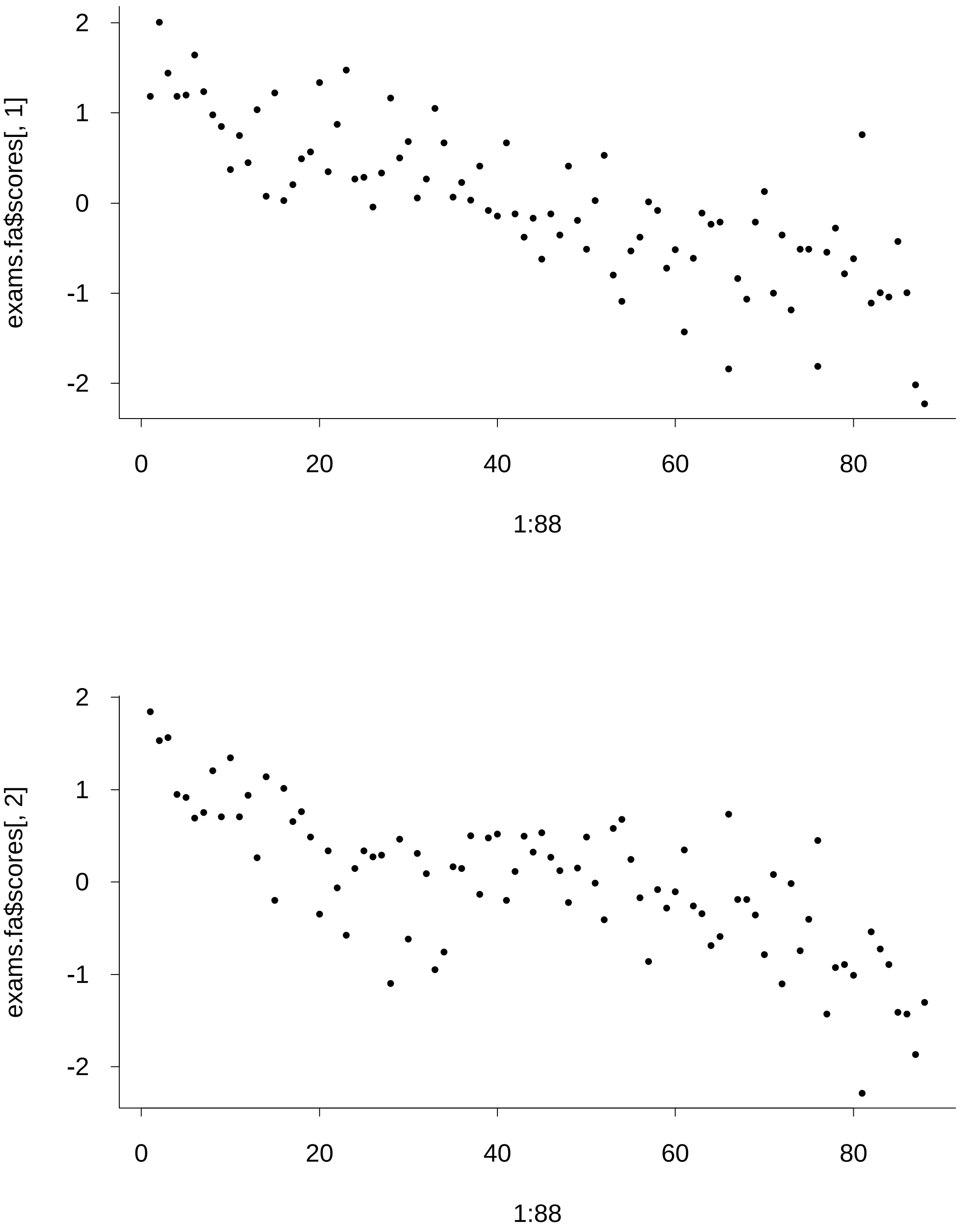

Fig. 4.2. Scores of the two factors in the two-factor model, MLE method.

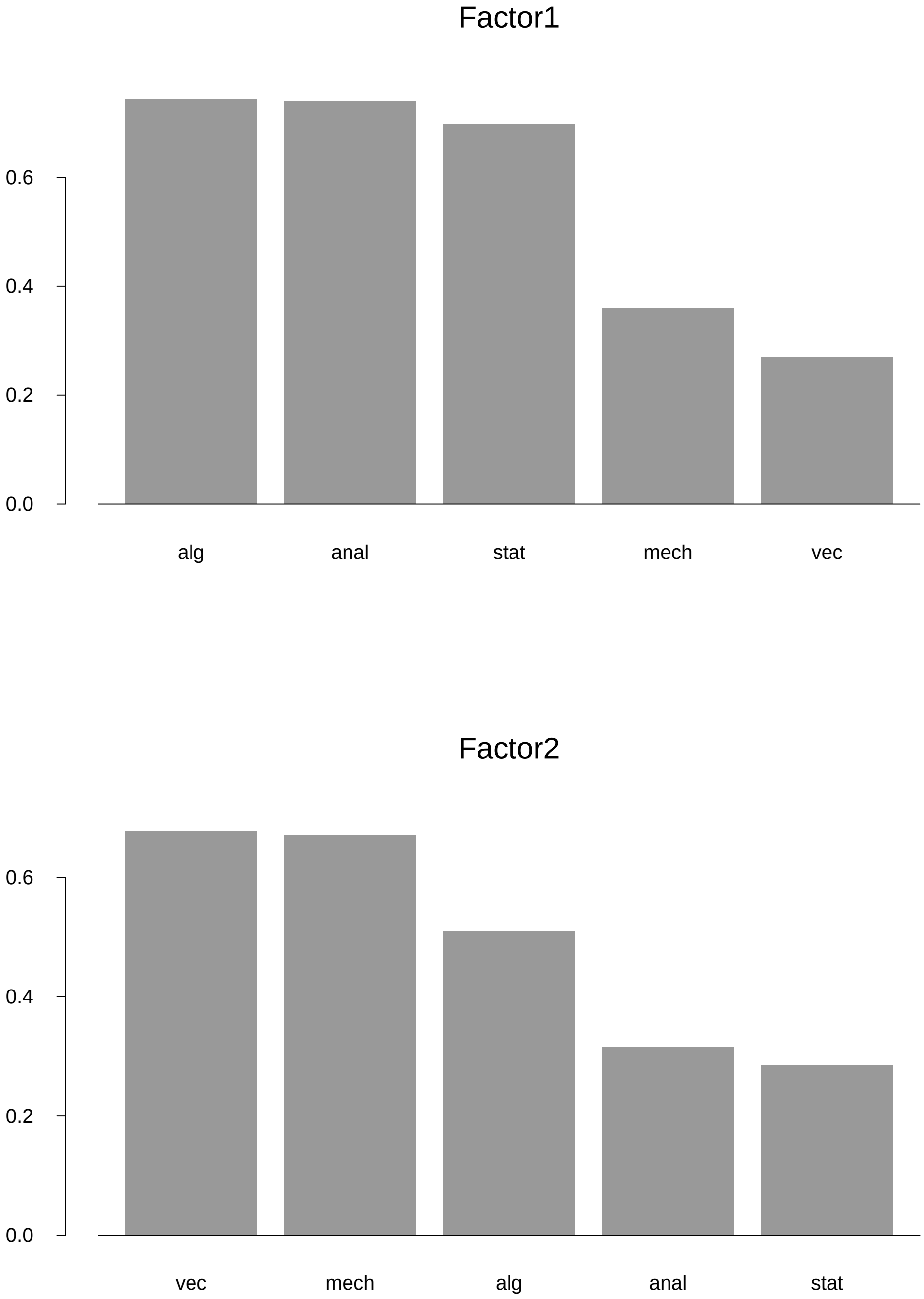

Fig. 4.3. Loadings for the two-factor model, MLE method.

5. CLUSTER ANALYSIS

Cluster analysis refers to a body of techniques concerned with grouping multidimensional data points into clusters of similar observations. The methods range from ones which are entirely *ad hoc*, not based on any statistical model or even much apparent application of any kind of statistical reasoning, to those which use well-defined models with all the usual paraphernalia of maximum likelihood estimation, testing for the number of clusters, etc., though even in model-based cases, unlike some of the other multivariate methods we have described, it is far from clear what *types* of models are appropriate and there are many different forms of cluster analysis depending on the models considered.

In this chapter we consider the two main approaches to cluster analysis, the largely model-free *hierarchical clustering* algorithms, followed by model-based approaches.

5.1 Hierarchical Clustering

Fig. 5.1 is an artifical data set due to Ruspini (1970), also discussed by Kaufman and Rousseeuw (1990). A plot of the (two-dimensional) data set reveals at least four apparent clusters.

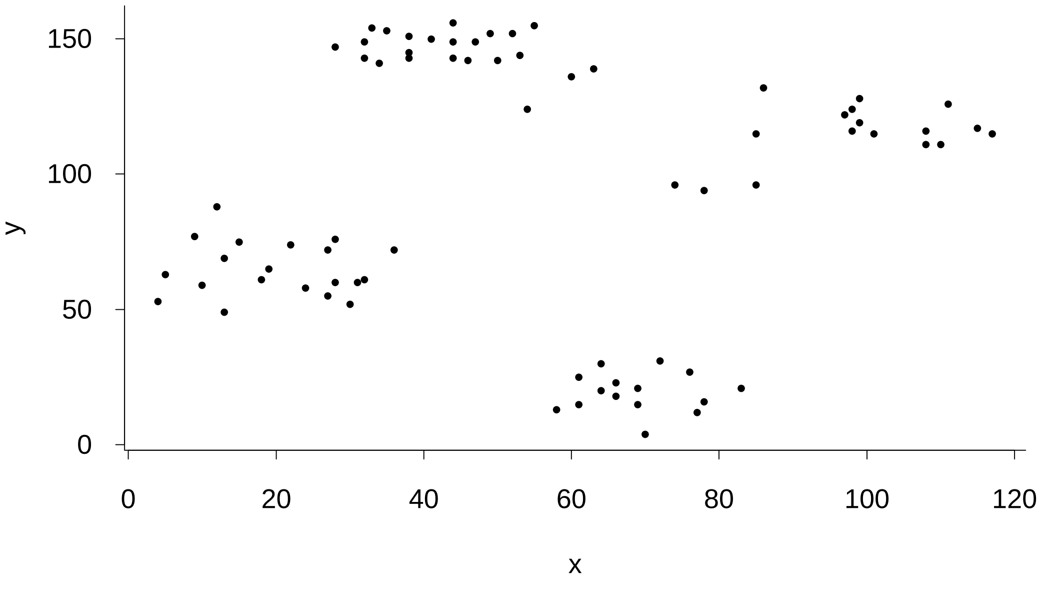

Fig. 5.1. Scatterplot of the Ruspini data.

A hierarchical clustering algorithm operates by starting from a single "cluster" covering all the data, and then successively splitting into smaller clusters according one of several criteria which we discuss in more detail below. One way to represent the result of this process is as a *dendogram* (Fig. 5.2).

Num.	x	y	Num.	x	y	Num.	x	y
1	4	53	26	41	150	51	98	124
2	5	63	27	38	145	52	99	119
3	10	59	28	38	143	53	99	128
4	9	77	29	32	143	54	101	115
5	13	49	30	34	141	55	108	111
6	13	69	31	44	156	56	110	111
7	12	88	32	44	149	57	108	116
8	15	75	33	44	143	58	111	126
9	18	61	34	46	142	59	115	117
10	19	65	35	47	149	60	117	115
11	22	74	36	49	152	61	70	4
12	27	72	37	50	142	62	77	12
13	28	76	38	53	144	63	83	21
14	24	58	39	52	152	64	61	15
15	27	55	40	55	155	65	69	15
16	28	60	41	54	124	66	78	16
17	30	52	42	60	136	67	66	18
18	31	60	43	63	139	68	58	13
19	32	61	44	86	132	69	64	20
20	36	72	45	85	115	70	69	21
21	28	147	46	85	96	71	66	23
22	32	149	47	78	94	72	61	25
23	35	153	48	74	96	73	76	27
24	33	154	49	97	122	74	72	31
25	38	151	50	98	116	75	64	30

Table 5.1. Ruspini data.

In the dendogram, the vertical scale on the left of the plot represents distance. Here, the distance between two clusters is defined as the minimum (Euclidean) distance between any two points in the cluster — as we shall see, there are a number of other definitions of distance between clusters but this is the one being used here.

For ease in interpreting the dendogram, the original data points are replotted using different plotting symbols, in Fig. 5.3.

The top horizontal bar, which corresponds to a distance of 44.9, represents the initial separation into two clusters — this separates the points marked + or × in Fig. 5.3 from the others. The next horizontal bar, at distance 40.5, separates the +'s from the ×'s. The third horizontal bar, at distance 24.0, separates the top half of the data into two clusters, with the squares and diamonds forming one cluster, and the triangles, octagons and stars

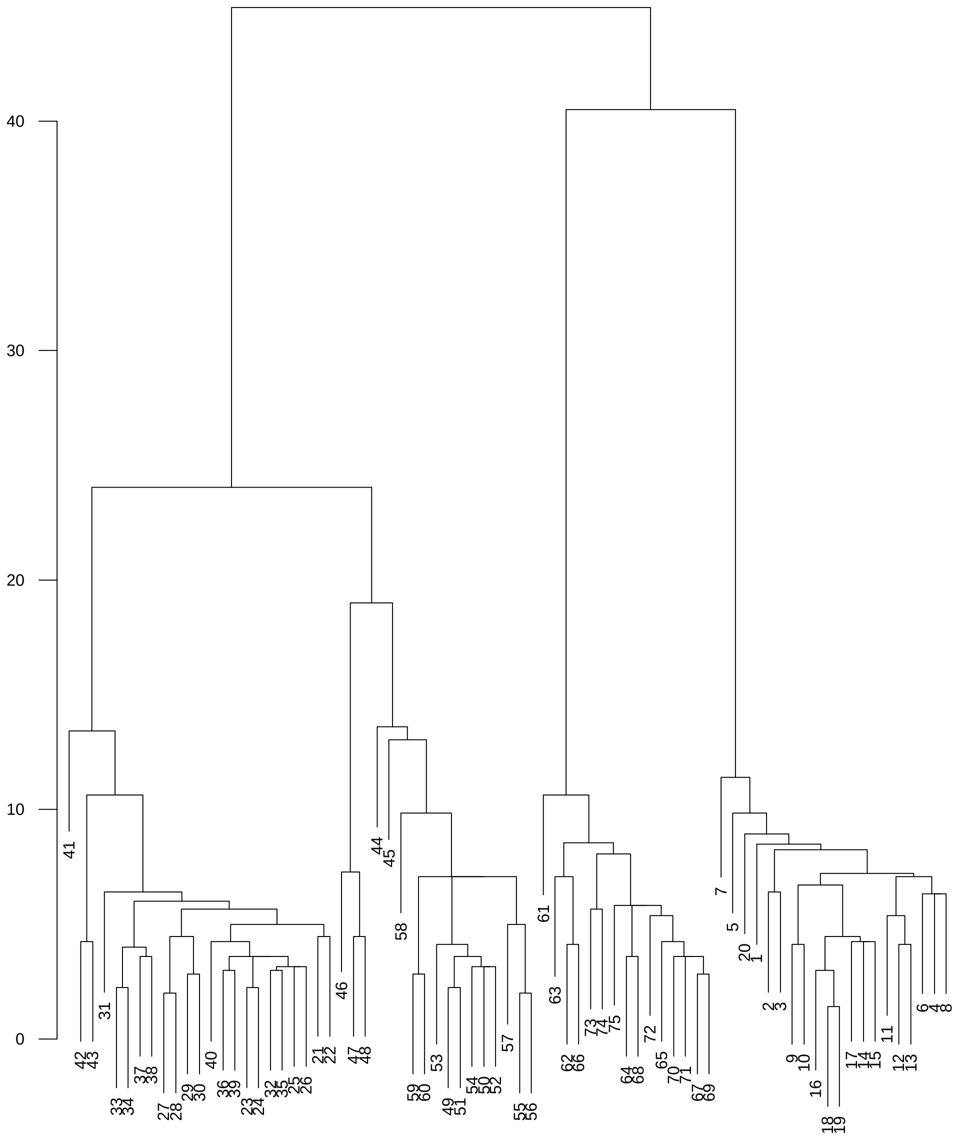

Fig. 5.2. Dendogram for the Ruspini data, hierarchical clustering, single linkage method.

the others. At this point, therefore, the four visually obvious clusters in the data have been identified by the algorithm. After that, the picture becomes more complicated. The fourth horizontal bar is at distance 19.0, and separates the stars from the triangles and octagons. Further horizontal bars occur at 13.6, 13.4, 13.0,..., until finally at distance 1.4, all 75 points lie in separate clusters. We could stop anywhere along the way, defining a critical distance d^* say, with the interpretation that two clusters whose distance apart is less than d^* are considered a single cluster. Thus, for example, setting $d^* = 20$ leads to exactly four clusters; $d^* = 15$ creates five clusters. The SPlus commands required to create these plots are

```
x<-matrix(scan(file='ruspini.dat'),ncol=2,byrow=T)
y<-dist(x,metric="euclidean")
y1<-hclust(y,method="connected")
plclust(y1)
```

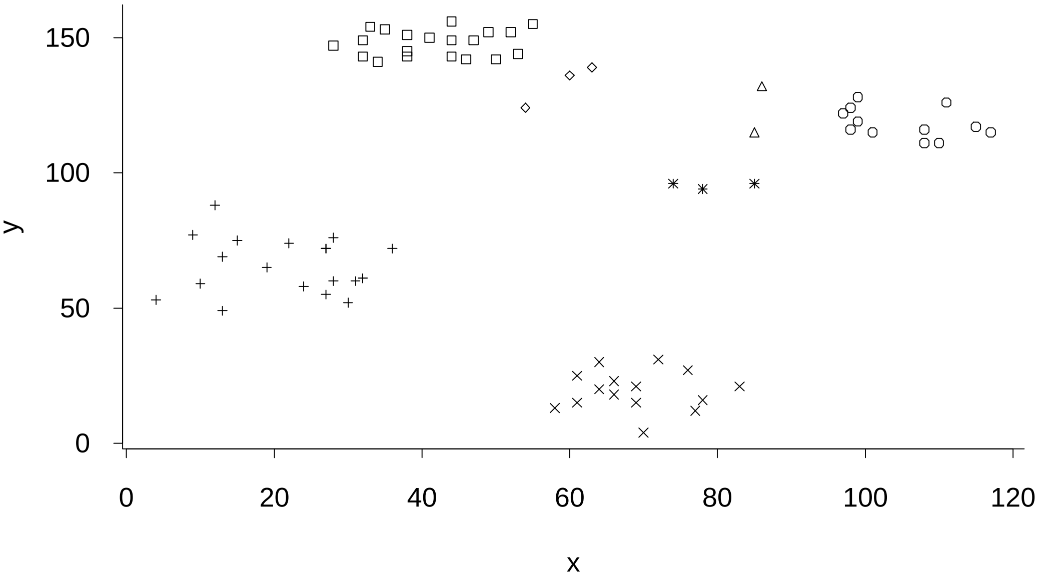

Fig. 5.3. Scatterplot of the Ruspini data with different plotting symbols.

Here, the first line reads the data into a matrix x, and the second creates a distance vector y. If x has n rows and $1 \leq i < j \leq n$, the distance between the i'th and j'th rows of x is the $\left\{n(i-1) - \frac{i(i-1)}{2} + j - i\right\}$'th entry of y. The distance here is the usual Euclidean distance between two vectors but we could also specify `metric="manhattan"` (the sum of absolute differences of the components of the two rows) or `metric="binary"` (the proportion of non-zero elements that the two vectors do not have in common). The third line applies the hierarchical clustering algorithm using the "connected" method (i.e. using the minimum distance between the points of two clusters to define the distance between the cluster — this is also called *single link clustering*) and the fourth line draws the dendogram.

Apart from the single link clustering algorithm just mentioned, there are a number of other hierarchical clustering algorithms based on different distance measures between clusters:

- *Average distance method* (`"average"` in SPlus) — the distance between two clusters is the average of the distances between the members of the clusters;

- *Complete linkage method* (`"compact"` in SPlus) — the distance between two clusters is the *maximum* of the individual distances between points of the cluster;

- *Centroid method* (not implemented as part of the SPlus `hclust` algorithm described above, but it is available through the `mclust` algorithm described in the next section, where it is given by the option `method="centroid"`) — the distance between two clusters is the distance between their centroids;

- *Sum of squares method* (also known as Ward's method, or the trace method: `method="sum of squares"` or `method="trace"` within the `mclust` algorithm) — this splits clusters in a way which minimizes the total within-cluster sum of squares.

As an example, Figs. 5.4 and 5.5 show dendograms for the Ruspini data that arise from the average distance and complete linkage methods. The average distance method agrees with the single linkage method for the first five clusters. According to the complete linkage method, the first three clusters are the same (the $+$'s, the $\times$'s and the rest), but then it splits the squares in Fig. 5.3 from the rest of the top half of the plot — in other words, the three points labelled with diamonds, formerly part of the left-hand cluster, and now in the right-hand cluster. The next step then separates the diamonds and the triangles to form one cluster, while the octagons and stars form another, and so on.

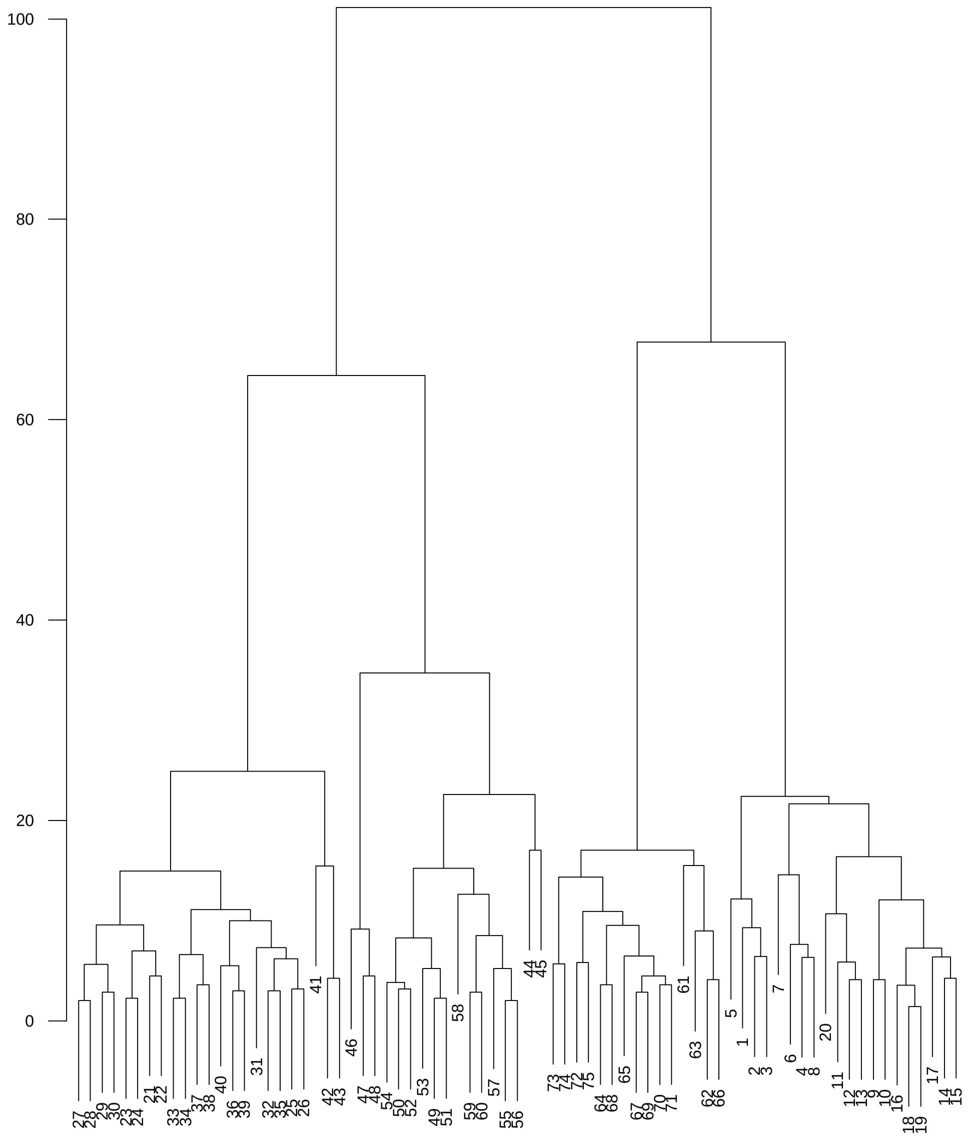

Fig. 5.4. Dendogram for the Ruspini data, hierarchical clustering, average distance method.

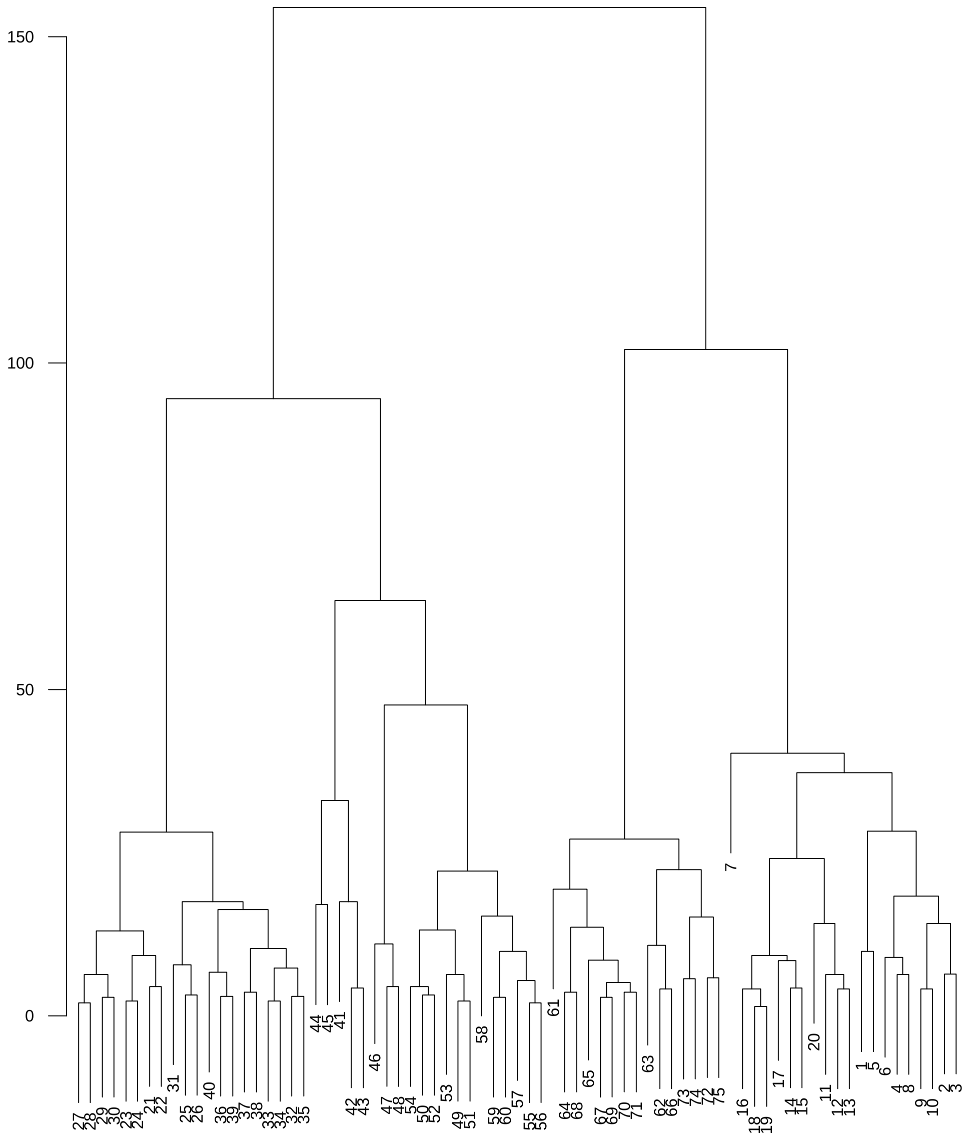

Fig. 5.5. Dendogram for the Ruspini data, hierarchical clustering, complete linkage method.

As a second example of these methods, consider the data in Table 5.2, taken from Kaufman and Rousseeuw (1990) and earlier Rousseeuw and Leroy (1987). They show the logarithmic temperature and logarithmic light intensity of a group of 47 stars in the direction of the constellation Cygnus. Fig. 5.6 shows a scatterplot of these data, in which the direction of the x axis is reversed in accordance with a common convention for this type of plot, known as the Hertzsprung-Russell diagram.

Num.	x	y	Num.	x	y	Num.	x	y
1	4.37	5.23	17	4.23	3.94	33	4.45	5.22
2	4.56	5.74	18	4.42	4.18	34	3.49	6.29
3	4.26	4.93	19	4.23	4.18	35	4.23	4.34
4	4.56	5.74	20	3.49	5.89	36	4.62	5.62
5	4.30	5.19	21	4.29	4.38	37	4.53	5.10
6	4.46	5.46	22	4.29	4.22	38	4.45	5.22
7	3.84	4.65	23	4.42	4.42	39	4.53	5.18
8	4.57	5.27	24	4.49	4.85	40	4.43	5.57
9	4.26	5.57	25	4.38	5.02	41	4.38	4.62
10	4.37	5.12	26	4.42	4.66	42	4.45	5.06
11	3.49	5.73	27	4.29	4.66	43	4.50	5.34
12	4.43	5.45	28	4.38	4.90	44	4.45	5.34
13	4.48	5.42	29	4.22	4.39	45	4.55	5.54
14	4.01	4.05	30	3.48	6.05	46	4.45	4.98
15	4.29	4.26	31	4.38	4.42	47	4.42	4.50
16	4.42	4.58	32	4.56	5.10			

Table 5.2. Astronomical data. Data are logarithmic surface temperature (x) and logarithmic light intensity (y) of 47 stars.

Initial inspection of the data shows two very obvious clusters, with four stars on the right hand side (so-called red giants) displaying completely different characteristics to the rest of the plot, all of which are “main sequence” stars. The dendogram in Fig. 5.7, which is again based on hierarchical clustering with a single-linkage model, clearly shows this split (any d^* between .46 and .78 results in a two-cluster separation) but also reveals a finer structure as the separation distance d^* is lowered, creating more clusters.

The average link method (Fig. 5.8, top) also shows the same initial split of the data, but the complete link method (Fig. 5.8, bottom) is different — here the initial split forms two large clusters and it is only at the next step that the four red giant stars are identified as a separate cluster.

Another feature of these methods that has not been mentioned is the initial *scaling* of the data — changing the relative scales of the different components will obviously affect the calculation of distances and hence the results of the analysis, so in practice, one should take care to ensure that all the components display about the same amount of variability on the chosen scale. In the astronomical example, for instance, the total range of data on the y scale is about twice that on the x scale, so there might be a case for rescaling here.

From these examples, it should be clear that there is a lot of arbitrariness about hierarchical clustering algorithms. There is no statistical theory to support which of these gives the "right" answer, and they should largely be thought of as crude data-analytic techniques. Probably their main virtue is in processing higher-dimensional data, when they might help to identify features of the data which crude inspection would not reveal.

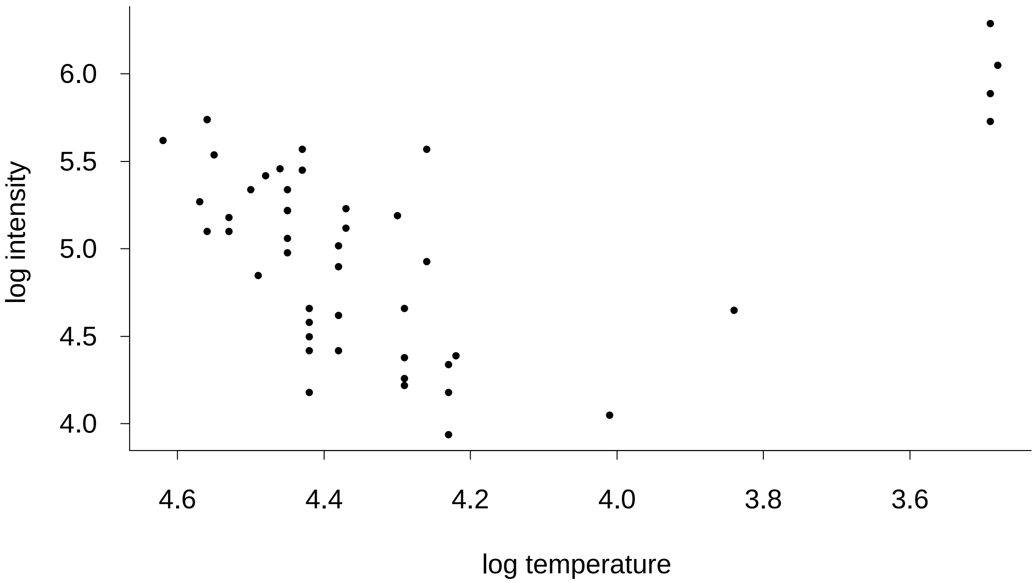

Fig. 5.6. Hertzsprung-Russell diagram based on 47 stars.

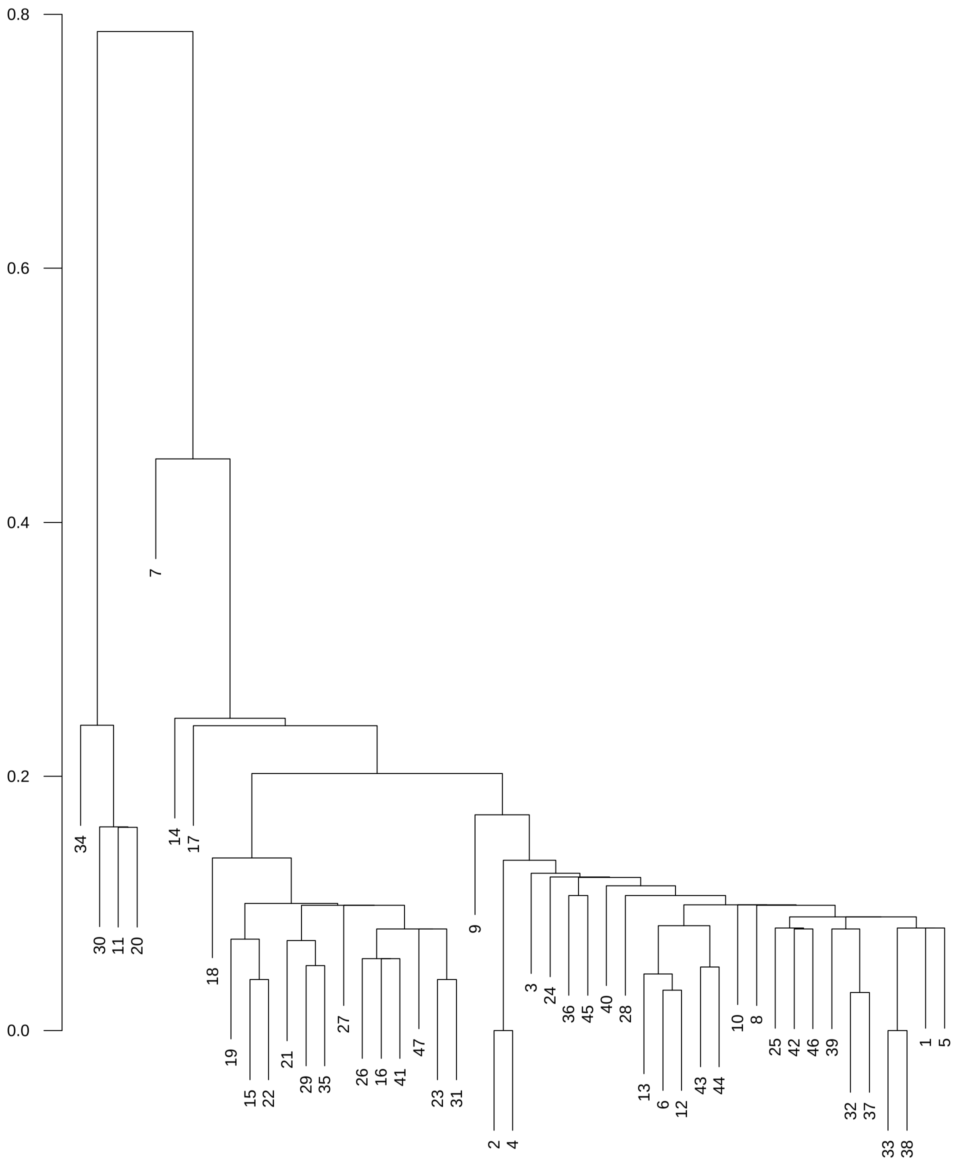

Fig. 5.7. Dendogram for the astronomical data, hierarchical clustering, single linkage method.

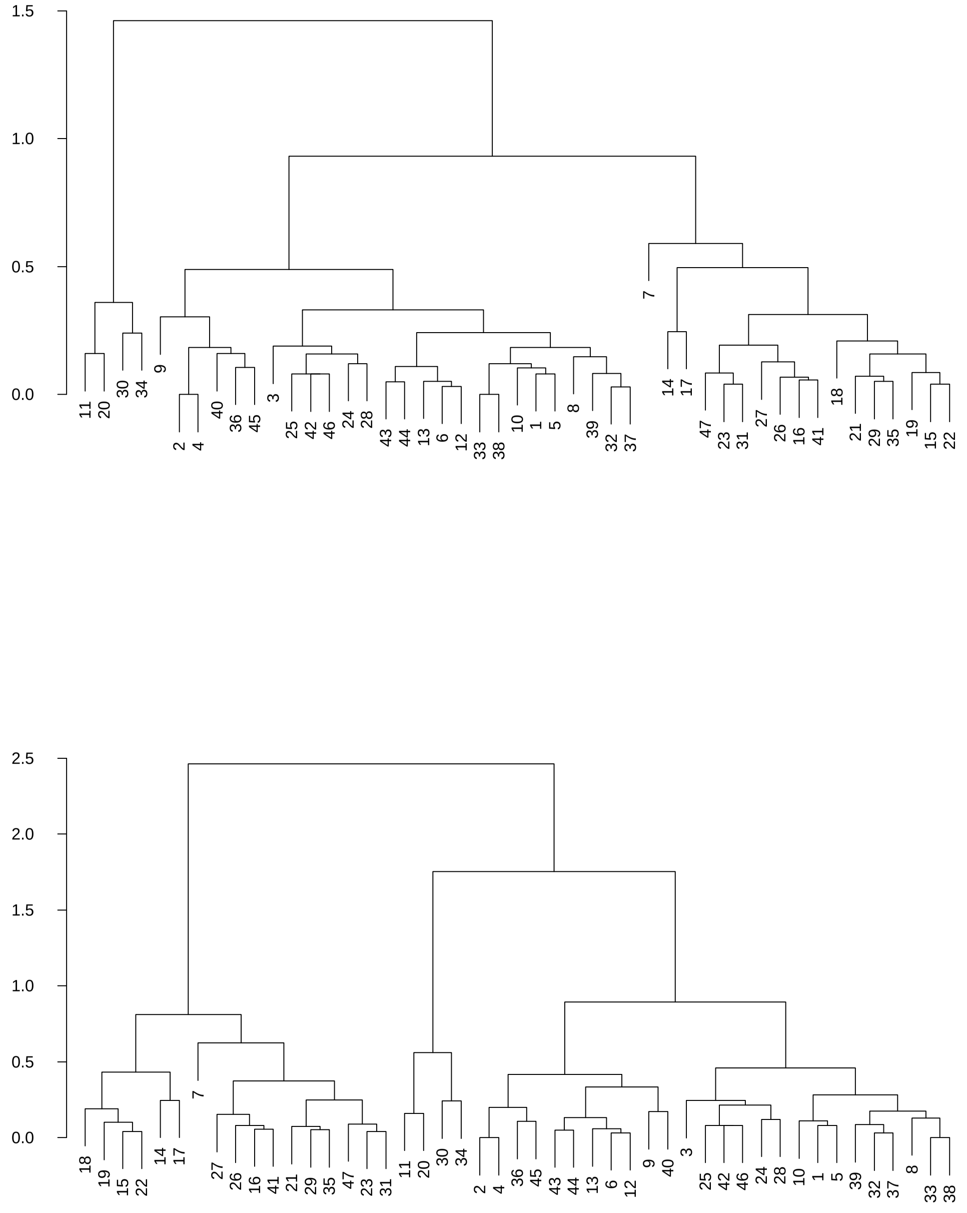

Fig. 5.8. Dendogram for the astronomical data, hierarchical clustering, average distance method (top) and complete linkage method (bottom).

5.2 Model-based clustering

The second broad strategy in cluster analysis is to assume some specific model, for example, that the observed data consist of a mixture of G multivariate normal distributions with different means and covariance matrices. We can then consider both how to allocate the observations optimally to the G groups, and also how to select the appropriate value of G when this is unknown. These techniques have been developed in a number of papers, for instance Scott and Symons (1971), Symons (1981) and in particular the paper of Banfield and Raftery (1993), which tied together a number of earlier methods and also proposed an approximate Bayes factor approach to estimating G. In more recent work, Richardson and Green (1997) proposed a fully Bayesian implementation using Markov chain Monte Carlo methods, but this is beyond the scope of the present discussion.

As a simple statement of the problem (following Mardia, Kent and Bibby (1979)), suppose we have G clusters labelled $C_1,\ C_2,...,C_G$, and an assignment function γ where $\gamma_i = k$ means that the i'th observation is assigned to the k'th group. If the G groups have population densities $f(x;\theta_k),\ 1 \le k \le G$, reflecting that the groups are all described by the same parametric family f but with different parameters $\theta_1,...,\theta_G$, then the likelihood of the model parameters is defined by

$$L(\gamma,\theta_1,...,\theta_G) = \prod_{x \in C_1} f(x;\theta_1) \prod_{x \in C_2} f(x;\theta_2) ... \prod_{x \in C_G} f(x;\theta_G). \tag{5.1}$$

In the multivariate normal case, we have $\theta_k = (\mu_k, \Sigma_k)$, where μ_k and Σ_k are the mean and covariance matrix of the k'th group. Given the allocation rule γ, we estimate μ_k and Σ_k in the obvious way, using the sample mean and covariance matrix in the k'th group. If the k'th group contains n_k observations and has sample covariance matrix S_k, then (5.1) reduces to

$$L(\gamma,\hat{\theta}_1,...,\hat{\theta}_G) \propto \prod_{k=1}^{G} |S_k|^{-n_k/2}. \tag{5.2}$$

Given G, the optimal allocation $\hat{\gamma}$ is therefore the one which maximizes (5.2). In practice, calculating $\hat{\gamma}$ exactly would involve searching over all G^n possible allocations of the n points into G groups, an impossible task when n is large, but the optimization is usually performed approximately using a hierarchical algorithm similar to the ones described in section 5.1. The difference is that (5.2) now defines a specific optimization criterion in place of the distance-based methods used earlier. Whether that is an improvement is, of course, open to debate.

The foregoing is called the unconstrained model, because there are no constraints on the μ_k and Σ_k values. In practice, a number of constraints are employed, for example that the Σ_k matrices are all the same. In that case, the maximum likelihood estimator of Σ_k is

$$W = \frac{\sum_{k=1}^{G} \sum_{i \in C_k} (X_i - \bar{X}_k)(X_i - \bar{X}_k)^T}{n},$$

with obvious notation, e.g. $\bar{X}_k$ denotes the mean of all the observations in the k'th group. The maximum likelihood allocation is then that which minimizes $|W|$.

Within the same context, we might consider testing for a single cluster, for example, by defining a null hypothesis that $\gamma_1 = ... = \gamma_n$ against that alternative that they are not all equal. The MLE of Σ under the null hypothesis is

$$T = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T}{n},$$

and the likelihood ratio test statistic is then

$$2 \log \frac{L_1}{L_0} = n \log \left\{ \max_{\gamma} \frac{|T|}{|W|} \right\}$$

but even in this simple form, the discrete maximization over γ frustrates the usual asymptotic theory so the distribution of this test statistic is intractible except by simulation. This intractability, along with the difficulty in more complicated situations of even defining a suitable hypothesis to test, means that in this field, standard hypothesis tests tend not to be used very much. Instead there has grown an extensive literature using Bayesian ideas, in particular Bayes factors.

We have already mentioned two possible forms of assumption about the Σ_k matrices: that they are unconstrained, or that they are all the same. The latter is also called the determinant rule in view of the $|W|$ optimality criterion. Another possibility is to assume $\Sigma_k = \sigma_k^2 I_p$ so that all the clusters are approximately spherical in shape, but with scaling constants σ_k^2 which may either be the same or different. When they are all the same the method is equivalent to Ward's method described in the previous section. Other possibilities are that the Σ_k are all constrained to be the same *shape* but with different orientations; again, the scaling constants may either be the same for all clusters or different. In this context, two covariance matrices may be said to be of the same shape if they have the same ratios of eigenvalues $\lambda_2/\lambda_1, ..., \lambda_p/\lambda_1$ where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$ are the ordered eigenvalues. SPlus allows the user to specify these ratios; for example taking all $\lambda_i/\lambda_1 = 0.2$ for $i = 2, ..., p$ is recommended as a reasonable compromise between forcing the clusters to be spherical and allowing them to become extremely thin and elongated. The different possibilities may be summarized in the following table (based on Banfield and Raftery 1993):

Criterion	Distribution	Orientation	Size	Shape
Ward	Spherical	N/A	Same	Same
Spherical	Spherical	N/A	Same	Same
Determinant	Ellipsoidal	Same	Same	Same
S	Ellipsoidal	Different	Same	Same
S^*	Ellipsoidal	Different	Different	Same
Unconstrained	Ellipsoidal	Different	Different	Different

Table 5.3. Different model criteria implemented in SPlus.

It remains to discuss the selection of number of clusters G. Banfield and Raftery (1993) proposed an *approximate weight of evidence* (AWE) criterion, as follows. It is based on the idea of a Bayes factor, which is defined as the ratio of posterior distributions of two models when the ratio of prior distributions is ignored. For the Bayes factor between two models containing G and $G+1$ clusters, when the model with $G+1$ clusters is formed from the model with G clusters by splitting one cluster into two, the approximate Bayes factor $B_{G,G+1}$ is given by

$$-2\log B_{G,G+1} \approx \lambda_G - \left\{\frac{3}{2} + \log(pn_{G,G+1})\right\} 2\delta_G, \tag{5.3}$$

where λ_G is the log likelihood ratio between the models with G and $G+1$ clusters, p is the dimension, $n_{G,G+1}$ is the number of observations in the merged cluster, and δ_G is the degrees of freedom (difference between the number of parameters in the G-cluster and $(G+1)$-cluster models). By adding (5.3) over $G = 1, 2, ...$, we compute an approximation to $2\log B_G$, where B_G is the approximate Bayes factor of the model with G clusters against the model with 1 cluster. The value of $2\log B_G$ is then called the approximate weight of evidence or AWE.

As pointed out by Banfield and Raftery, the whole calculation leading to the AWE involves a number of simplifying approximations and therefore it should not be treated too literally. What typically happens is that AWE rises rapidly with G when G is small, then levels off and remains approximately constant for several values of G. These values of G can then all be thought of as equally justified under the model; parsimony would suggest that one of the smaller values of G, within this group that have about the same AWE, should be chosen.

Banfield and Raftery also consider an extension which allows for noise in the form of occasional outlying observations which do not belong to any cluster. This is included in their model in the form of an additional Poisson process which is uniform over the space of observation.

Implementation in SPlus

The main function for model-based clustering is `mclust`. Unlike the `hclust` function, which takes a distance matrix as its input, `mclust` uses the data matrix directly. A typical call to the function is

```
x1<-mclust(x,method="sum",noise=F)
```

in which Ward's (sum of squares) method is applied to the data matrix `x` with no noise. To add Poisson noise, use `noise=T`. Other method options include `"spherical"`, `"determinant"`, `"S"`, `"S*"` and `"unconstrained"`, to correspond to the different criteria in Table 5.3. Also permitted as part of the `mclust` command are the various hierarchical criteria mention in section 5.1, namely "centroid", "group average link", "farthest neighbor" and "nearest neighbor". Abbreviations of these methods are allowed.

As an example, a typical call to the S^* method is

```
x4<-mclust(x,method="S*",noise=F,shape=c(1,rep(0.2,(dim(x)[2]-1))))
```

in which the shape specifies the ratio of eigenvalues referred to earlier, that $\lambda_k/\lambda_1 = 0.2$ for all $k > 1$.

A typical dendogram plot resulting from the call to mclust is

```
plclust(x1$tree)
```

though unlike the dendogram resulting from hclust, the vertical scale of the plot does not use exact distances (a point in favor of the hclust method).

Finally, we can plot the AWE for the number of clusters, e.g.

```
plot(1:10,x1$awe[1:10],xlab="Number of clusters",ylab="awe")
```

to plot the AWE for $G = 1, ..., 10$.

As an example, Fig. 5.9 shows the AWE plots for four of the model-based clustering criteria, Ward's method without noise (a), Ward's method with noise (b), the spherical method without noise (c) and the S^* method without noise (d). The dendograms are not shown since at least as far as the fourth cluster, they agree with the single linkage clustering in the cases of models (a), (b) and (c), and with complete linkage clustering in the case of model (d). In all four cases, inspection of the AWE plots by the criteria discussed earlier suggest that the plots level off at $G = 4$ clusters, supporting the four-cluster conclusion. However it should be pointed out that the conclusion is not quite so simple if one looks at the detailed numbers — for example, in method (a) the AWE rises by 16 between $G = 4$ and $G = 5$, and in method (b), the rise over the same interval is 21. Since AWE is being calculated on a scale which makes it comparable with log-likelihood values (or AIC, BIC, etc.) one might feel that such differences should not be ignored.

For the astronomical data, we show the dendograms and AWE plots for Ward's method without noise (Fig. 5.10) and for the spherical method (Fig. 5.11). In the case of Fig. 5.10, the first few AWE values are 0, 48, 100, 109, 104,..., suggesting the need for at least three clusters and maybe four. On the other hand, for Fig. 5.11 the AWE values begin 0, –2, –13, –46,..., suggesting no need for any clusters at all (or in other words, the data are consistent with one big cluster). Given results like these from the model-based approach, the reader would be entitled to ask whether the supposed *ad hockery* of the hierarchical clustering approach was such a bad thing after all.

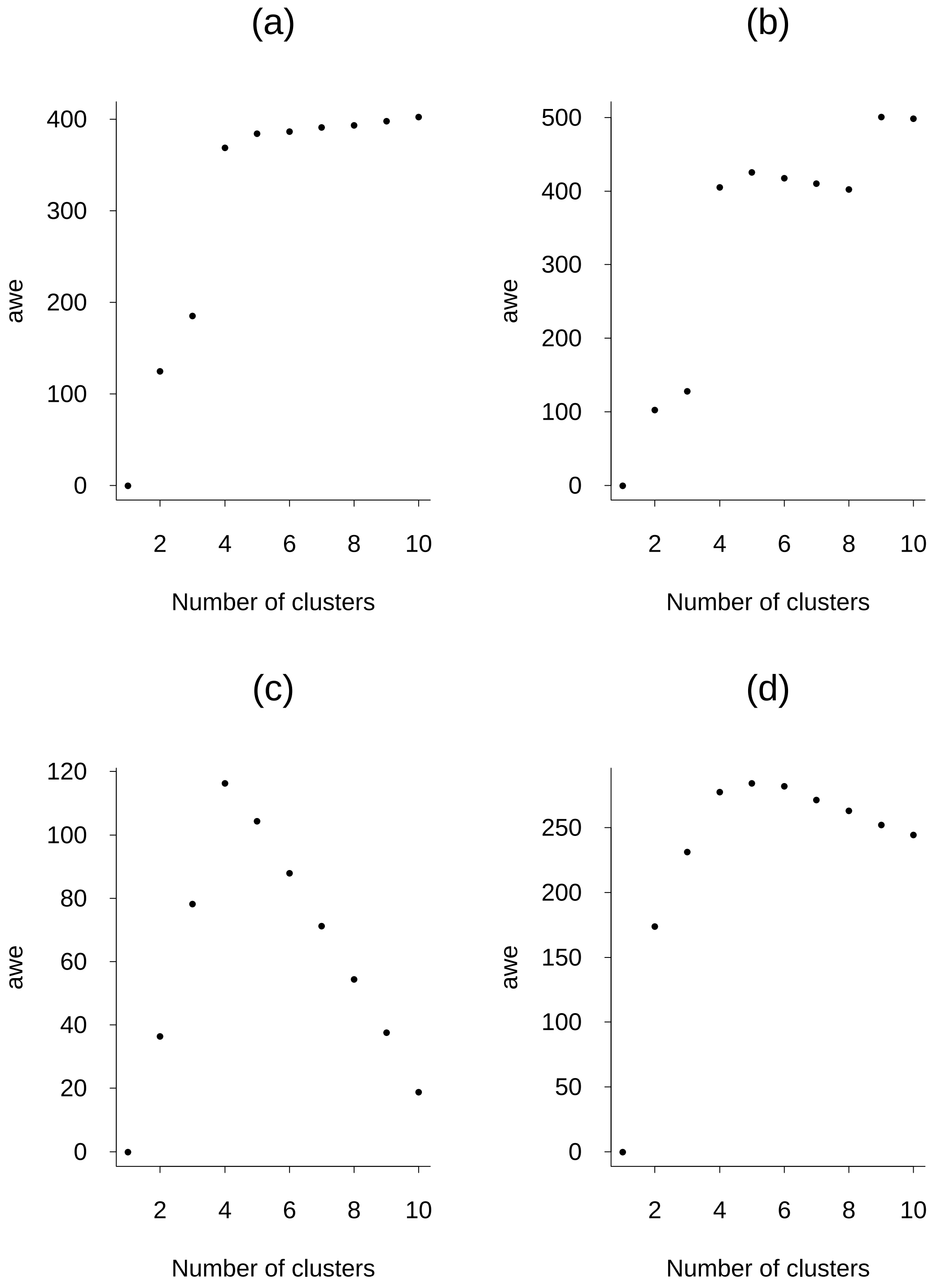

Fig. 5.9. AWE plots for the Ruspini data based on four models, (a) Ward's method without noise, (b) Ward's method with noise, (c) spherical method, (d) S^* method.

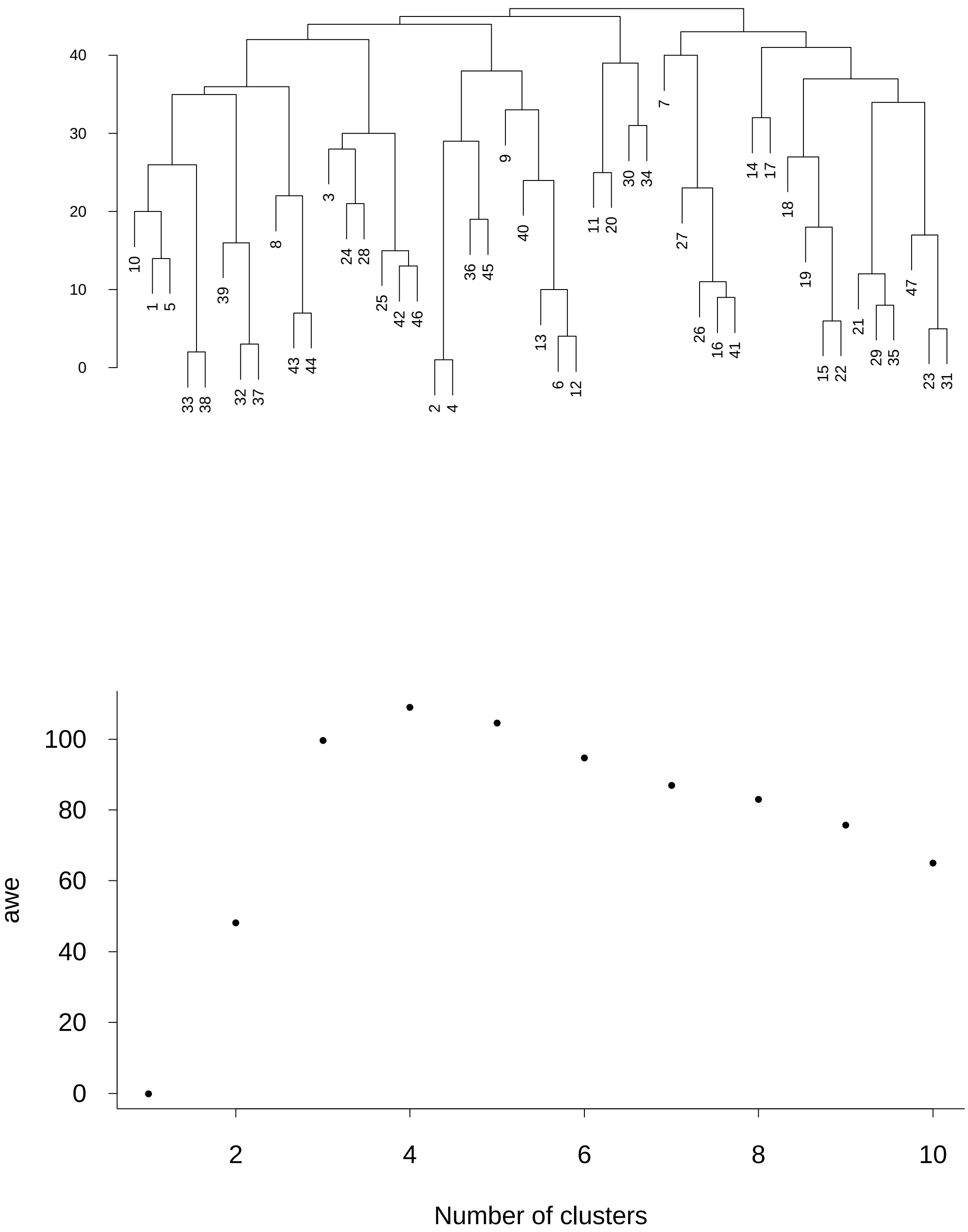

Fig. 5.10. Dendogram and AWE for the astronomical data, model-based clustering, Ward's method.

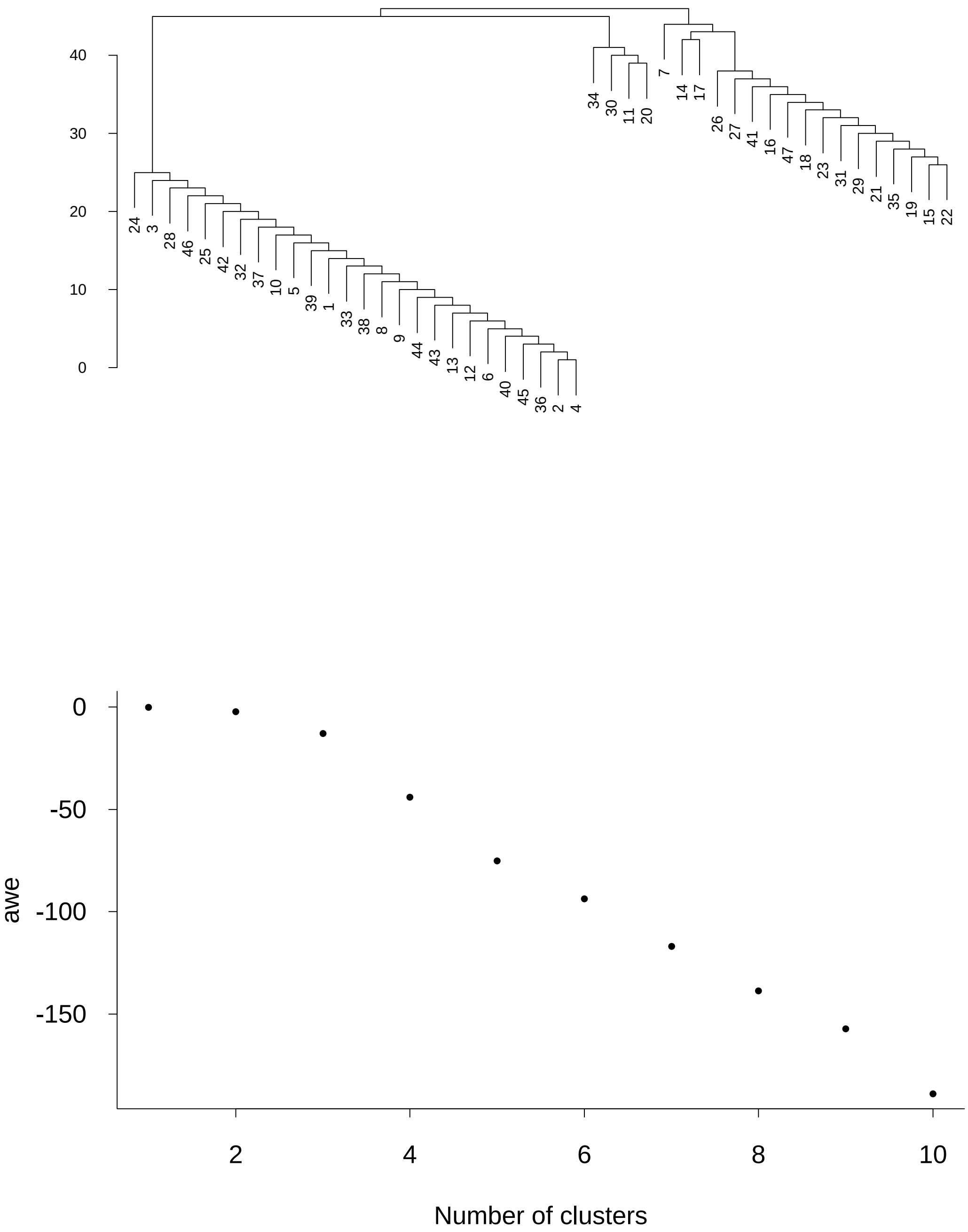

Fig. 5.11. Dendogram and AWE for the astronomical data, model-based clustering, spherical method.

5.3 Other clustering procedures

Besides the procedures which have been described, there are many other clustering algorithms which have been implemented in SPlus. For example, the `mclass` function may be used to classify objects using the output of `mclust`. Following that, `mreloc` may be used to look for improvements in the model-based criteria based on iterative relocation of elements. The `kmeans` algorithms follows a different approach to re-allocation, due to Hartigan (1975). Other procedures include `clorder` (reordering the leaves of the tree), `cutree` (creating groups from hierarchical clustering), `labclust` (labelling the leaves of the tree) and `subtree` (extracting part of a tree). All of these are described in the SPlus documentation.

Finally, we should mention that the book by Kaufman and Rousseeuw (1990) has described a completely different approach to clustering, whose routines are also implemented as part of an SPlus library.

REFERENCES

Banfield, J.D. and Raftery, A.E. (1993), Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803–822.

Berger, J.O. (1985), *Statistical Decision Theory and Bayesian Analysis* (second edition). Springer, New York.

Chatfield, C. and Collins, A.J. (1980), *Introduction to Multivariate Analysis.* Chapman and Hall, London

Fisher, R.A. (1936), The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **7**, 179–188.

Hartigan, J.A. (1975), *Clustering Algorithms.* Wiley, New York.

James, W. and Stein, C. (1961), Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium in Mathmatical Statistics and Probability*, Vol. 1. Berkeley: University of California Press, 361–379.

Jolliffe, I.T. (1986), *Principal Components Analysis.* Springer, New York.

Kaiser, H.F. (1958), The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **23**, 187–200.

Kaufman, L. and Rousseeuw, P.J. (1990), *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley, New York.

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979), *Multivariate Analysis.* Academic Press, London.

Morrison, D.F. (1990), *Multivariate Statistical Methods.* Third edition, McGraw-Hill, New York.

Richardson, S. and Green, P.J. (1997), On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J.R. Statist. Soc. B* **59**, 731–792.

Rousseeuw, P.J. and Leroy, A.M. (1987), *Robust Regression and Outlier Detection.* Wiley-Interscience, New York.

Ruspini, E.H. (1970), Numerical methods for fuzzy clustering. *Inform. Sci.* **2**, 319–350.

Scott, A.J. and Symons, M.J. (1971), Clustering methods based on likelihood ratio criteria. *Biometrics* **27**, 387–397.

Symons, M.J. (1981), Clustering criteria and multivariate normal mixtures. *Biometrics* **37**, 35–43.