Chapter 3: Principle Component Analysis

Contents

1. Introduction

The possible problem of too many variables is sometimes known as the *curse of dimensionality. principal components analysis* (PCA), is a multivariate technique with the central aim of reducing the dimensionality of a multivariate data set while accounting for as much of the original variation as possible present in the data set. This aim is achieved by transforming to a new set of variables, the principal components, that are linear combinations of the original variables, which are uncorrelated and are ordered so that the first few of them account for most of the variation in all the original variables[1]. So

Question: What is PCA?

Answer: It is a technique for transforming a set of observed correlated variables, describing the variation, into a new set of variables that are uncorrelated with one another.

Question: What is the benefit?

Answer: Dimension reduction[2]

Principal components may be useful when

- There are too many explanatory variables relative to the number of observations.
- The explanatory variables are highly correlated.

An application for principal components analysis arises in the field of economics, where complex data are often summarized by some kind of index number; for example, indices of prices, wage rates, cost of living, and so on. When assessing changes in prices over time, the economist will wish to allow for the fact that prices of some commodities are more variable than others, or that the prices of some of the commodities are considered more important than others; in each case the index will need to be weighted accordingly. In such examples, the first principal component can often satisfy the investigator's requirements.

But it is not always the first principal component that is of most interest to a researcher. A taxonomist, for example, when investigating variation in morphological measurements on animals for which all the pairwise correlations are likely to be positive, will often be more concerned with the second and subsequent components since these might provide a convenient description of aspects of an animal's "shape". The latter will often be of more interest to the researcher than aspects of an animal's "size" which here, because of the positive correlations, will be reacted in the first principal component. For essentially the same reasons, the first principal component derived from, say, clinical psychiatric scores on patients may only provide an index of the severity of symptoms, and it is the remaining components that will give the psychiatrist important information about the "pattern" of symptoms.

---

[1] PCA is a widely used data analytic technique that aims to reduce the dimensionality of the data for simplifying further analysis and visualization. It achieves its goal by constructing a sequence of orthogonal linear combinations of the original variables, called the principal components, that have maximum variance.

[2] The reduction in dimensionality that can often be achieved by a principal components analysis is possible only if the original variables are correlated; if the original variables are independent of one another a principal components analysis cannot lead to any simplification.

2. Principal components analysis

The basic goal of principal components analysis is to describe variation in a set of correlated variables, $\boldsymbol{x}' = (x_1, \dots, x_q)$, in terms of a new set of uncorrelated variables, $\boldsymbol{y}' = (y_1, \dots, y_q)$, each of which is a linear combination of the $\boldsymbol{x}$ variables. The new variables are derived in decreasing order of "importance" in the sense that y_1 accounts for as much as possible of the variation in the original data amongst all linear combinations of $\boldsymbol{x}$. Then y_2 is chosen to account for as much as possible of the remaining variation, subject to being uncorrelated with y_1, and so on. The new variables defined by this process, $y_1, \dots, y_q$, are the principal components.
The general hope of principal components analysis is that the first few components will account for a substantial proportion of the variation in the original variables, $x_1, \dots, x_q$, and can, consequently, be used to provide a convenient lower-dimensional summary of these variables that might prove useful for a variety of reasons.

1. Finding the population principle components

Let $\boldsymbol{X}$ be an $q \times 1$ random vector with mean $\boldsymbol{\mu}$ and positive definite covariance matrix $\boldsymbol{\Sigma}$. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q \ (\geq 0)$ be the latent roots[3] of $\boldsymbol{\Sigma}$ and let $\boldsymbol{H} = [\boldsymbol{h}_1 \dots \boldsymbol{h}_q]$ be an $q \times q$ orthogonal matrix[4] such that

$$\boldsymbol{H}'\boldsymbol{\Sigma}\boldsymbol{H} = \boldsymbol{\Lambda}^{5} = diag(\lambda_1, \dots, \lambda_q),$$

so that $\boldsymbol{h}_i$ is an eigenvector of $\boldsymbol{\Sigma}$ corresponding to the latent root λ_i. Now put $\boldsymbol{U} = \boldsymbol{H}'\boldsymbol{X} = (U_1, \dots, U_q)'$; then $Cov(\boldsymbol{U}) = \boldsymbol{\Lambda}$, so that $U_1, \dots, U_q$ are all
uncorrelated, and $Var(U_i) = \lambda_i, \ i = 1, \dots, q$. The components $U_1, \dots, U_q$ of $\boldsymbol{U}$ are called the *principal components* of $\boldsymbol{X}$. The first principal component is $U_1 = \boldsymbol{h}_1'\boldsymbol{X}$ and its variance is λ_1; the second principle component is $U_2 = \boldsymbol{h}_2'\boldsymbol{X}$, with variance λ_2; and etc. Moreover, the principal components have the following optimality property: The first principal component U_1**,** is the normalized linear combination of the components of $\boldsymbol{X}$ with the largest possible variance, and this maximum variance is λ_1,; then out of all normalized linear combinations of the components of $\boldsymbol{X}$ which are uncorrelated with U_1, the second principal component U_2, has maximum variance, namely, λ_2, and etc. In general, out of all normalized linear combinations which are uncorrelated with $U_1, \dots, U_{k-1}$ **,**the kth principal component U_k has maximum variance λ_k, with $k = 1, \dots, q$. To prove this, first note that the variance of an arbitrary linear function $\boldsymbol{\alpha}'\boldsymbol{X}$ of $\boldsymbol{X}$ is $Var(\boldsymbol{\alpha}'\boldsymbol{X}) = \boldsymbol{\alpha}'\boldsymbol{\Sigma}\boldsymbol{\alpha}$ and that the condition that $\boldsymbol{\alpha}'\boldsymbol{X}$ be uncorrelated with the ith principle component $U_i = h_i{}'X$ is

$$0 = Cov(\boldsymbol{\alpha}'\boldsymbol{X}, \boldsymbol{h}_i'\boldsymbol{X}) = \boldsymbol{\alpha}'\boldsymbol{\Sigma}\boldsymbol{h}_i = \lambda_i\boldsymbol{\alpha}'\boldsymbol{h}_i,$$

since $\boldsymbol{\Sigma}\boldsymbol{h}_i = \lambda_i\boldsymbol{h}_i$, so that $\boldsymbol{\alpha}$ must be orthogonal to $\boldsymbol{h}_i$. The above optimality property of the principal components is a direct consequence of the following theorem.

---

[3] Or characteristic roots Or eigenvalues
[4] $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{H}\boldsymbol{H}' = \boldsymbol{I}_q$
[5] Resulted from spectral decomposition

Theorem 2-1: Let $\boldsymbol{H} = [\boldsymbol{h_1} \dots \boldsymbol{h_q}] \in \boldsymbol{O(q)}$[6] be such that $\boldsymbol{H'\Sigma H} = \boldsymbol{\Lambda} = \boldsymbol{diag}(\boldsymbol{\lambda_1}, \dots, \boldsymbol{\lambda_q})$, where $\boldsymbol{\lambda_1} \geq \boldsymbol{\lambda_2} \geq \cdots \geq \boldsymbol{\lambda_q}$ $(\geq \boldsymbol{0})$. Then

$$\lambda_k = \max_{\substack{\boldsymbol{\alpha'\alpha}=1 \\ \boldsymbol{\alpha'h_i}=0 \\ i=1,\dots,k-1}} \boldsymbol{\alpha'\Sigma\alpha} = \boldsymbol{h_k'\Sigma h_k}$$

Proof: First note that with $\boldsymbol{\beta} = \boldsymbol{H'\alpha} = (\beta_1, \dots, \beta_q)'$ we have

$\boldsymbol{\alpha'\Sigma\alpha} = \boldsymbol{\alpha'HH'\Sigma HH'\alpha} = \boldsymbol{\beta'\Lambda\beta} = \sum_{i=1}^{q} \lambda_i \beta_i^2$.

As a consequence, if $\boldsymbol{\alpha'\alpha} = 1$, so that $\boldsymbol{\beta'\beta} = 1$, $\boldsymbol{\alpha'\Sigma\alpha} = \boldsymbol{\beta'\Lambda\beta} \leq \lambda_1 \sum_{i=1}^{q} \beta_i^2 = \lambda_1$, with equality when $\boldsymbol{\beta} = (1,0, \dots ,0)'$, i.e., when $\boldsymbol{\alpha} = \boldsymbol{h_1}$. Hence $\lambda_1 = \max_{\boldsymbol{\alpha'\alpha}=1} \boldsymbol{\alpha'\Sigma\alpha} = \boldsymbol{h_1'\Sigma h_1}$.

Next, the condition that $\boldsymbol{\alpha'h_1} = 0$ is equivalent to $\boldsymbol{\beta'H'h_1} = 0$, that is, $\beta_1 = 0$, since $\boldsymbol{\beta'H'h_1} = \beta_1 h_1'h_1 + \beta_2 h_2'h_1 + \cdots + \beta_q h_q'h_1 = \beta_1$[7]. So that, when this holds and when $\boldsymbol{\alpha'\alpha} = \boldsymbol{\beta'\beta} = 1$ we have

$\boldsymbol{\alpha'\Sigma\alpha} = \boldsymbol{\beta'\Lambda\beta} = \sum_{i=2}^{q} \lambda_i \beta_i^2 \leq \lambda_2 \sum_{i=2}^{q} \beta_i^2 = \lambda_2$, with equality when $\boldsymbol{\beta} = (0,1,0, \dots ,0)'$ i.e., when $\boldsymbol{\alpha} = \boldsymbol{h_2}$. Hence $\lambda_2 = \max_{\substack{\boldsymbol{\alpha'\alpha}=1 \\ \boldsymbol{\alpha'h_1}=0}} \boldsymbol{\alpha'\Sigma\alpha} = \boldsymbol{h_2'\Sigma h_2}$. The rest of the proof follows in exactly the same way.■

What a principal components analysis attempts to do is "explain" the variability in $\boldsymbol{X}$. To do this, some overall measure of the "total variability" in $\boldsymbol{X}$ is required; two such measures are $tr\boldsymbol{\Sigma}$ and $\det \boldsymbol{\Sigma}$, with the former being more commonly used since $\det \boldsymbol{\Sigma}$ has the disadvantage of being very sensitive to any small latent roots even though the others may be large. Note that in transforming to principal components these measures of total variation are unchanged, for

$$tr\boldsymbol{\Sigma} = tr\boldsymbol{H'\Sigma H} = tr\boldsymbol{\Lambda} = \sum_{i=1}^{q} \lambda_i,$$

and

$$\det \boldsymbol{\Sigma} = \det \boldsymbol{H'\Sigma H} = \det \boldsymbol{\Lambda} = \prod_{i=1}^{q} \lambda_i.$$

Note also that $\lambda_1 + \cdots + \lambda_k$ is the variance of the first k principal components; in a principal components analysis the hope is that for some small k, $\lambda_1 + \cdots + \lambda_k$ is close to $tr\boldsymbol{\Sigma}$. If this is so, the first k principal components explain most of the variation in $\boldsymbol{X}$ and the remaining $m - k$ principal components contribute little, since these have small variances. Of course, in most practical situations, the covariance matrix $\boldsymbol{\Sigma}$ is unknown, and hence so are its roots and vectors.

[6] Group of orthogonal matrices of order q

[7] Note that $h_i'h_i = 1$, $i = 1, \dots, q$ and $h_ih_j = 0$ for $i \neq j$.

2. Finding the sample principal components

The first principal component of the observations is that linear combination of the original variables whose sample variance is greatest amongst all possible such linear combinations. The second principal component is defined as that linear combination of the original variables that accounts for a maximal proportion of the remaining variance subject to being uncorrelated with the first principal component. Subsequent components are defined similarly.
The first principal component of the observations, y_1, is the linear combination

$$y_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1q}x_q$$

whose sample variance is greatest among all such linear combinations. Because the variance of y_1 could be increased without limit simply by increasing the coefficients $\boldsymbol{a}_1' = (a_{11}, \dots, a_{1q})$, a restriction must be placed on these coefficients as $\sum_{i=1}^{q} a_{1i}^2 = \boldsymbol{a}_1'\boldsymbol{a}_1 = 1$. To find a_1, we need to maximize $Var(y_1) = Var(\boldsymbol{a}_1'\boldsymbol{x}) = \boldsymbol{a}_1'\boldsymbol{S}\boldsymbol{a}_1$, where $\boldsymbol{x}' = (x_1, \dots, x_q)$ and $\boldsymbol{S}$ is the sample covariance matrix.
Thus we want to find the solution of the following optimization problem that can be done via the method of Lagrange multipliers

$$\max_{\boldsymbol{a}_1} \boldsymbol{a}_1'\boldsymbol{S}\boldsymbol{a}_1 \qquad \text{subject to} \qquad \boldsymbol{a}_1'\boldsymbol{a}_1 = 1.$$

According to Theorem 2-1, $\arg\left(\max_{\boldsymbol{a}_1} \boldsymbol{a}_1'\boldsymbol{S}\boldsymbol{a}_1 \; s.t. \; \boldsymbol{a}_1'\boldsymbol{a}_1 = 1\right) = \boldsymbol{a}_{max}$ where $\boldsymbol{a}_{max}$ is the eigenvector or characteristic vector of $\boldsymbol{S}$ corresponding to it largest eigenvalue.

The second principal component, y_2, is defined to be the linear combination

$$y_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2q}x_q = \boldsymbol{a}_2'\boldsymbol{x}$$

where $\boldsymbol{a}_2' = (a_{21}, \dots, a_{2q})$. That has the greatest variance subject to the following two conditions:

$$\begin{cases} \boldsymbol{a}_2'\boldsymbol{a}_2 = 1 \\ \boldsymbol{a}_2'\boldsymbol{a}_1 = 0 \end{cases}$$

(The second condition ensures that y_1 and y_2 are uncorrelated; i.e., that the sample correlation is zero.)
Similarly, the jth principal component is that linear combination $y_j = \boldsymbol{a}_j{}'\boldsymbol{x}$ that has the greatest sample variance subject to the conditions $\boldsymbol{a}_j'\boldsymbol{a}_j = 1$, and $\boldsymbol{a}_j'\boldsymbol{a}_i = 0$ $(i < j)$.
Application of the Lagrange multiplier technique demonstrates that the vector of coefficients defining the jth principal component, $\boldsymbol{a}_j$, is the eigenvector of $\boldsymbol{S}$ associated with its jth largest eigenvalue (Theorem 2-1).
If the q eigenvalues of $\boldsymbol{S}$ are denoted by $\lambda_1, \dots, \lambda_q$, then by requiring that $\boldsymbol{a}_i'\boldsymbol{a}_i = 1$ it can be shown that the variance of the ith principal component is given by λ_i. The total variance of the q principal components will equal the total variance of the original variables so that

$\sum_{i=1}^{q} \lambda_i = s_1^2 + s_2^2 + \cdots + s_q^2 = tr(\boldsymbol{S})$,
where s_i^2 is the sample variance of x_i. So,

$$Y = A'X, \Sigma = A\Lambda A'$$

Consequently, the jth principal component accounts for a proportion P_j of the total variation of the original data, where

$$P_j = \frac{\lambda_j}{tr(\boldsymbol{S})}.$$

The first m principal components, where $m < q$ account for a proportion $P^{(m)}$ of the total variation in the original data, where

$$P^{(m)} = \frac{\sum_{j=1}^{m} \lambda_j}{tr(\boldsymbol{S})}.$$

In geometrical terms, the first principal component defines the line of best fit (in the sense of minimizing residuals orthogonal to the line) to the q-dimensional observations in the sample. These observations may therefore be represented in one dimension by taking their projection onto this line; that is, finding their first principal component score. If the observations happen to be collinear in q dimensions, this representation would account completely for the variation in the data and the sample covariance matrix would have only one non-zero eigenvalue. In practice, of course, such collinearity is extremely unlikely, and an improved representation would be given by projecting the q-dimensional observations onto the space of the best fit, this being defined by the first two principal components. Similarly, the first m components give the best fit in m dimensions. If the observations fit exactly into a space of m dimensions, it would be indicated by the presence of $q - m$ zero eigenvalues of the covariance matrix. This would imply the presence of $q - m$ linear relationships between the variables. Such constraints are sometimes referred to as *structural relationships*. In practice, in the vast majority of applications of principal components analysis, all the eigenvalues of the covariance matrix will be non-zero.

One problem with principal components analysis is that it is not scale invariant.
So if we imagine a set of multivariate data where the variables are of completely different types, for example length, temperature, blood pressure, or anxiety rating, then the structure of the principal components derived from the covariance matrix will depend upon the essentially arbitrary choice of units of measurement; thus in practice, when variables are on very different scales or have very different variances, a principal components analysis of the data should be performed on the correlation matrix, not on the covariance matrix. And choosing to work with $\boldsymbol{R}$ rather than with $\boldsymbol{S}$ involves a definite but possibly arbitrary decision to make variables "equally important".

3. Principal components of bivariate data with correlation coefficient r

Assume $Corr(x_1, x_2) = r$. Then the sample correlation matrix is

$$\boldsymbol{R} = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}.$$

The eigenvalues are found as the roots of the equation $\det(R - \lambda I) = 0$. Consequently we have the following quadratic function in λ

$$(1-\lambda)^2 - r^2 = 0, \quad \Longrightarrow \quad \lambda_1 = 1 + r, \quad \lambda_2 = 1 - r.^8$$

The eigenvector corresponding to λ_1 is obtained by solving the equation

$$\boldsymbol{R}\boldsymbol{a}_1 = \lambda_1 \boldsymbol{a}_1.$$

This leads to the equations

$$\begin{cases} a_{11} + r a_{12} = (1+r) a_{11} \\ r a_{11} + a_{12} = (1+r) a_{12} \end{cases}, \quad \Longrightarrow \quad a_{11} = a_{12}.$$

If we now introduce the normalization constraint $\boldsymbol{a}_1' \boldsymbol{a}_1 = 1$, we find that $a_{11} = a_{12} = \frac{1}{\sqrt{2}}$. Similarly, we find the second eigenvector is given by $a_{21} = \frac{1}{\sqrt{2}}$ and $a_{22} = -\frac{1}{\sqrt{2}}$. The two principal components are then given by

$$y_1 = \tfrac{1}{\sqrt{2}}(x_1 + x_2), \qquad y_2 = \tfrac{1}{\sqrt{2}}(x_1 - x_2).$$

Then we have $Var(y_1) = 1 + r$, and $Var(y_2) = 1 - r$. Notice that if $r < 0$, the order of the eigenvalues and hence of the principal components is reversed; if $r = 0$, the eigenvalues are both equal to 1 and any two solutions at right angles could be chosen to represent the two components. Two further points should be noted:

- There is an arbitrary sign in the choice of the elements of $\boldsymbol{a}_i$. It is customary (but not universal) to choose a_{i1} to be positive.
- The coefficients that define the two components do not depend on r, although the proportion of variance explained by each does change with r. As r tends to 1, the proportion of variance accounted for by y_1, namely $(1 + r)/2$, also tends to one. When $r = 1$, the points all align on a straight line and the variation in the data is unidimensional.

[8] Note that $\lambda_1 + \lambda_2 = tr\boldsymbol{R}$

4. How the principal components predict the observed covariance matrix

Now we will look at how the principal components reproduce the observed covariance or correlation matrix from which they were extracted.

Let the initial vectors $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_q$, that define the principle components be used to form a $q \times q$ matrix, $\boldsymbol{A} = (\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_q)$; we assume that these are vectors extracted from the covariance matrix, $\boldsymbol{S}$, and scaled so that $\boldsymbol{a}_i'\boldsymbol{a}_i = 1$. The spectral decomposition of S is given by

$$\boldsymbol{S} = \boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{A}'$$

Suppose $\boldsymbol{a}_i^* = \lambda_i^{\frac{1}{2}}\boldsymbol{a}_i$; then

$$\boldsymbol{S} = \boldsymbol{A}^*\boldsymbol{A}^{*\prime}, \qquad \boldsymbol{A}^* = (\boldsymbol{a}_1^*, \ldots, \boldsymbol{a}_q^*)$$

If the matrix $\boldsymbol{A}_m^*$ is formed from, say, the first m components rather than from all q, then $\boldsymbol{A}_m^*\boldsymbol{A}_m^{*\prime}$ gives the predicted value of S based on these m components.

Question: How many components are needed to provide an adequate summary of a given data set?

Answers:

- Retain just enough components to explain some specified large percentage of the total variation of the original variables. Values between 70% and 90% are usually suggested, although smaller values might be appropriate as q or n, the sample size, increases.
- Exclude those principal components whose eigenvalues are less than the average, $\sum_{i=1}^{q}\frac{\lambda_i}{q}$. Since $\sum_{i=1}^{q}\lambda_i = tr\boldsymbol{S}$, the average eigenvalue is also the average variance of the original variables. This method then retains those components that account for more variance than the average for the observed variables.
- When the components are extracted from the correlation matrix, $tr\boldsymbol{R} = q$, and the average variance is therefore one, so applying the rule in the previous bullet point, components with eigenvalues less than one are excluded. But on the basis of a number of simulation studies, a more appropriate procedure would be to exclude components extracted from a correlation matrix whose associated eigenvalues are less than 0.7.
- Examination of the plot of the λ_i against i, the so called scree diagram. The number of components selected is the value of i corresponding to an "elbow" in the curve, i.e., a change of slope from "steep" to "shallow". It should also be remembered that the scree diagram suggested in the context of factor analysis rather than applied to principal components analysis.
- A modification of the scree diagram is the log eigenvalue diagram consisting of a plot of $\log\lambda_i$ against i.

5. Calculating principal components scores

If we decide that we need, say, m principal components to adequately represent our data, then we will generally wish to calculate the scores on each of these components for each individual in our sample. If, for example, we have derived the components from the covariance matrix, $\boldsymbol{S}$, then the m principal components scores for individual i with original $q \times 1$ vector of variable values $\boldsymbol{x}_i$ are obtained as

$$\begin{aligned} y_{i1} &= \boldsymbol{a}_1' \boldsymbol{x}_i \\ y_{i2} &= \boldsymbol{a}_2' \boldsymbol{x}_i \\ &\vdots \\ y_{im} &= \boldsymbol{a}_m' \boldsymbol{x}_i \end{aligned}$$

If the components are derived from the correlation matrix, then $\boldsymbol{x}_i$ would contain individual i's standardised scores for each variable.

The principal components scores calculated as above have variances equal to λ_j for $j = 1, \ldots, m$. Many investigators might prefer to have scores with mean zero and variance equal to unity. Such scores can be found as

$$\boldsymbol{z} = \boldsymbol{\Lambda}_m^{-1} \boldsymbol{A}_m' \boldsymbol{x}, \qquad \boldsymbol{\Lambda}_m = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \ddots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_m \end{pmatrix}, \quad \boldsymbol{A}_m = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m)$$

and $\boldsymbol{x}$ is the $q \times 1$ vector of standardized scores.

We should note here that the first m principal components scores are the same whether we retain all possible q components or just the first m. As we shall see later, this is not the case with the calculation of factor scores.

6. Some examples

Head data

The data in Table 3.1 give the head lengths and head breadths (in millimetres) for each of the first two adult sons in 25 families.
R code for producing Table 3.1

```
> head1 = c(191, 195, 181, 183, 176, 208, 189, 197, 188, 192, 179, 183, 174, 190, 188, 163,
195, 186, 181, 175, 192, 174, 176, 197, 190)
> breath1 = c(155, 149, 148, 153, 144, 157, 150, 159, 152, 150, 158, 147, 150, 159, 151, 137,
155, 153, 145, 140, 154, 143, 139, 167, 163)
> head2 = c(179, 201, 185, 188, 171, 192, 190, 189, 197, 187, 186, 174, 185, 195, 187, 161,
183, 173, 182, 165, 185, 178, 176, 200, 187)
> breath2 = c(145, 152, 149, 149, 142, 152, 149, 152, 159, 151, 148, 147, 152, 157, 158, 130,
158, 148, 146, 137, 152, 147, 143, 158, 150)
> head_breath = data.frame(head1, breath1, head2, breath2)
```

Here we only use the head measurements. Thus

```
> head_dat = head_breath[ , c("head1", "head2")]
```

The mean vector and covariance matrix of the head length measurements are found using

```
> colMeans (head_dat)
head1 head2
185.7 183.8

> cov (head_dat)
        head1 head2
head1 95.29  69.66
head2 69.66 100.81
```

The principal components of these data, extracted from their covariance matrix, can be found using

```
> head_pca = princomp(x = head_dat)
> head_pca
Call:
princomp(x = head_dat)
Standard deviations:
   Comp.1    Comp.2
12.690766  5.215406
 2  variables and  25 observations.

> print(summary(head_pca), loadings = TRUE)

Importance of components:
                            Comp.1    Comp.2
Standard deviation       12.6907660 5.2154059
Proportion of Variance    0.8555135 0.1444865
Cumulative Proportion     0.8555135 1.0000000
```

```
Loadings:
      Comp.1 Comp.2
head1  0.693 -0.721
head2  0.721  0.693
```

and are

$$y_1 = 0.693x_1 + 0.721x_2, \qquad y_2 = -0.721x_1 + 0.693x_2$$

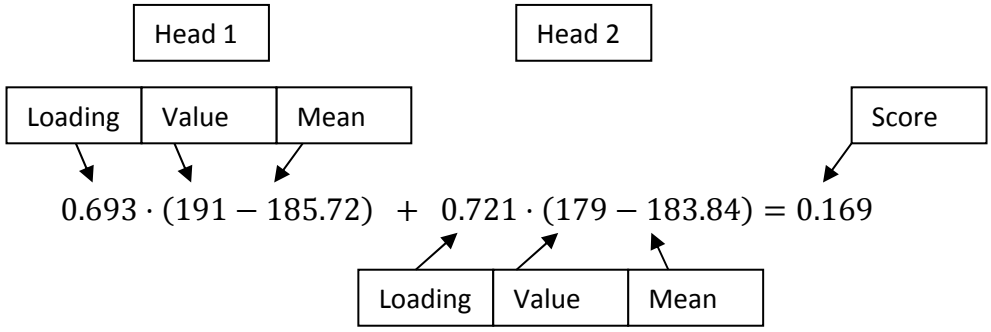
with variances 167.77 and 28.33. The first principal component accounts for a proportion

$$\frac{167.77}{167.77 + 28.33} = 0.86$$

of the total variance in the original variables. Note that the total variance of the principal components is 196.10, which as expected is equal to the total variance of the original variables.

How should the two derived components be interpreted? The first component is essentially the sum of the head lengths of the two sons, and the second component is the difference in head lengths. Perhaps we can label the first component "size" and the second component "shape", but later we will have some comments about trying to give principal components such labels.

To calculate an individual's score on a component for family 1:



and on the second component the score is

$$-0.721 \cdot (191 - 185.72) + 0.693 \cdot (179 - 183.84) = -7.61$$

We can plot the data showing the axes corresponding to the principal components. The first axis passes through the mean of the data and has slope $\frac{0.721}{0.693}$, and the second axis also passes through the mean and has slope $-\frac{0.693}{0.721}$. The plot is shown in Figure 3.2.

```
> a1 = 183.84-(0.721*185.72/0.693)
> b1 = 0.721/0.693
> a2 = 183.84-(-0.693*185.72/0.721)
> b2 = -0.693/0.721
> plot(head_dat, xlab = "First son's head length (mm)", ylab = "Second son's head length")
> abline(a1, b1)
> abline(a2, b2, lty = 2)
```

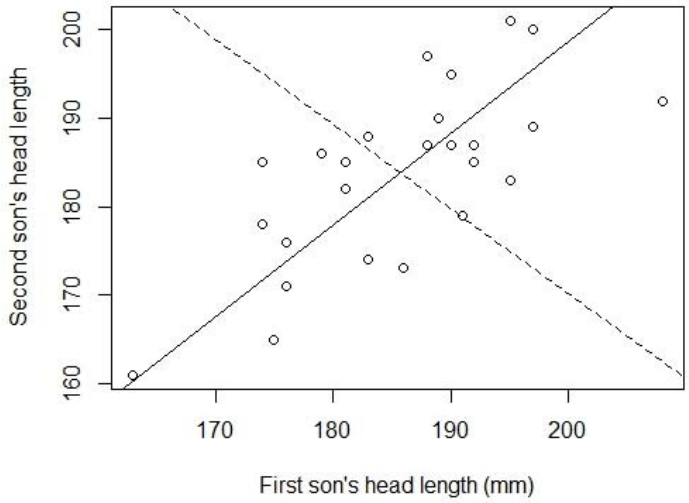


Fig. 3.2. Head length of first and second sons, showing axes corresponding to the principal components of the sample covariance matrix of the data.

This example illustrates that a principal components analysis is essentially simply a rotation of the axes of the multivariate data scatter. And we can also plot the principal components scores to give Figure 3.3.

```
> xlim <- range(head_pca$scores[,1])
> plot(head_pca$scores, xlim = xlim, ylim = xlim)
```

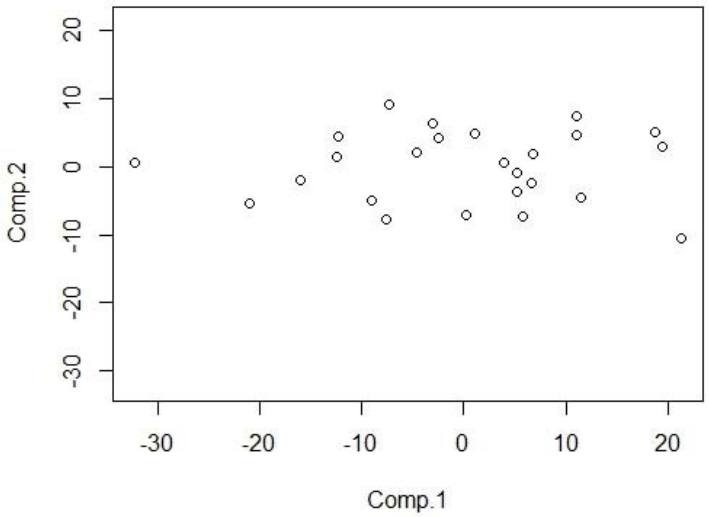


Fig. 3.3. Plot of the _rst two principal component scores for the head size data.

(Note that in this figure the range of the x-axis and the range for the y-axis have been made the same to account for the larger variance of the first principal component.)
We can use the principal components analysis of the head size data to demonstrate how the principal components reproduce the observed covariance matrix. We first need to rescale the principal components we have at this point by multiplying them by the square roots of their respective variances to give the new components

$$y_1 = 12.952(0.693x_1 + 0.721x_2), \quad i.e., \quad y_1 = 8.976x_1 + 9.338x_2$$

and

$$y_2 = 5.325(-0.721x_1 + 0.693x_2), \quad i.e., \quad y_2 = -3.837x_1 + 3.688x_2$$

leading to the matrix A_2^* as

$$A_2^* = \begin{pmatrix} 8.976 & -3.837 \\ 9.338 & 3.688 \end{pmatrix}.$$

Multiplying this matrix by its transpose should recreate the covariance matrix of the head length data (as an exercise, the readers predict covariance matrix using only the first component);

$$A_2^* A_2^{*\prime} = \begin{pmatrix} 95.29 & 69.66 \\ 69.66 & 100.81 \end{pmatrix}.$$

Olympic heptathlon results

The pentathlon[9] for women was first held in Germany in 1928. Initially this consisted of the shot put, long jump, 100 m, high jump, and javelin events, held over two days. In the 1964 Olympic Games, the pentathlon became the first combined Olympic event for women, consisting now of the 80 m hurdles, shot, high jump, long jump, and 200 m. In 1977, the 200 m was replaced by the 800 m run, and from 1981 the IAAF brought in the seven-event heptathlon in place of the pentathlon, with day one containing the events 100 m hurdles, shot, high jump, and 200 m run, and day two the long jump, javelin, and 800 m run. A scoring system is used to assign points to the results from each event, and the winner is the woman who accumulates the most points over the two days. The event made its first Olympic appearance in 1984.
In the 1988 Olympics held in Seoul, the heptathlon was won by one of the stars of women's athletics in the USA, Jackie Joyner-Kersee. The results for all 25 competitors in all seven disciplines are given in Table 3.2. R code to make Table 3.2

```
> athlete = c('Joyner-Kersee (USA)', 'John (GDR)', 'Behmer (GDR)', 'Sablovskaite (URS)',
'Choubenkova (URS)', 'Schulz (GDR)', 'Fleming (AUS)', 'Greiner (USA)', 'Lajbnerova (CZE)',
'Bouraga (URS)', 'Wijnsma (HOL)', 'Dimitrova (BUL)', 'Scheider (SWI)', 'Braun (FRG)',
'Ruotsalainen (FIN)', 'Yuping (CHN)', 'Hagger (GB)', 'Brown (USA)', 'Mulliner (GB)', 'Hautenauve
(BEL)', 'Kytola (FIN)', 'Geremias (BRA)', 'Hui-Ing (TAI)', 'Jeong-Mi (KOR)', 'Launa (PNG)')
> hurdles = c(12.69, 12.85, 13.20, 13.61, 13.51, 13.75, 13.38, 13.55, 13.63, 13.25, 13.75,
13.24, 13.85, 13.71, 13.79, 13.93, 13.47, 14.07, 14.39, 14.04, 14.31, 14.23, 14.85, 14.53,
16.42)
> highjump = c(1.86, 1.80, 1.83, 1.80, 1.74, 1.83, 1.80, 1.80, 1.83, 1.77, 1.86, 1.80, 1.86,
1.83, 1.80, 1.86, 1.80, 1.83, 1.71, 1.77, 1.77, 1.71, 1.68, 1.71, 150)
> shot = c(15.80, 16.23, 14.20, 15.23, 14.76, 13.50, 12.88, 14.13, 14.28, 12.62, 13.01, 12.88,
11.58, 13.16, 12.32, 14.21, 12.75, 12.69, 12.68, 11.81, 11.66, 12.95, 10.00, 10.83, 11.78)
```

[9] sporting contest consisting of seven different track-and-field events

```
> run200m = c(22.56, 23.65, 23.10, 23.92, 23.93, 24.65, 23.59, 24.48, 24.86, 23.59, 25.03,
23.59, 24.87, 24.78, 24.61, 25.00, 25.47, 24.83, 24.92, 25.61, 25.69, 25.50, 25.23, 26.61,
26.16)
> longjump = c(7.27, 6.71, 6.68, 6.25, 6.32, 6.33, 6.37, 6.47, 6.11, 6.28, 6.34, 6.37, 6.05, 6.12,
6.08, 6.40, 6.34, 6.13, 6.10, 5.99, 5.75, 5.50, 5.47, 5.50, 4.88)
> javelin = c(45.66, 42.56, 44.54, 42.78, 47.46, 42.82, 40.28, 38.00, 42.20, 39.06, 37.86,
40.28, 47.50, 44.58, 45.44, 38.60, 35.76, 44.34, 37.76, 35.68, 39.48, 39.64, 39.14, 39.26,
46.38)
> run800m = c(128.51, 126.12, 124.20, 132.24, 127.90, 125.79, 132.54, 133.65, 136.05,
134.74, 131.49, 132.54, 134.93, 142.82, 137.06, 146.67, 138.48, 146.43, 138.02, 133.90,
133.35, 144.02, 137.30, 139.17, 163.43)
> score = c(7291, 6897, 6858, 6540, 6540, 6411, 6351, 6297, 6252, 6252, 6205, 6171, 6137,
6109, 6101, 6087, 5975, 5972, 5746, 5734, 5686, 5508, 5290, 5289, 4566)
> heptathlon = data.frame(hurdles, highjump, shot, run200m, longjump, javelin, run800m,
score)
> rownames(heptathlon) = athlete
```

It will help to score all seven events in the same direction so that “large” values are indicative of a “better” performance.

```
> heptathlon$hurdles <- with(heptathlon, max(hurdles)-hurdles)
> heptathlon$run200m <- with(heptathlon, max(run200m)-run200m)
> heptathlon$run800m <- with(heptathlon, max(run800m)-run800m)
```

Correlation

```
> score <- which(colnames(heptathlon) == "score")
> round(cor(heptathlon[,-score]), 2)
```

	hurdles	highjump	shot	run200m	longjump	javelin	run800m
hurdles	1.00	-0.73	0.65	0.77	0.91	0.01	0.78
highjump	-0.73	1.00	-0.19	-0.32	-0.56	0.29	-0.69
shot	0.65	-0.19	1.00	0.68	0.74	0.27	0.42
run200m	0.77	-0.32	0.68	1.00	0.82	0.33	0.62
longjump	0.91	-0.56	0.74	0.82	1.00	0.07	0.70
javelin	0.01	0.29	0.27	0.33	0.07	1.00	-0.02
run800m	0.78	-0.69	0.42	0.62	0.70	-0.02	1.00

The scatterplot matrix appears in Figure 3.4.

```
> plot(heptathlon[,-score])
```

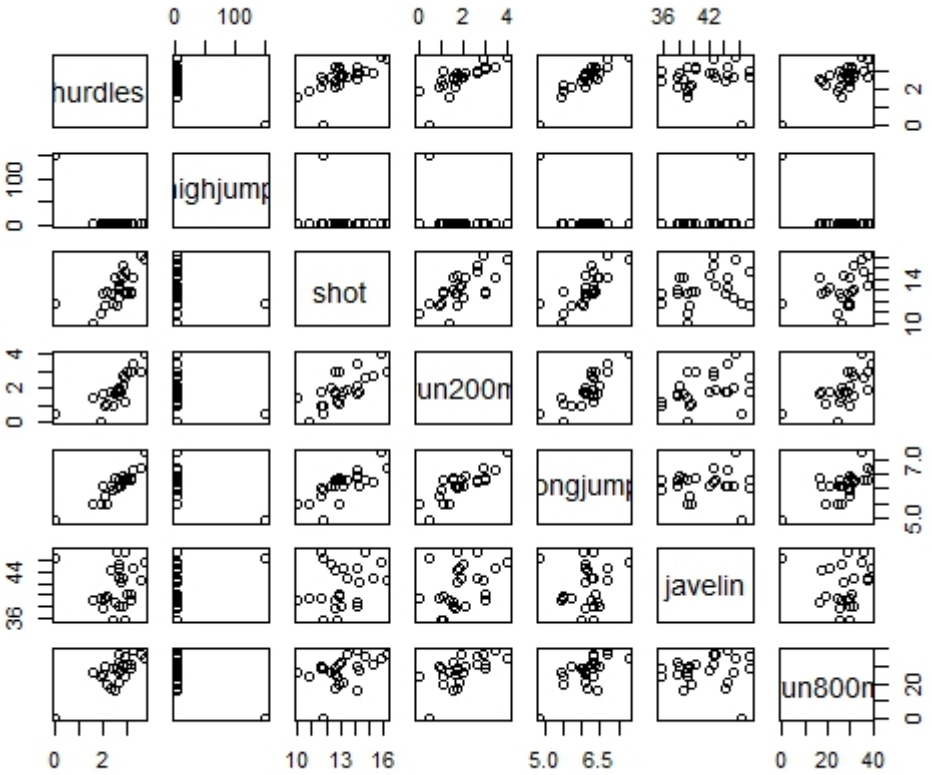


Fig. 3.4. Scatterplot matrix of the seven heptathlon events after transforming some variables so that for all events large values are indicative of a better performance.

Examination of results:

- Most pairs of events are positively correlated, some moderately (for example, high jump and shot) and others relatively highly (for example, high jump and hurdles)
- Almost all the correlations of javelin event and the others are close to zero[10]
- For all events except the javelin there is an outlier who is very much poorer than the other athletes at these six events, and this is the competitor from Papua New Guinea (PNG), who finished last in the competition in terms of points scored. But surprisingly, in the scatterplots involving the javelin, it is this competitor who again stands out, but in this case she has the third highest value for the event.

Removing the competitor from PNG;

```
> heptathlon <- heptathlon[-grep("PNG", rownames(heptathlon)),]
> score <- which(colnames(heptathlon) == "score")
> round(cor(heptathlon[,-score]), 2)
> plot(heptathlon[,-score], pch = ".", cex = 1.5)
```

[10] The javelin is a very “technical” event and perhaps the training for the other events does not help the competitors in the javelin

```
         hurdles highjump  shot run200m longjump javelin run800m
hurdles     1.00    -0.58 -0.77    0.83    -0.89   -0.33    0.56
highjump   -0.58     1.00  0.46   -0.39     0.66    0.35   -0.15
shot       -0.77     0.46  1.00   -0.67     0.78    0.34   -0.41
run200m     0.83    -0.39 -0.67    1.00    -0.81   -0.47    0.57
longjump   -0.89     0.66  0.78   -0.81     1.00    0.29   -0.52
javelin    -0.33     0.35  0.34   -0.47     0.29    1.00   -0.26
run800m     0.56    -0.15 -0.41    0.57    -0.52   -0.26    1.00
```

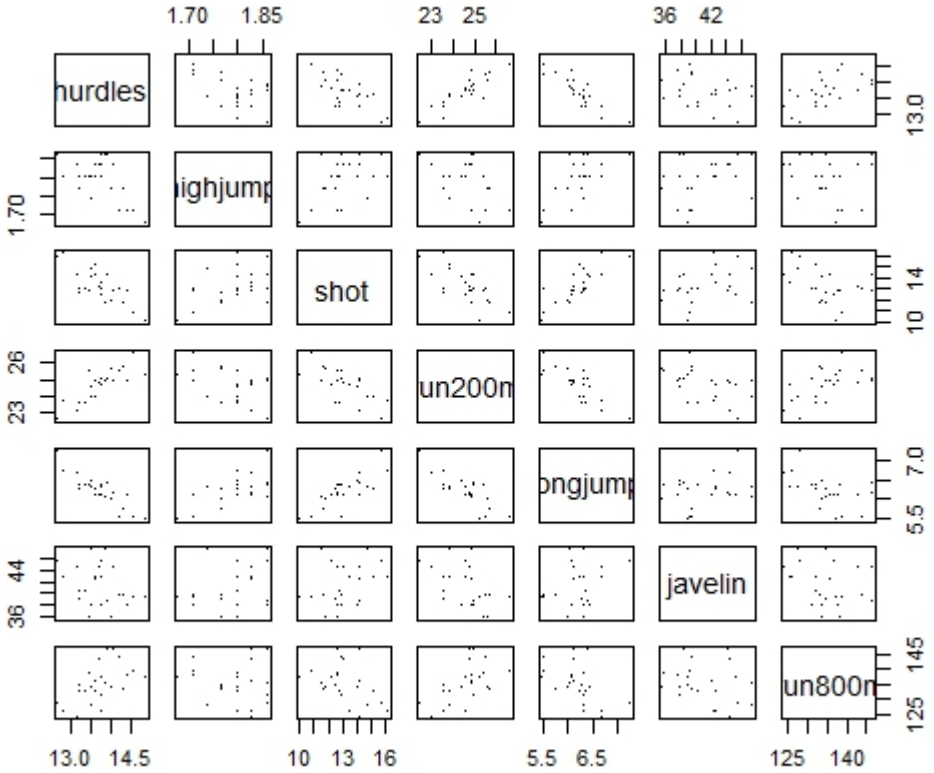


Fig. 3.5. Scatterplot matrix for the heptathlon data after removing observations of the PNG competitor.

Several of the correlations are changed to some degree from those shown before removal of the PNG competitor, particularly the correlations involving the javelin event, where the very small correlations between performances in this event and the others have increased considerably. Given the relatively large overall change in the correlation matrix produced by omitting the PNG competitor, we shall extract the principal components of the data from the correlation matrix after this omission. The principal components can now be found using

```
> heptathlon_pca <- prcomp[11](heptathlon[, -score], scale = TRUE)
> print(heptathlon_pca)

Standard deviations:
[1] 2.0793370 0.9481532 0.9109016 0.6831967 0.5461888 0.3374549 0.2620420
Rotation:
              PC1         PC2         PC3         PC4         PC5         PC6
hurdles   0.4503876 -0.05772161 -0.1739345  0.04840598  0.19889364 -0.84665086
highjump -0.3145115 -0.65133162  0.2088272  0.55694554  0.07076358 -0.09007544
shot     -0.4024884 -0.02202088  0.1534709 -0.54826705  0.67166466 -0.09886359
run200m   0.4270860 -0.18502783  0.1301287  0.23095946  0.61781764  0.33279359
longjump -0.4509639 -0.02492486  0.2697589  0.01468275 -0.12151793 -0.38294411
javelin  -0.2423079 -0.32572229 -0.8806995 -0.06024757  0.07874396  0.07193437
run800m   0.3029068 -0.65650503  0.1930020 -0.57418128 -0.31880178  0.05217664
              PC7
hurdles   0.06961672
highjump  0.33155910
shot      0.22904298
run200m  -0.46971934
longjump -0.74940781
javelin  -0.21108138
run800m  -0.07718616
```

The summary method can be used for further inspection of the details:

```
> summary(heptathlon_pca)

Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7
Standard deviation     2.0793 0.9482 0.9109 0.68320 0.54619 0.33745 0.26204
Proportion of Variance 0.6177 0.1284 0.1185 0.06668 0.04262 0.01627 0.00981
Cumulative Proportion  0.6177 0.7461 0.8646 0.93131 0.97392 0.99019 1.00000
```

The linear combination for the first principal component is

```
> a1 <- heptathlon_pca$rotation[,1]
> a1

  hurdles   highjump       shot    run200m   longjump    javelin    run800m
0.4503876 -0.3145115 -0.4024884  0.4270860 -0.4509639 -0.2423079  0.3029068
```

[11] There are three ways to perform PCA in R: **princomp()** , **prcomp()** and **pca()** in labdsv library . Essentially, they compute the same values (technically, princomp() and labdsv package computes an eigen analysis and prcomp() computes a singular value decomposition.). The **prcomp()** function is a numerically stable routine that returns a "prcomp object" that contains the square-root of the eigenvalues ("sdev"), the eigenvectors ("rotation"), and the scores. And so the preferred method is **prcomp()**. The **princomp()** function is slightly less stable, but has more features. It returns a "princomp object" that contains the square-root of the eigenvalues ("sdev"), the eigenvectors ("loadings"), the means for each variable ("center") and the scores ("scores"), as well as some other things. Typing summary(princomp) or summary(prcomp) will return the percent of variation explained.

We see that the hurdles and long jump events receive the highest weight but the javelin result is less important. For computing the first principal component, extract the first from all pre-computed principal components:

```
> predict(heptathlon_pca)[,1]
```

```
Joyner-Kersee (USA)        John (GDR)        Behmer (GDR)  Sablovskaite (URS)
       -4.757530189      -3.147943402      -2.926184760      -1.288135516
 Choubenkova (URS)        Schulz (GDR)       Fleming (AUS)       Greiner (USA)
       -1.503450994      -0.958467101      -0.953445060      -0.633239267
  Lajbnerova (CZE)       Bouraga (URS)       Wijnsma (HOL)      Dimitrova (BUL)
       -0.381571974      -0.522322004      -0.217701500      -1.075984276
    Scheider (SWI)        Braun (FRG)  Ruotsalainen (FIN)        Yuping (CHN)
        0.003014986       0.109183759       0.208868056       0.232507119
      Hagger (GB)        Brown (USA)      Mulliner (GB)    Hautenauve (BEL)
        0.659520046       0.756854602       1.880932819       1.828170404
     Kytola (FIN)      Geremias (BRA)       Hui-Ing (TAI)      Jeong-Mi (KOR)
        2.118203163       2.770706272       3.901166920       3.896847898
```

The first two components account for 75% of the variance. A barplot of each component's variance (see Figure 3.6) shows how the first two components dominate.

```
> plot(heptathlon_pca)
```

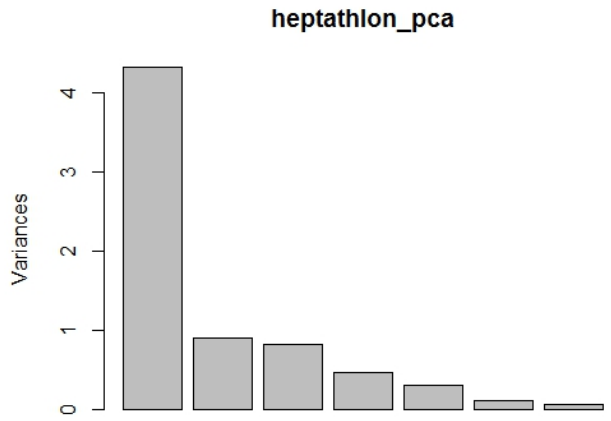


Fig. 3.6. Barplot of the variances explained by the principal components (with observations for PNG removed).

The correlation between the score given to each athlete by the standard scoring system used for the heptathlon and the first principal component score can be found from

```
> cor(heptathlon$score, heptathlon_pca$x[,1])
[1] -0.9931168[12]
```

[12] The fact that the correlation is negative is unimportant here because of the arbitrariness of the signs of the coefficients defining the first principal component; it is the magnitude of the correlation that is important.

This implies that the first principal component is in good agreement with the score assigned to the athletes by official Olympic rules; a scatterplot of the official score and the first principal component is given in Figure 3.7.

```
> plot(heptathlon$score, heptathlon_pca$x[,1])
```

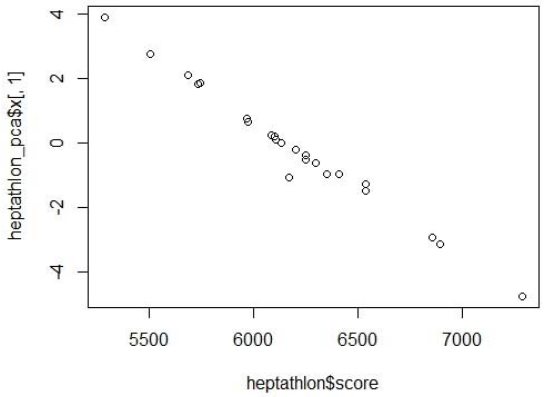


Fig. 3.7. Scatterplot of the score assigned to each athlete in 1988 and the first principal component.

Air pollution in US cities

The data were originally collected to investigate the determinants of pollution, presumably by regressing SO2 on the six other variables. Here, however, we shall examine how principal components analysis can be used to explore various aspects of the data, and will then look at how such an analysis can also be used to address the determinants of pollution question.
To begin we shall ignore the SO2 variable and concentrate on the others, two of which relate to human ecology (popul, manu) and four to climate (temp, Wind, precip, predays). A case can be made to use negative temperature values in subsequent analyses since then all six variables are such that high values represent a less attractive environment. This is, of course, a personal view, but as we shall see later, the simple transformation of temp does aid interpretation.

```
> data("USairpollution", package = "HSAUR2")
> panel.hist <- function(x, ...) {
                                    usr <- par("usr"); on.exit(par(usr))
                                    par(usr = c(usr[1:2], 0, 1.5) )
                                    h <- hist(x, plot = FALSE)
                                    breaks <- h$breaks; nB <- length(breaks)
                                    y <- h$counts; y <- y/max(y)
                                    rect(breaks[-nB], 0, breaks[-1], y, col="grey", ...)
                                  }
> USairpollution$negtemp <- USairpollution$temp * (-1)
> USairpollution$temp <- NULL
> pairs(USairpollution[,-1], diag.panel = panel.hist, pch = ".", cex = 1.5)
```

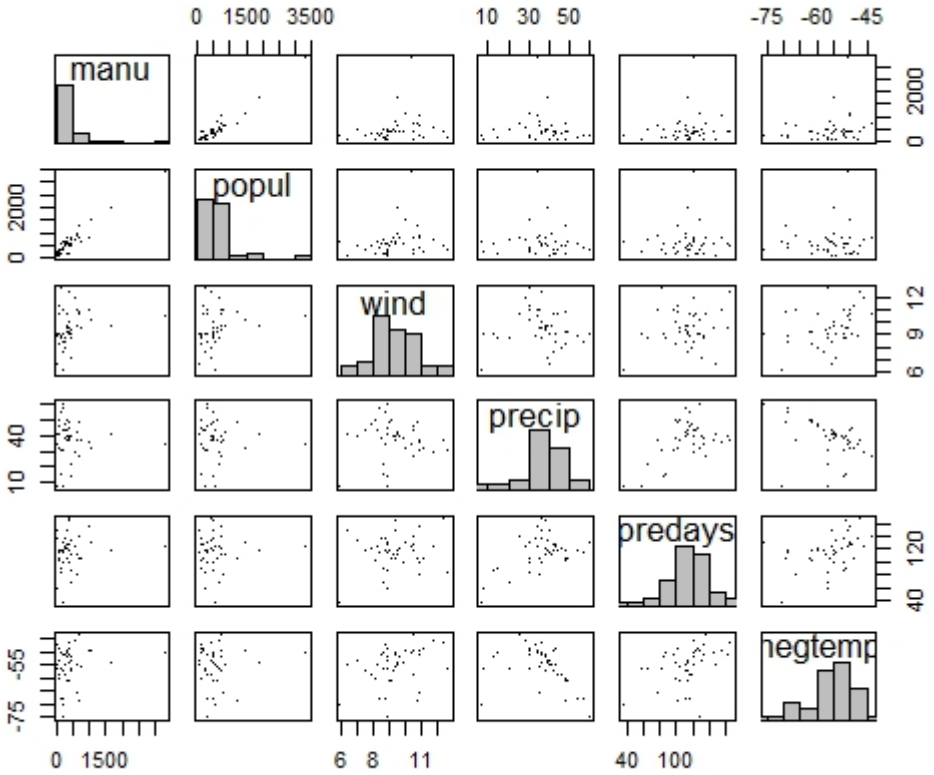


Fig. 3.8. Scatterplot matrix of six variables in the air pollution data.

A clear message from Figure 3.8 is that there is at least one city, and probably more than one, that should be considered an outlier. On the manu variable, for example, Chicago, with a value of 3344, has about twice as many manufacturing enterprises employing 20 or more workers as the city with the second highest number (Philadelphia). We shall return to this potential problem later in the chapter, but for the moment we shall carry on with a principal components analysis of the data for all 41 cities.

For the data in Table 1.5, it seems necessary to extract the principal components from the correlation rather than the covariance matrix, since the six variables to be used are on very different scales. The correlation matrix and the principal components of the data can be obtained in R using the following command line code:

```
> cor(USairpollution[,-1])
             temp       manu       popul       wind      precip     predays
temp     1.00000000 -0.19004216 -0.06267813 -0.34973963  0.38625342 -0.43024212
manu    -0.19004216  1.00000000  0.95526935  0.23794683 -0.03241688  0.13182930
popul   -0.06267813  0.95526935  1.00000000  0.21264375 -0.02611873  0.04208319
wind    -0.34973963  0.23794683  0.21264375  1.00000000 -0.01299438  0.16410559
precip   0.38625342 -0.03241688 -0.02611873 -0.01299438  1.00000000  0.49609671
predays -0.43024212  0.13182930  0.04208319  0.16410559  0.49609671  1.00000000

> usair_pca <- princomp(USairpollution[,-1], cor = TRUE)
> summary(usair_pca, loadings = TRUE)
```

```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5      Comp.6
Standard deviation     1.4819456 1.2247218 1.1809526 0.8719099 0.33848287 0.185599752
Proportion of Variance 0.3660271 0.2499906 0.2324415 0.1267045 0.01909511 0.005741211
Cumulative Proportion  0.3660271 0.6160177 0.8484592 0.9751637 0.99425879
1.000000000
```

```
Loadings:[13]
       Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
temp    0.330 -0.128  0.672 -0.306 -0.558 -0.136
manu   -0.612 -0.168  0.273  0.137  0.102 -0.703
popul  -0.578 -0.222  0.350               0.695
wind   -0.354  0.131 -0.297 -0.869 -0.113
precip         0.623  0.505 -0.171  0.568
predays -0.238  0.708         0.311 -0.580
```

One thing to note about the correlations is the very high values for manu and popul. We see that the first three components all have variances (eigenvalues) greater than one and together account for almost 85% of the variance of the original variables. Scores on these three components might be used to graph the data with little loss of information.

You might be tempted to search for an interpretation of the derived components that allows them to be "labeled" in some sense. This requires examining the coefficients defining each component. We see that the first component might be regarded as some index of "quality of life", with high values indicating a relatively poor environment. The second component is largely concerned with a city's rainfall having high coefficients for precip and predays and might be labelled as the "wet weather" component. Component three is essentially a contrast between precip and negtemp and will separate cities having high temperatures and high rainfall from those that are colder but drier. A suitable label might be simply "climate type".

Main goal when collecting the air pollution data:

determining which of the climate and human ecology variables are the best predictors of the degree of air pollution in a city as measured by the sulphur dioxide content of the air. This question would normally be addressed by multiple linear regression, but there is a potential problem with applying this technique to the air pollution data, and that is the very high correlation between the manu and popul variables. We might, of course, deal with this problem by simply dropping either manu or popul, but here we will consider a possible alternative approach, and that is regressing the SO2 levels on the principal components derived from the six other variables in the data (see Figure 3.10).

```
> out <- sapply(1:6, function(i) {
                                        plot(USairpollution$SO2,usair_pca$scores[,i],
                                        xlab = paste("PC", i, sep = ""),
                                        ylab = "Sulphur dioxide concentration")
                                       })
```

[13] these are scaled so that their sums of squares equal unity-"blanks" indicate near zero values

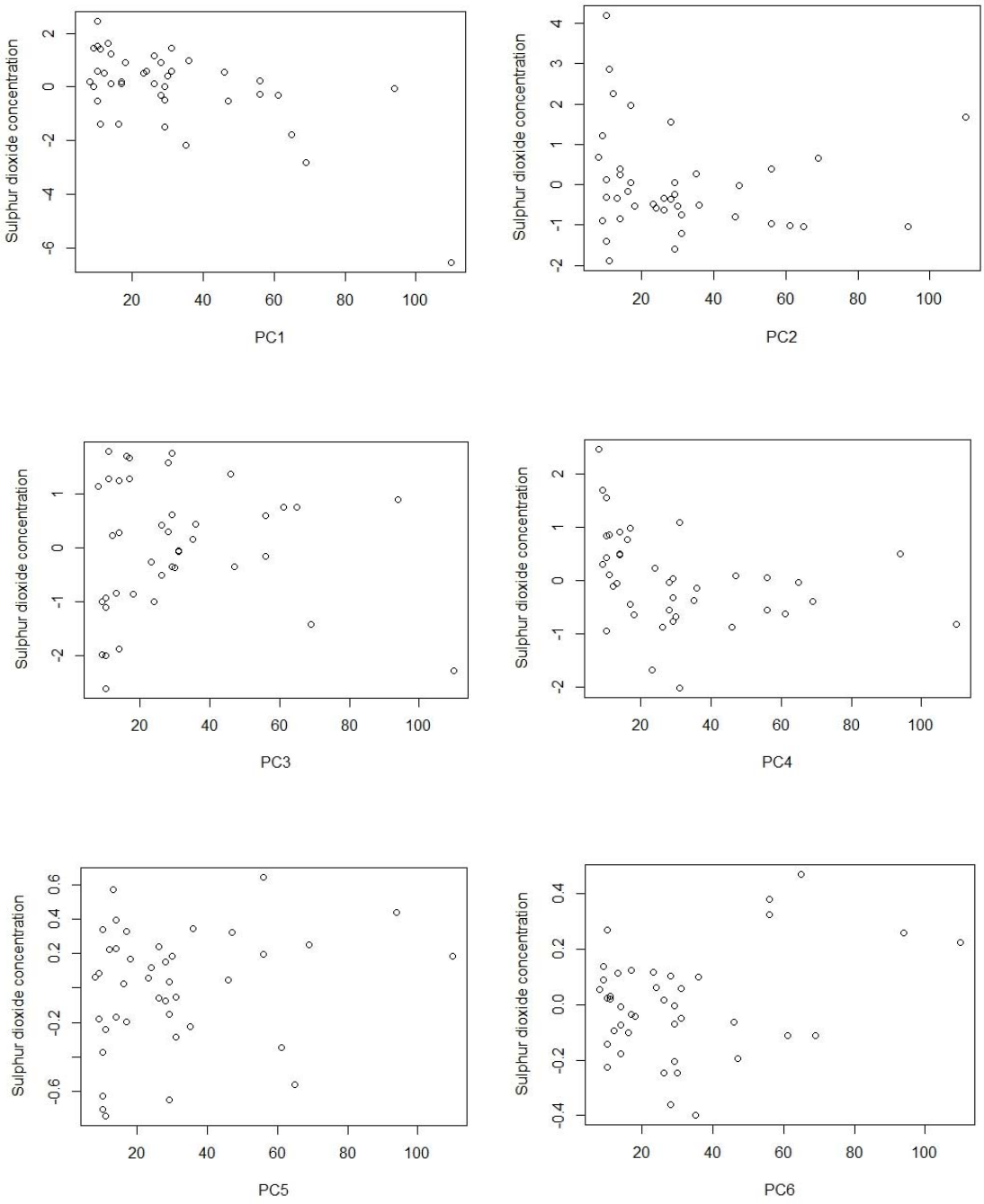


Fig. 3.10. Sulphur dioxide concentration depending on principal components.

The first question we need to ask is "**how many principal components should be used as explanatory variables in the regression?**" The obvious answer to this question is to use the number of principal components that were identified as important in the original analysis; for example, those with eigenvalues greater than one. But this is a case where the obvious answer is not necessarily correct. We will regress the SO2 variables on all six principal components; the necessary R code is given as

```
> usair_reg <- lm(SO2 ~ usair_pca$scores, data = USairpollution)
> summary(usair_reg)
Call:
lm(formula = SO2 ~ usair_pca$scores, data = USairpollution)
Residuals:
   Min     1Q Median     3Q    Max
-23.004  -8.542  -0.991   5.758  48.758
Coefficients:
                          Estimate Std. Error  t value  Pr(>|t|)
(Intercept)                30.049     2.286    13.146   6.9e-15
usair_pca$scoresComp.1     -9.942     1.542    -6.446   2.28e-07
usair_pca$scoresComp.2     -2.240     1.866    -1.200   0.23845
usair_pca$scoresComp.3     -0.375     1.935    -0.194   0.84752
usair_pca$scoresComp.4     -8.549     2.622    -3.261   0.00253
usair_pca$scoresComp.5     15.176     6.753     2.247   0.03122
usair_pca$scoresComp.6     39.271    12.316     3.189   0.00306

Residual standard error: 14.64 on 34 degrees of freedom
Multiple R-squared: 0.6695,     Adjusted R-squared: 0.6112
F-statistic: 11.48 on 6 and 34 DF,  p-value: 5.419e-07
```

Clearly, the first principal component score is the most predictive of sulphur dioxide concentration, but it is also clear that components with small variance do not necessarily have small correlations with the response.

7. The bi[14]plot

The biplot is a procedure that graphically describes both relationships among the q-dimensional observations $x_1, x_2, \dots, x_n$ and relationships among the variables. It is based on the standard result, demonstrated below, that any $n \times m$ matrix $\boldsymbol{B} = (b_{ij})$ of rank r can be factorized (nonuniquely) as

$$\boldsymbol{B} = \boldsymbol{G}\boldsymbol{H}'$$

where $\boldsymbol{G}$ and $\boldsymbol{H}$ are $n \times r$ and $m \times r$ matrices, respectively, of rank r. Thus

$$b_{ij} = \boldsymbol{g}_i'\boldsymbol{h}_j$$

where $\boldsymbol{g}_i'$ and $\boldsymbol{h}_j'$ are the rows of $\boldsymbol{G}$ and $\boldsymbol{H}$, respectively, and we have a representation of the b_{ij} in terms of r-dimensional vectors. This factorization assigns each $\boldsymbol{g}_i$ to a row of $\boldsymbol{B}$ and each $\boldsymbol{h}_j$ to a column of $\boldsymbol{B}$. If $r = 2$, we can plot the $n + m$ vectors in tow dimensions and obtain the biplot. For $r > 2$ it may be possible to approximate $\boldsymbol{B}$ by a rank 2 matrix. It can be seen that

$$\boldsymbol{B} = (\boldsymbol{G}\boldsymbol{\Gamma}')(\boldsymbol{H}\boldsymbol{\Gamma}^{-1})'$$

[14] The "bi" reflects that the technique displays in a single diagram the variances and covariances of the variables and the distances between units.

for any nonsingular $\boldsymbol{\Gamma}$. Note that this orthogonal transformation (rotation or reflection), does not change the relations between vectors. Clearly, we would wish to choose a factorization in which the $\boldsymbol{G}$ and $\boldsymbol{H}$ have meaningful properties. $\boldsymbol{A}$ natural approach is to use the singular value decomposition.

A biplot is a two-dimensional representation of a data matrix obtained from eigenvalues and eigenvectors of the covariance matrix and obtained as

$$X_2 = (p_1, p_2)\begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{pmatrix}\begin{pmatrix}q_1'\\ q_2'\end{pmatrix},$$

where X_2 is the "rank two" approximation of the data matrix X, λ_1 and λ_2 are the first two eigenvalues of the matrix nS, and q_1 and q_2 are the corresponding eigenvectors. The vectors p_1 and p_2 are obtained as $p_i = \frac{1}{\sqrt{\lambda_i}} Xq_i;\ \ i = 1{,}2.$

The biplot is the plot of the n rows of $\sqrt{n}(p_1 + p_2)$ and the q rows of $n^{-\frac{1}{2}}(\sqrt{\lambda_1}q_1, \sqrt{\lambda_2}q_2)$ represented as vectors. The distance between the points representing the units reflects the generalized distance between the units, the length of the vector from the origin to the coordinates representing a particular variable reflects the variance of that variable, and the correlation of two variables is reflected by the angle between the two corresponding vectors for the two variables-the smaller the angle, the greater the correlation. The biplot for the heptathlon data omitting the PNG competitor is shown in Figure 3.11.

```
> biplot(heptathlon_pca, col = c("gray", "black"))
```

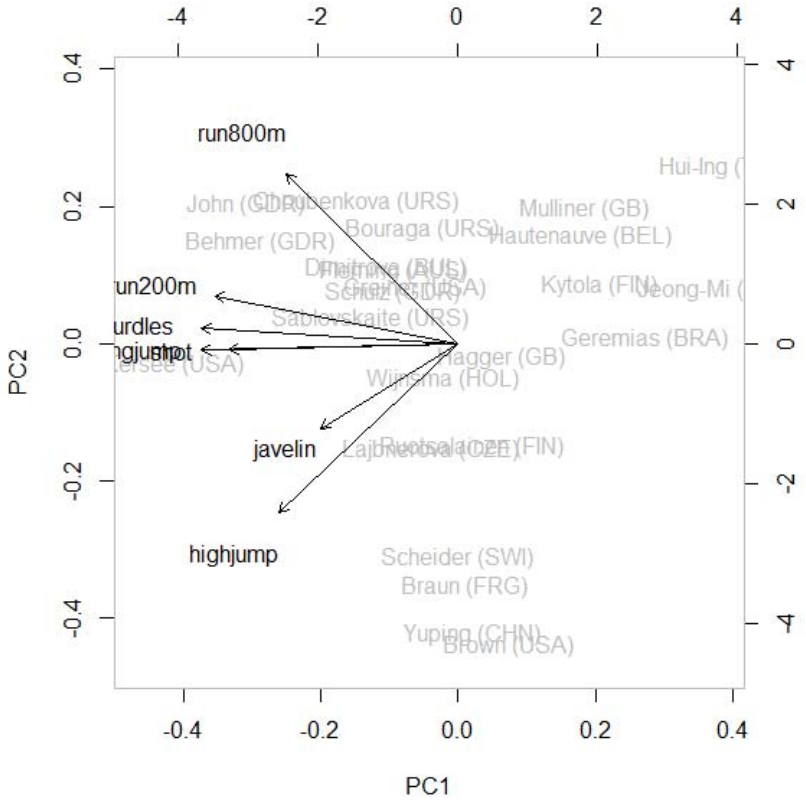


Fig. 3.11. Biplot of the (scaled) first two principal components (with observations for PNG removed).

The plot in Figure 3.11 clearly shows that the winner of the gold medal, Jackie Joyner-Kersee, accumulates the majority of her points from the three events long jump, hurdles, and 200 m. We can also see from the biplot that the results of the 200 m, the hurdles and the long jump are highly correlated, as are the results of the javelin and the high jump; the 800 m time has relatively small correlation with all the other events and is almost uncorrelated with the high jump and javelin results.
The first component largely separates the competitors by their overall score, with the second indicating which are their best events; for example, John, Choubenkova, and Behmer are placed near the end of the vector, representing the 800 m event because this is, relatively speaking, the event in which they give their best performance. Similarly Yuping, Scheider, and Braun can be seen to do well in the high jump.

8. Canonical correlation analysis

Principal components analysis considers interrelationships **within** a set of variables. But there are situations where the researcher may be interested in assessing the relationships **between** two sets of variables. For example, in psychology, an investigator may measure a number of aptitude variables and a number of achievement variables on a sample of students and wish to say something about the relationship between "aptitude" and "achievement". Consider an example in which an agronomist has taken, say, q_1 measurements related to the yield of plants (e.g., height, dry weight, number of leaves) at each of n sites in a region and at the same time may have recorded q_2 variables related to the weather conditions at these sites (e.g., average daily rainfall, humidity, hours of sunshine). The whole investigation thus consists of taking $(q_1 + q_2)$ measurements on n units, and the question of interest is the measurement of the association between "yield" and "weather". One technique for addressing such questions is *canonical correlation analysis* (CCA).

> **PCA** is used to investigate one set of variables with or without covariates. The original variables are replaced by a set of variates called principal components. The components are created to account for maximal variation among the original variables. A generalization of PCA, is **CCA**. The method is developed to investigate relationships between two sets of variables with one or more sets of covariates.

One way to view canonical correlation analysis is as an extension of multiple regression where a single variable (the response) is related to a number of explanatory variables and the regression solution involves finding the linear combination of the explanatory variables that is most highly correlated with the response. In canonical correlation analysis where there is more than a single variable in each of the two sets, the objective is to find the linear functions of the

variables in one set that maximally correlate with linear functions of variables in the other set. Extraction of the coefficients that define the required linear functions has similarities to the process of finding principal components.

The purpose of CCA is to characterize the independent statistical relationships that exist between two sets of variables, $\boldsymbol{x}' = (x_1, x_2, \dots, x_{q_1})$ and $\boldsymbol{y}' = (y_1, y_2, \dots, y_{q_2})$. The overall $(q_1 + q_2) \times (q_1 + q_2)$ correlation matrix contains all the information on associations between pairs of variables in the two sets, but attempting to extract from this matrix some idea of the association between the two sets of variables is not straightforward. This is because the correlations between the two sets may not have a consistent pattern, and these between-set correlations need to be adjusted in some way for the within-set correlations. The question of interest is "how do we quantify the association between the two sets of variables $\boldsymbol{x}$ and $\boldsymbol{y}$?"

The approach adopted in CCA is to take the association between $\boldsymbol{x}$ and $\boldsymbol{y}$ to be the largest correlation between two single variables, u_1 and v_1, derived from $\boldsymbol{x}$ and $\boldsymbol{y}$, with u_1 being a linear combination of $x_1, x_2, \dots, x_{q_1}$ and v_1 being a linear combination of $y_1, y_2, \dots, y_{q_2}$. But often a single pair of variables (u_1, v_1) is not sufficient to quantify the association between the $\boldsymbol{x}$ and $\boldsymbol{y}$ variables, and we may need to consider some or all of s pairs $(u_1, v_1), (u_2, v_2), \dots, (u_s, v_s)$ to do this, where $s = \min(q_1, q_2)$. Each u_i is a linear combination of the variables in $\boldsymbol{x}$, $u_i = \boldsymbol{a}_i'\boldsymbol{x}$, and each v_i is a linear combination of the variables $\boldsymbol{y}$, $v_i = \boldsymbol{b}_i'\boldsymbol{x}$, with the coefficients $(\boldsymbol{a}_i, \boldsymbol{b}_i)$ $(i = 1,2, \dots, s)$ being chosen so that the u_i and v_i satisfy the following:

- The u_i are mutually uncorrelated; i.e., $Cov(u_i, u_j) = 0$ for $i \neq j$.
- The v_i are mutually uncorrelated; i.e., $Cov(v_i, v_j) = 0$ for $i \neq j$.
- The correlation between u_i and v_i is R_i for $i = 1,2, \dots, s$, where $R_1 > R_2 > \cdots > R_s$. The R_i are the *canonical correlations*[15].
- The u_i are uncorrelated with all v_j except v_i; i.e., $Cov(u_i, v_j) = 0$ for $i \neq j$.

The vectors $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$, $i = 1, \dots, s$, which define the required linear combinations of the $\boldsymbol{x}$ and $\boldsymbol{y}$ variables, are found as the eigenvectors of matrices $\boldsymbol{E}_1$ $(q_1 \times q_1)$ (the $\boldsymbol{a}_i$) and $\boldsymbol{E}_2$ $(q_2 \times q_2)$ (the $\boldsymbol{b}_i$), defined as

$$\boldsymbol{E}_1 = \boldsymbol{R}_{11}^{-1}\boldsymbol{R}_{12}\boldsymbol{R}_{22}^{-1}\boldsymbol{R}_{21}, \qquad \boldsymbol{E}_2 = \boldsymbol{R}_{22}^{-1}\boldsymbol{R}_{21}\boldsymbol{R}_{11}^{-1}\boldsymbol{R}_{12},$$

Where $\boldsymbol{R}_{11}$ is the correlation matrix of the variables in $\boldsymbol{x}$, $\boldsymbol{R}_{22}$ is the correlation matrix of the variables in $\boldsymbol{y}$, and $\boldsymbol{R}_{12} = \boldsymbol{R}_{21}$ is the $q_1 \times q_2$ matrix of correlations across the two sets of variables. The canonical correlations $R_1, R_2, \dots, R_s$ are obtained as the square roots of the non-zero eigenvalues of either $\boldsymbol{E}_1$ or $\boldsymbol{E}_2$.

The s canonical correlations $R_1, R_2, \dots, R_s$ express the association between the $\boldsymbol{x}$ and $\boldsymbol{y}$ variables after removal of the within-set correlation.

Inspection of the coefficients of each original variable in each canonical variate can provide an interpretation of the canonical variate in much the same way as interpreting principal components. Such interpretation of the canonical variates may help to describe just how the two sets of original variables are related.

[15] Since the correlation involves the canonical variates u_i and v_i.

Head measurements

Consider the data on head length and head breadth for each of the first two adult sons in 25 families shown in Table 3.1. The question of interest is whether there is a relationship between the head measurements for pairs of sons. We shall address this question by using canonical correlation analysis. Here we shall develop the canonical correlation analysis from first principles as detailed above. Assuming the head measurements data are contained in the data frame headsize, the necessary R code is

```
> headsize <- data.frame(head1, breath1, head2, breath2)
> headsize.std <- sweep[16](headsize, 2, apply(headsize, 2, sd), FUN = "/")
> R <- cor(headsize.std)
> r11 <- R[1:2, 1:2]
> r22 <- R[-(1:2), -(1:2)]
> r12 <- R[1:2, -(1:2)]
> r21 <- R[-(1:2), 1:2]
> (E1 <- solve(r11) %*% r12 %*% solve(r22) %*%r21)

          head1    breath1
head1   0.3225003 0.3168319
breath1 0.3018705 0.3021324


>  (E2 <- solve(r22) %*% r21 %*% solve(r11) %*%r12)

          head2    breath2
head2   0.3013980 0.3002082
breath2 0.3185347 0.3232347


> (e1 <- eigen(E1))
$values
[1] 0.621744734 0.002887956

$vectors
          [,1]       [,2]
[1,] 0.7269968 -0.7040109
[2,] 0.6866408  0.7101892

> (e2 <- eigen(E2))

$values
[1] 0.621744734 0.002887956

$vectors
           [,1]       [,2]
[1,] -0.6837994 -0.7091095
[2,] -0.7296700  0.7050984
```

[16] We use sweep function to standardize head measurements by dividing columns of data matrix by the appropriate standard deviation.

Here the four linear functions are found to be

$$u_1 = +0.73x_1 + 0.69x_2,$$
$$u_2 = -0.70x_1 + 0.71x_2,$$
$$v_1 = -0.68x_3 - 0.73x_4,$$
$$v_2 = -0.71x_3 + 0.71x_4.$$

```
> girth1 <- headsize.std[,1:2] %*% e1$vectors[,1]
> girth2 <- headsize.std[,3:4] %*% e2$vectors[,1]
> shape1 <- headsize.std[,1:2] %*% e1$vectors[,2]
> shape2 <- headsize.std[,3:4] %*% e2$vectors[,2]
> (g <- cor(girth1, girth2))

          [,1]
[1,] -0.7885079

> (s <- cor(shape1, shape2))

          [,1]
[1,] 0.0537397

> plot(girth1, girth2)
> plot(shape1, shape2)
```

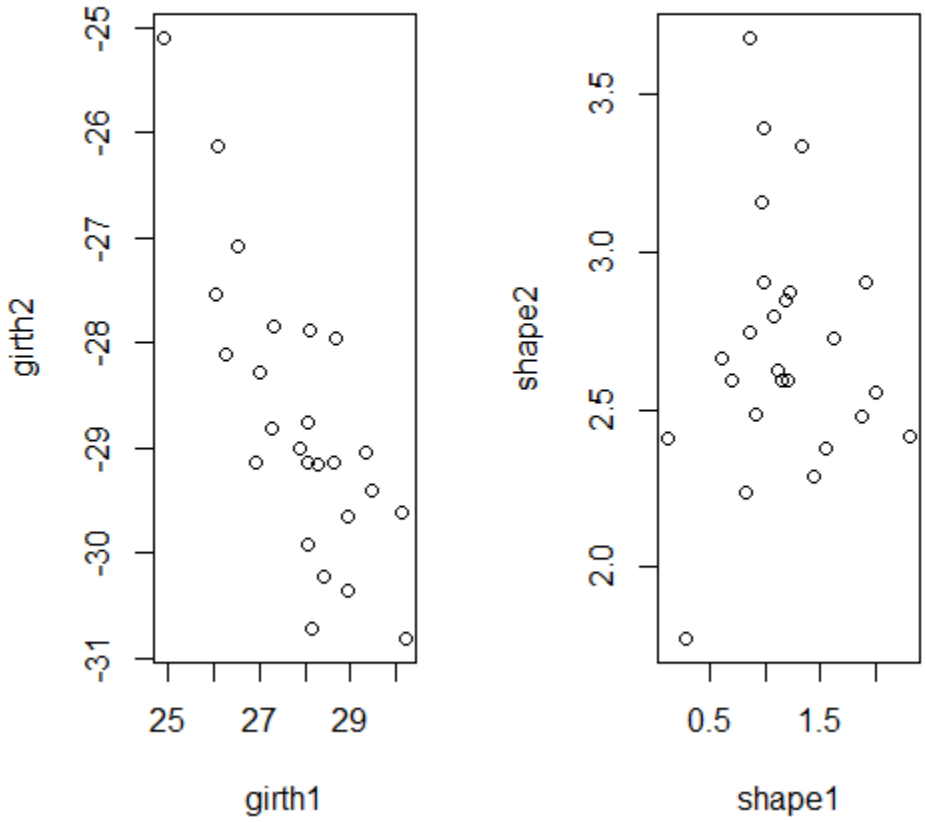


Fig. 3.12. Scatterplots of girth and shape for _rst and second sons.

The first canonical variate for both first and second sons is simply a weighted sum of the two head measurements and might be labeled "girth"; these two variates have a correlation of -0.79. (The negative value arises because of the arbitrariness of the sign in the first coefficient of an eigenvector here both coefficients for girth in first sons are positive and for second sons they are both negative. The correlation can also be found as the square root of the first eigenvalue of $\boldsymbol{E}_1$ (and $\boldsymbol{E}_2$), namely 0.6217.) Each second canonical variate is a weighted difference of the two head measurements and can be interpreted roughly as head "shape"; here the correlation is 0.05 (which can also be found as the square root of the second eigenvalue of $\boldsymbol{E}_1$, namely 0.0029). (Girth and shape are defined to be uncorrelated bioth within and between first and second sons.)
In this example, it is clear that the association between the two head measurements of first and second sons is almost entirely expressed through the "girth" variables, with the two "shape" variables being almost uncorrelated. The association between the two sets of measurements is essentially one-dimensional. A scatterplot of girth for first and second sons and a similar plot for shape reinforce this conclusion. Both plots are shown in Figure 3.12.

Health and personality

The data arise from a study of depression amongst 294 respondents in Los Angeles. The two sets of variables of interest are "health" variables, namely the CESD (the sum of 20 separate numerical scales measuring different aspects of depression) and a measure of general health and "personal" variables, of which there are four: gender, age, income, and educational level (numerically coded from the lowest "less than high school", to the highest, "finished doctorate"). The sample correlation matrix between these variables is given in Table 3.3.

Table 3.3: `LAdepr` data. Los Angeles Depression Data.

CESD	Health	Gender	Age	Edu	Income
1.000	0.212	0.124	-0.164	-0.101	-0.158
0.212	1.000	0.098	0.308	-0.207	-0.183
0.124	0.098	1.000	0.044	-0.106	-0.180
-0.164	0.308	0.044	1.000	-0.208	-0.192
-0.101	-0.207	-0.106	-0.208	1.000	0.492
-0.158	-0.183	-0.180	-0.192	0.492	1.000

Here the maximum number of canonical variate pairs is two, and they can be found using the following R code:

```
> r11 <- LAdepr[1:2, 1:2]
> r22 <- LAdepr[-(1:2), -(1:2)]
> r12 <- LAdepr[1:2, -(1:2)]
> r21 <- LAdepr[-(1:2), 1:2]
> (E1 <- solve(r11) %*% r12 %*% solve(r22) %*%r21)

            CESD      Health
CESD    0.08356175 -0.04312499
Health -0.03256316  0.13338168

> (E2 <- solve(r22) %*% r21 %*% solve(r11) %*%r12)

            Gender         Age         Edu      Income
Gender  0.015477871 -0.001482893 -0.01812515 -0.02224400
Age    -0.002445329  0.147069299 -0.03980907 -0.01599010
Edu    -0.014914914 -0.026019366  0.02544284  0.02508096
Income -0.021163110  0.013027177  0.02158611  0.02895343

> (e1 <- eigen(E1))

$values
[1] 0.15346941 0.06347403

$vectors
           [,1]       [,2]
[1,]  0.5250234 -0.9064831
[2,] -0.8510878 -0.4222421


> (e2 <- eigen(E2))

$values
[1]  1.534694e-01  6.347403e-02 -5.226077e-18  2.915492e-18

$vectors
            [,1]       [,2]       [,3]       [,4]
[1,]  0.002607046  0.4903563  0.1540525 -0.8455776
[2,]  0.980095357 -0.3207531 -0.1127338 -0.1623963
[3,] -0.185801475 -0.4270437 -0.7016704 -0.4697861
[4,]  0.069886386 -0.6886957  0.6864529 -0.1947484
```

(Note that the third and fourth eigenvalues of E2 are essentially zero, as we would expect in this case.) The first canonical correlation is 0.409, calculated as the square root of the first eigenvalue of E1, which is given above as 0.15347.

If tested as outlined in #9, it has an associated p-value that is very small; there is strong evidence that the first canonical correlation is significant. The corresponding variates, in terms of standardized original variables, are

$$u_1 = 0.53\ \text{CESD} - 0.85\ \text{Health}$$

$$v_1 = -0.00\ \text{Gender} - 0.98\ \text{Age} + 0.19\ \text{Education} - 0.07\ \text{Income}$$

Since the higher value of the Gender variable is for females, the interpretation here is that relatively young, poor, and uneducated females are associated with higher depression scores and, to a lesser extent, with poor perceived health.

9. Test of independence

Let $\boldsymbol{X} \sim N_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and partition $\boldsymbol{X}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as

$$\boldsymbol{X}' = (\boldsymbol{X}_1' \; \boldsymbol{X}_2' \; \dots \; \boldsymbol{X}_k'), \boldsymbol{\mu}' = (\boldsymbol{\mu}_1' \; \boldsymbol{\mu}_2' \dots \; \boldsymbol{\mu}_k') \text{ and } \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \dots & \boldsymbol{\Sigma}_{1k} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} & \dots & \boldsymbol{\Sigma}_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{\Sigma}_{k1} & \boldsymbol{\Sigma}_{k2} & \dots & \boldsymbol{\Sigma}_{kk} \end{bmatrix}$$

where $\boldsymbol{X}_i$ and $\boldsymbol{\mu}_i$ are $m_i \times 1$ and $\boldsymbol{\Sigma}_{ii}$ is $q_i \times q_i$ $(i = 1, \dots, k)$, with $\sum_{i=1}^{k} q_i = q$. We wish to test the null hypothesis H that the subvectors $X_1, \dots, X_k$ are independent, i.e.,

$$H: \boldsymbol{\Sigma}_{ij} = 0, \quad i, j = 1, \dots, k; i \neq j$$

against the alternative K that H is not true. Let $\overline{\boldsymbol{X}}$ and $\boldsymbol{S}$ be, respectively, the sample mean vector and covariance matrix formed from a sample of $N = n + 1$ observations on $\boldsymbol{X}$, and let $\boldsymbol{A} = n\boldsymbol{S}$ and partition $\overline{\boldsymbol{X}}$ and $\boldsymbol{A}$ as

$$\overline{\boldsymbol{X}}' = (\overline{\boldsymbol{X}}_1' \; \overline{\boldsymbol{X}}_2' \; \dots \; \overline{\boldsymbol{X}}_k'), \boldsymbol{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \dots & \mathbf{A}_{1k} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \dots & \mathbf{A}_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{A}_{k1} & \mathbf{A}_{k2} & \dots & \mathbf{A}_{kk} \end{bmatrix}$$

where $\overline{\boldsymbol{X}}_i$ and $\boldsymbol{\mu}_i$ are $q_i \times 1$ and $\mathbf{A}_{ii}$ is $q_i \times q_i$.

Theorem 9-1: The likelihood ration test of level α for testing the null hypothesis H of independence is

$$\boldsymbol{\Lambda} = \frac{(\det \boldsymbol{A})^{\frac{N}{2}}}{\prod_{i=1}^{k} (\det \boldsymbol{A}_{ii})^{\frac{N}{2}}}$$

Then H rejects if $\Lambda \leq c_\alpha$ where c_α is chosen so that the significant level of the test is α.

Theorem 9-2: Under the assumptions of Theorem 9-1, for testing $H: \boldsymbol{\Sigma} = diag\ (\boldsymbol{\Sigma}_{11}, \dots, \boldsymbol{\Sigma}_{kk})$, then asymptotically $-2log\ \Lambda \sim \chi_f^2$ where $f = \frac{1}{2}\left(q^2 - \sum_{i=1}^{k} q_i^2\right)$.

Information about the distribution of the likelihood ratio statistic Λ can be obtained from a study of its moments.

Theorem 9-3: When H is true, the hth moment of $W = \Lambda^{\frac{2}{N}}$ is

$$E(W^h) = \frac{\Gamma_q\left(\frac{1}{2}n + h\right)}{\Gamma_q\left(\frac{1}{2}n\right)} \prod\nolimits_{i=1}^{k} \frac{\Gamma_{q_i}\left(\frac{1}{2}n\right)}{\Gamma_{q_i}\left(\frac{1}{2}n + h\right)}$$

where $n = N - 1$.

Suppose that k= 2 and that $q_1 \leq q_2$, $(q_1 + q_2 = q)$. We want to test the independency between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, i.e., testing the null hypothesis $\boldsymbol{\Sigma}_{12} = 0$. Then the test statistic is

$$W = \Lambda^{\frac{2}{N}} = \frac{\det \mathbf{A}}{\det \mathbf{A}_{11} \det \mathbf{A}_{22}} = \prod\nolimits_{i=1}^{q_1} (1 - r_i^2)$$

where $r_1^2, r_2^2, \ldots, r_{q_1}^2$ are the latent roots of $\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}$. Actually those are the squares of the sample canonical correlation coefficients.