8 Principal Component Analysis

8.1 Introduction

We search for a set of *mutually uncorrelated* variables, each one being a *linear combination* (or *linear compound*) of the original set of p variables. One of the motivations for determining such a collection is that, if we derive a set that concentrates the overall variability into the first few variables, it is perhaps easier to see what accounts for the variation in the data. Indeed, if just a few of the new variables seem to account for most of the variation in the data, then it could be argued that the effective dimensionality is less than p - and this could result in a simplified analysis based on a smaller set of variables. If, in addition, the new variables can be interpreted, then the transformation can lead to a greater understanding of the data.

In this course we shall consider the derivation of such variables (or *principal components*) for a given sample as just a method for investigating and summarizing that particular data set, and not as a method for estimating population characteristics. Assume that we have a random sample $\mathbf{x}_1$, $\mathbf{x}_2$, $\ldots$, $\mathbf{x}_n$ of observations on the $p \times 1$ random vector $\mathbf{X}$ and let $\mathbf{S}$ be the corresponding sample covariance matrix. [We shall generally use the unbiased estimate with divisor $(n-1)$.] We shall derive a collection of linear compounds with the aforesaid properties. These new variables take the form $y_i = \mathbf{a}_i'\mathbf{x}$, $i = 1, 2, \ldots, p$, where $\mathbf{x}$ is the general sample vector of observations and y_i is the general sample value of the i'th principal component.

Definition 8.1 (Principal Components)
Suppose that $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$ is the general observation vector in a random sample of size n with sample covariance matrix $\mathbf{S}$. Then the principal components, denoted by $y_1, \ldots, y_p$, satisfy the following conditions:

(I) $y_j = a_{1j}x_1 + a_{2j}x_2 + \ldots + a_{pj}x_p = \mathbf{a}_j'\mathbf{x}$, for $j = 1, \ldots, p$, where $\mathbf{a}_j = (a_{1j}, a_{2j}, \ldots, a_{pj})'$ is a vector of constants satisfying

$$\|\mathbf{a}_j\|^2 = \mathbf{a}_j'\mathbf{a}_j = \sum_{k=1}^{p} a_{kj}^2 = 1;$$

(II) $\mathrm{Var}(y_j) = \mathbf{a}_j'\mathbf{S}\mathbf{a}_j$ is the maximum possible sample variance of any linear combination $\mathbf{a}'\mathbf{x}$ that is uncorrelated (within the sample) with all preceding components y_i, $i = 1, \ldots, j-1$.

8.2 Finding Principal Components

Derivation of y_1

By definition, the first principal component is the variable $y_1 = \mathbf{a}_1'\mathbf{x}$, with $\mathbf{a}_1'\mathbf{a}_1 = 1$, that has the maximum possible sample variance $s_{y_1}^2 = \mathbf{a}_1'\mathbf{S}\mathbf{a}_1$, where $\mathbf{S}$ is the sample covariance matrix for a sample $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$.

This is a standard problem in constrained optimization and may be solved using the method of *Lagrange multipliers*. To use this method, form the *Lagrangian*:

$$L_1(\mathbf{a}) = \mathbf{a}'\mathbf{S}\mathbf{a} - \lambda(\mathbf{a}'\mathbf{a} - 1). \tag{1}$$

The required $\mathbf{a}_1$ is the value of $\mathbf{a}$ that is a stationary point of (1).

Now $L_1(\mathbf{a})$ is a real-valued function of the $p \times 1$ vector $\mathbf{a}$. The derivative of such a function with respect to $\mathbf{a}$ is defined to be the $p \times 1$ vector

$$\frac{\partial L_1}{\partial \mathbf{a}}(\mathbf{a}) \;=\; \left(\frac{\partial L_1}{\partial a_i}\right)$$

and a stationary point of (1) must therefore satisfy:

$$\frac{\partial L_1}{\partial \mathbf{a}}(\mathbf{a}) \;=\; \mathbf{0}.$$

It is easy to show that

$$\frac{\partial}{\partial \mathbf{a}}(\mathbf{a}'\mathbf{S}\mathbf{a}) = 2\mathbf{S}\mathbf{a}$$

and

$$\frac{\partial}{\partial \mathbf{a}}(\mathbf{a}'\mathbf{a}) = 2\mathbf{a}$$

so that

$$\frac{\partial L_1}{\partial \mathbf{a}}(\mathbf{a}) \;=\; 2\mathbf{S}\mathbf{a} \;-\; 2\lambda\mathbf{a}$$

and equating to zero we see that $\mathbf{a}_1$ must satisfy

$$\mathbf{S}\mathbf{a}_1 \;=\; \lambda\mathbf{a}_1 \tag{2}$$

or

$$(\mathbf{S} - \lambda I)\mathbf{a}_1 \;=\; \mathbf{0} \tag{2'}$$

Then, for non-null $\mathbf{a}_1$, we require $|S - \lambda I| = 0$, where $|\cdot|$ is the *determinant* operator, so that λ is an *eigenvalue* of S, with $\mathbf{a}_1$ a corresponding *eigenvector*.

Note

• Since $\mathbf{S}$ is a $p \times p$ symmetric matrix there can be up to p distinct eigenvalues, all of which must be real.

- Since $\mathbf{S}$ is positive (semi-)definite, then all of its eigenvalues are non-negative.

Assume, for the moment, that $\mathbf{S}$ has distinct eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_p \geq 0$ (with corresponding eigenvectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p$), and recall that we are trying to maximize the sample variance of y_1. Premultiplying (2) by $\mathbf{a}_1'$ gives

$$s_{y_1}^2 \;=\; \mathbf{a}_1' S \mathbf{a}_1 \;=\; \mathbf{a}_1' \lambda \mathbf{a}_1 \;=\; \lambda.$$

Hence we choose $\lambda = \lambda_1$ to maximize the sample variance, with $\mathbf{a}_1$ being the corresponding eigenvector of S scaled so that $\mathbf{a}_1'\mathbf{a}_1 = 1$. Then $s_{y_1}^2 = \lambda_1$ with $y_1 = \mathbf{a}_1'\mathbf{x}$.

Derivation of y_2

The second principal component, y_2, takes the form:

$$y_2 = a_{12}x_1 + a_{22}x_2 + \ldots + a_{p2}x_p = \mathbf{a}_2'\mathbf{x}$$

where $\mathbf{a}_2 = (a_{12}, \ldots, a_{p2})'$ and $\mathbf{a}_2'\mathbf{a}_2 = 1$. Since the sample covariance of y_1 and y_2 is required to be zero, we also have

$$S_{y_2 y_1} \;=\; \text{Cov}(\mathbf{a}_2'\mathbf{x}, \mathbf{a}_1'\mathbf{x}) \;=\; \mathbf{a}_2'\mathbf{S}\mathbf{a}_1 = 0. \tag{3}$$

(c.f. Exercises 1, question 1.)

However, since $\mathbf{a}_1$ is the eigenvector corresponding to eigenvalue λ_1, we have

$$\mathbf{S}\mathbf{a}_1 = \lambda_1 \mathbf{a}_1$$

and so, together with (3), this implies

$$\lambda_1 \mathbf{a}_2'\mathbf{a}_1 = 0 \quad \text{or} \quad \mathbf{a}_2'\mathbf{a}_1 = 0. \tag{4}$$

Thus, to find $\mathbf{a}_2$ we need to maximize $\mathbf{a}'\mathbf{S}\mathbf{a}$ with respect to $\mathbf{a}$, subject to the following constraints:

$$\mathbf{a}'\mathbf{a} \;= 1 \;\; \text{normalizing condition}$$
$$\mathbf{a}'\mathbf{a}_1 = 0 \text{ orthogonality condition.}$$

Again, form the Lagrangian:

$$L_2(\mathbf{a}) = \mathbf{a}'\mathbf{S}\mathbf{a} - \lambda(\mathbf{a}'\mathbf{a} - 1) - \gamma(\mathbf{a}'\mathbf{a}_1 - 0).$$

The maximizing $\mathbf{a}_2$ is a stationary point of $L_2(\mathbf{a})$, i.e. that value of $\mathbf{a}$ satisfying

$$\frac{\partial L_2}{\partial \mathbf{a}}(\mathbf{a}) = \mathbf{0}. \tag{5}$$

It is easy to see that

$$\frac{\partial L_2}{\partial \mathbf{a}} = 2\mathbf{S}\mathbf{a}_2 - 2\lambda\mathbf{a}_2 - \gamma\mathbf{a}_1 = \mathbf{0}. \tag{6}$$

Pre-multiplying each term of (6) by $\mathbf{a}_1'$ yields

$$2\mathbf{a}_1'\mathbf{S}\mathbf{a}_2 - 2\lambda\mathbf{a}_1'\mathbf{a}_2 - \gamma\mathbf{a}_1'\mathbf{a}_1 = 0$$

and so, upon invoking (3), the orthogonality condition, and $\mathbf{a}_1'\mathbf{a}_1 = 1$, we have

$$\gamma = 0$$

and hence (6) reduces to

$$(\mathbf{S} - \lambda I)\mathbf{a}_2 = \mathbf{0}.$$

such that

$$\mathbf{a}_2'\mathbf{a}_2 = 1, \text{ and } \mathbf{a}_2'\mathbf{a}_1 = 0$$

This implies that $|\mathbf{S} - \lambda I| = 0$ so that again λ is an eigenvalue of S, with corresponding eigenvector $\mathbf{a}_2$.

Clearly, $\lambda = \lambda_2$, where $\mathbf{a}_2$ is a corresponding eigenvector, so that

$$s_{y_2}^2 = \mathbf{a}_2'\mathbf{S}\mathbf{a}_2 = \lambda_2, \quad \mathbf{a}_2'\mathbf{a}_2 = 1 \text{ and } \mathbf{a}_2'\mathbf{a}_1 = 0$$

In general, the p principal components of the sample are given by $y_i = \mathbf{a}_i'\mathbf{x}$, $i = 1, \ldots, p$, where $s_{y_i}^2 = \lambda_i$, the i'th largest eigenvalue of $\mathbf{S}$ and $\mathbf{a}_i$ is the corresponding, normalized eigenvector.

Remarks 8.2 (Eigenvalues not distinct)
In the case where some of the eigenvalues are not distinct, then there is no unique way of choosing the corresponding eigenvectors; however, it can be shown that as long as these eigenvectors are chosen to be mutually orthogonal, then the (suitably modified) procedure carries through.

8.3 Some Further Results and R Computations

Let $\mathbf{A}$ be a $p \times p$ matrix whose columns are made up from the $\{\mathbf{a}_i\}$, as constructed in the previous section, i.e.

$$\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p].$$

Then, clearly, the matrix $\mathbf{A}$ is orthogonal so that $\mathbf{A}'\mathbf{A} = \mathbf{A}\mathbf{A}' = \mathbf{I}$. The general $p \times 1$ vector $\mathbf{y}$ of principal components corresponding to the general vector of observations, $\mathbf{x}$, may be calculated using the formula

$$\mathbf{y} = \mathbf{A}'\mathbf{x}.$$

Let $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$. By construction

$$\mathbf{S_y} \;=\; \text{sample covariance matrix of } \mathbf{y} \;=\; \mathbf{\Lambda}.$$

On the other hand,

$$\mathbf{S_y} \;=\; \text{sample covariance matrix of } \mathbf{A}'\mathbf{x} = \mathbf{A}'\mathbf{S}\mathbf{A},$$

so we have the identity

$$\mathbf{A}'\mathbf{S}\mathbf{A} \;=\; \mathbf{\Lambda}. \tag{7}$$

Pre- and post-multiplying by $\mathbf{A}$ and $\mathbf{A}'$ respectively yields

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

or

$$\mathbf{S} \;=\; \mathbf{A}\mathbf{\Lambda}\mathbf{A}' \;=\; \sum_{i=1}^{p} \lambda_i \mathbf{a}_i \mathbf{a}_i^{'} \tag{8}$$

Proposition 8.3 (Identity for the sum of variances)
Suppose that $\mathbf{x}$ is the general $p \times 1$ observation vector (with sample covariance matrix $\mathbf{S}$), and $\mathbf{y}$ is the corresponding vector of principal components. Then

$$\text{trace}(\mathbf{S}) \;=\; \sum_{i=1}^{p} s_{x_i}^2 \;=\; \sum_{i=1}^{p} s_{y_i}^2 \;=\; \sum_{i=1}^{p} \lambda_i \;=\; \text{trace}(\mathbf{S_y}). \tag{9}$$

Proof
Since $\mathbf{\Lambda}$ is the sample covariance matrix of $\mathbf{y}$, then

$$\sum_{i=1}^{p} s_{y_i}^2 = \sum_{i=1}^{p} \lambda_i = \text{trace}(\mathbf{\Lambda}).$$

On the other hand

$$\begin{aligned}&\text{trace}(\mathbf{\Lambda}) = \text{trace}(\mathbf{A}'\mathbf{S}\mathbf{A}) = \text{trace}((\mathbf{A}'\mathbf{S})\mathbf{A}) = \text{trace}(\mathbf{A}(\mathbf{A}'\mathbf{S}))\\ &= \text{trace}((\mathbf{A}\mathbf{A}')\mathbf{S}) = \text{trace}(\mathbf{I}\mathbf{S}) = \text{trace}(\mathbf{S}).\end{aligned}$$

However,

$$\text{trace}(\mathbf{S}) = \sum_{i=1}^{p} s_{x_i}^2.$$

□

Remarks 8.4
(i) In the light of Proposition 8.3, if we regard the total variation in the data to be given

by $\sum_{i=1}^{p} \lambda_i$, then the proportion of this quantity that can be attributed to the i-th principal component is

$$\lambda_i \Big/ \sum_{j=1}^{p} \lambda_j \qquad i=1,\ldots,p$$

and, to the first m principal components,

$$\sum_{i=1}^{m} \lambda_i \Big/ \sum_{j=1}^{p} \lambda_j \qquad m \leq p.$$

(ii) The principal components of $x_1, \ldots, x_p$ are *not* independent of the scale on which each variable is measured. [Unlike regression models, where if the variables are rescaled then the coefficients are correspondingly rescaled so as to give equivalent predictions.] One could decide to derive principal components on the basis of the correlation matrix $\mathbf{R}$, so that variables have been scaled to have equal variances. This makes all variables equally important (in terms of our criterion of size of variance) and yields a different set of principal components to those obtained from $\mathbf{S}$. The use of $\mathbf{R}$ may be desirable in some cases, particularly when the components of $\mathbf{x}$ have variances of quite different orders of magnitude. The choice of S vs R will be considered later. [See also Exercises 1, question 2.]

(iii) The derivations given in this chapter are for the *sample* only. However it is possible to regard the PCs calculated from the sample covariance matrix $\mathbf{S}$ as estimates of the PCs in the population. Thus, the eigenvalues and eigenvectors of $\mathbf{S}$ could be regarded as estimates of the eigenvalues and eigenvectors of $\mathbf{\Sigma}$.

Consider a very small example. Some market researchers were trying to assess income and spending patterns of households within a particular region. As a trial run, the researchers randomly selected a number of households and managed to ascertain how much was spent on groceries(X_1), the amount spent on leisure activities(X_2), and also the total (combined) income(X_3), in a particular month. The data are summarized below:

Groceries(£)	Leisure(£)	Income(£)
227.01	96.98	741.29
241.42	140.44	854.07
188.08	85.13	812.07
238.23	158.22	813.69
235.86	103.06	731.42

We consider whether there is any scope for reducing the dimensionality of the data by constructing a new set of variables accounting for most of the variation. Principal components can be found in R as follows:

```
> groceries <- c(227.01, 241.42, 188.08, 238.23, 235.86)
> leisure <- c(96.98, 140.44, 85.13, 158.22, 103.06)
> income <- c(741.29, 854.07, 812.07, 813.69, 731.42)
> spend <- data.frame(groceries, leisure, income)

> m.spend <- apply(spend, 2, mean)
> m.spend
groceries   leisure    income
  226.120   116.766   790.508

> cov.spend <- var(spend)
> cov.spend
          groceries  leisure     income
groceries 480.86085 479.1369  -46.57675
leisure   479.13690 964.7673  891.82637
income    -46.57675 891.8264 2739.06402

> var.spend <- apply(spend, 2, var)
> var.spend
groceries   leisure    income
 480.8608  964.7673 2739.0640

> cor.spend <- cor(spend)
> cor.spend
            groceries   leisure      income
groceries  1.00000000 0.7034587 -0.04058433
leisure    0.70345873 1.0000000  0.54861532
income    -0.04058433 0.5486153  1.00000000

> spend.eig.cov <- eigen(var(spend))
> spend.eig.cov$values
[1] 3117.62546  997.08606   69.98063

> spend.eig.cov$vectors
           [,1]       [,2]       [,3]
[1,] 0.05511988 -0.6581783  0.7508416
[2,] 0.39257810 -0.6771368 -0.6223891
[3,] 0.91806549  0.3290700  0.2210627

> spend.eig.cor <- eigen(cor(spend))
> spend.eig.cor$values
[1] 1.87268832 1.03935677 0.08795491

> spend.eig.cor$vectors
           [,1]        [,2]       [,3]
[1,] -0.5565053 -0.61454914  0.5591344
[2,] -0.7148146  0.01112539 -0.6992255
[3,] -0.4234878  0.78880009  0.4454801
```

On the basis of the *sample covariance matrix*, we find that the principal components are

$$\begin{aligned}
y_1 &= \ \ 0.05511988x_1 + 0.39257810x_2 + 0.91806549x_3\\
y_2 &= -0.6581783x_1 - 0.6771368x_2 + 0.3290700x_3\\
y_3 &= \ \ 0.7508416x_1 - 0.6223891x_2 + 0.2210627x_3
\end{aligned}$$

with corresponding sample variances 3117.62546, 997.08606, and 69.98063, respectively. Thus the total variance is 4184.69215, and the principal components account successively for 74.5%, 23.8% and 1.7% of the total variance.

Similarly, the principal components, based on the *sample correlation matrix*, are given by

$$\begin{aligned}
\widetilde{y}_1 &= \ \ 0.5565053\widetilde{x}_1 + 0.7148146\widetilde{x}_2 + 0.4234878\widetilde{x}_3\\
\widetilde{y}_2 &= -0.61454914\widetilde{x}_1 + 0.01112539\widetilde{x}_2 + 0.78880009\widetilde{x}_3\\
\widetilde{y}_3 &= -0.5591344\widetilde{x}_1 + 0.6992255\widetilde{x}_2 - 0.4454801\widetilde{x}_3
\end{aligned}$$

where $\widetilde{x}_i = x_i/S_i$ (i.e. x_i/s_{x_i}) for $i = 1, 2, 3$. The sample variances of the new principal components, $\widetilde{y}_1, \widetilde{y}_2, \widetilde{y}_3$, are 1.87268832, 1.03935677, and 0.08795491, respectively.

In this case the total variance is 3, and the principal components account successively for 62.4%, 34.7% and 2.9% of the total variance.

These results can also be obtained using the command `princomp`.

```
> spend.pca.cov <- princomp(spend)
> summary(spend.pca.cov)
Importance of components:
                           Comp.1     Comp.2   Comp.3
Standard deviation     49.9409689 28.2430319 7.482279
Proportion of Variance  0.7450071  0.2382699 0.016723
Cumulative Proportion   0.7450071  0.9832770 1.000000
> loadings(spend.pca.cov)
Loadings:
          Comp.1 Comp.2 Comp.3
groceries        -0.658  0.751
leisure    0.393 -0.677 -0.622
income     0.918  0.329  0.221

               Comp.1 Comp.2 Comp.3
SS loadings     1.000  1.000  1.000
Proportion Var  0.333  0.333  0.333
Cumulative Var  0.333  0.667  1.000

> spend.pca.cor <- princomp(spend, cor = T)
> summary(spend.pca.cor)
Importance of components:
                          Comp.1    Comp.2    Comp.3
```

```
Standard deviation     1.3684620 1.0194885 0.2965719
Proportion of Variance 0.6242294 0.3464523 0.0293183
Cumulative Proportion  0.6242294 0.9706817 1.0000000

> loadings(spend.pca.cor)
Loadings:
          Comp.1 Comp.2 Comp.3
groceries -0.557 -0.615  0.559
leisure   -0.715        -0.699
income    -0.423  0.789  0.445

               Comp.1 Comp.2 Comp.3
SS loadings     1.000  1.000  1.000
Proportion Var  0.333  0.333  0.333
Cumulative Var  0.333  0.667  1.000
```

The reporting of the $\{\mathbf{a}_i\}$ using `princomp` is less precise; moreover, very small coefficients are 'suppressed'. [1]

We see that the 1st two PCs computed on the basis of the sample covariance matrix account for just over 98% of the variation; similarly the 1st two PCs derived from the sample correlation matrix account for just over 97%.
Aside Using the `eigen` function, we found that the eigenvalues of the sample covariance matrix were 3117.62546, 997.08606, and 69.98063. The square root of these values, i.e. the standard deviations of the PCs, are 55.835701, 31.576670, and 8.365443 respectively. However, from the `princomp` function, we find that the standard deviations are 49.9409689, 28.2430319, and 7.482279. Why the discrepancy?

For some reason `princomp` analyses the sample covariance matrix calculated using divisor n rather than $(n-1)$. On the other hand, the `var` function uses $(n-1)$. Thus if we scale the eigenvalues in our own calculation by $(n-1)/n$ before taking the square root, we obtain the results from `princomp`.

This is demonstrated by the following code and output:

```
> n <- 5
> sqrt(((n - 1)/n) * spend.eig.cov$values)
[1] 49.940969 28.243032  7.482279
```

8.4 Principal Components Uses

1. Interpretation
What we do with components very often depends on their interpretation, and in some contexts it can be the seemingly unimportant components that we are interested in. We try to interpret

[1] Unless otherwise stated, you should feel free to use either the `eigen` or `princomp` method for carrying out PCA.

components by looking at the coefficients associated with the original variables. However we need to be careful, since the contribution of a variable to a component depends on the values taken by the variable as well as the value of the coefficient. The coefficients are not necessarily comparable.

2. Reduction in dimensionality

If an eigenvalue is zero, then there is an exact linear relationship between the variables. One variable is therefore redundant. (This can happen with percentage data if all variables are included. e.g. geological data consisting of percentages of each component in the sample.) If many eigenvalues are 'small', or equivalently if the first few components account for most of the variation, then the effective dimensionality of the data may be $k << p$. Such a reduction in dimensionality has a variety of uses.

For example,

(i) The first k (uncorrelated) components could be used in a subsequent analysis (e.g. multiple regression or discriminant analysis) instead of the original p correlated variables. This would give more stable estimates with little loss of information.

(ii) The n data units can be displayed graphically using fewer dimensions. For example, they can be plotted in 2 dimensions using the first two components. This gives the best 2-dimensional representation in the least-squares sense and may reveal groupings or patterns of interest. [Use $Y = XA$ to get a new data matrix of component scores, and use the first k columns.]

Example 8.5

Data are measurements on shells of painted turtles: x_1 =length, x_2 =width and x_3 =height in mm. The data have the following sample covariance matrix

$$\mathbf{S} = \begin{array}{cc} \begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array} & \begin{pmatrix} 451.4 & & \\ 271.2 & 171.7 & \\ 168.7 & 271.2 & 66.7 \end{pmatrix} \end{array}$$

which shows that Total Variance = trace($\mathbf{S}$) = $451.4 + 171.7 + 66.7 = 689.8$.

	Principal Components		
Variable	y_1	y_2	y_3
x_1	0.81	-0.55	-0.21
x_2	0.50	0.83	-0.25
x_3	0.30	0.10	0.95
Variance	680.4	6.5	2.86
%Total Var.	98.64%	0.94%	0.41%

In this example the interpretation is critical, and the use made of the principal components is a function of this.

PC1: this accounts for 98.6% of the total variance is a measure of 'size'. We could say that this component is dominant and forget the others. However 'size' could just be a nuisance factor, perhaps representing 'stage of growth', or reflecting environmental factors. We might wish to omit this component and consider only shape differences which could indicate whether there are different species present.

PC2: This is a measure of 'shape' comparing *length* with *width.*

PC3: This is also a measure of 'shape'comparing *height* with the average of *length* and *width.*

Note that in a 'shape' component, it is the *relative* sizes that are important. e.g.

$$\begin{aligned} y_2 = -0.55x_1 + 0.83x_2 = 0 &\Rightarrow \frac{\text{length}}{\text{width}} = \frac{0.83}{0.55} \approx 1.5 \\ > 0 &\Rightarrow \frac{\text{length}}{\text{width}} < 1.5 \\ < 0 &\Rightarrow \frac{\text{length}}{\text{width}} > 1.5 \end{aligned}$$

Methods for choosing components to be retained

When carrying out a principal component analysis, we need to decide how many components to retain to represent the data; the other components can then be discarded. We shall consider a variety of methods. Note first, however, that before discarding components we may wish to consider whether their interpretation is of interest, in which case we may wish to retain or discard on the basis of that interpretation.

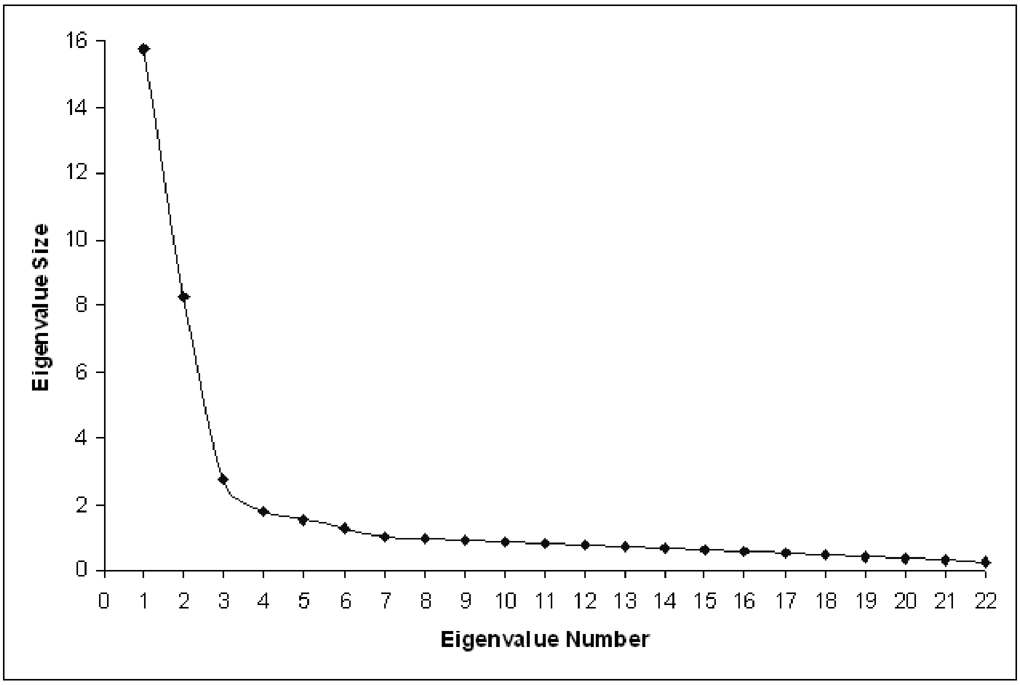
Suppose now that we are not looking to interpret the components, but just to reduce the dimensionality. If we assume multivariate normality then we may wish to use large-sample methods for the sampling distributions of the eigenvalues. It is then possible to test, for any subset of components $1, \ldots, k$, whether the remaining eigenvalues $\lambda_{k+1}, \ldots, \lambda_p$ are all equal. If this hypothesis is accepted then there is no point in extracting further components.

If we make no distributional assumptions then there is no *objective* way of deciding how many to take. If we are using $\mathbf{R}$, then one 'rule of thumb' is to retain components so long as the eigenvalues remain ≥ 1. (If no direction is more important than any other, i.e. if there are no correlations, then all the eigenvalues are 1).

Percentage of Variance accounted for: The number of components retained is often based on the percentage of the variance accounted for. We might like to achieve a relatively high percentage, say 70%-90%. Note, however, that if the reduction in dimensionality is for the purpose of input into some other analysis such as regression or MANOVA, it may be advisable to keep more components than if we are merely aiming for descriptive simplification.

Average Eigenvalue: We could retain those components whose eigenvalues are greater than the average eigenvalue $\bar{\lambda} = \frac{1}{p}\sum_{j=1}^{p}\lambda_j$, which is also the average variance of the variables since $\sum_{j=1}^{p}\lambda_j = \text{trace}(\mathbf{S})$. [For a correlation matrix $\bar{\lambda} = 1$.] The average eigenvalue method often works well in practice. When this method errs, it is likely to be on the side of retaining too many components. The method is fairly accurate when the number of variables is ≤ 30 and the variables are rather highly correlated. For larger numbers of variables that are not as highly correlated, the technique tends to overestimate the number of components.

Scree Graph: We could plot the eigenvalues in an attempt to find a visual break (a gap or an 'elbow') between the 'large' eigenvalues and the 'small' eigenvalues. This plot is called a *scree graph.* (The term 'scree' refers to the geographical term for the debris at the bottom of a rocky cliff.



An ideal scree graph is shown above in which it is easy to distinguish the large eigenvalues from the small ones. The first three eigenvalues form a steep curve, the remaining eigenvalues exhibit a linear trend with small slope. In such a case, it is clear that we should delete the components corresponding to the small eigenvalues on the straight line. In practice, this ideal pattern may not appear and this approach may not be conclusive.

8.5 Choice of R versus S

The principal components of **S** are preferred to those of **R** for many purposes. However we may wish to use **R** in situations in which the variances differ widely. In such cases , the principal components from **S** may be close to the original variables, and the principal components from **R** may be more useful and interpretable since all variables contribute more evenly.

We look now at some of the properties of principal components extracted from **R**.

1. Principal components from **R** differ from those obtained from **S** because of the lack of scale invariance of components of **S**. If we express the components from **R** in terms of the original variables they will not be the same as the components from **S**, and the

corresponding coefficient vectors will no longer be orthogonal. The percentage of variance accounted for by the components of **R** will not agree with the analogous percentages for **S**. For **R** the total variance, and hence the sum of the eigenvalues is p, and all variables are constrained to contribute the same amount to the total variance.

2. Principal components extracted from **R** are scale invariant, since a change in the scale of the variables does not affect the r_{ij}'s in **R**.

3. The components of **R** do not change if we multiply the off-diagonal elements of **R** by any non-zero constant k such that $-1 \leq kr_{ij} \leq 1$ for $i \neq j$, as in

$$\mathbf{R}_k = \begin{pmatrix} 1 & kr_{12} & \dots & kr_{1p} \\ kr_{21} & 1 & \dots & kr_{2p} \\ \vdots & & & \\ kr_{p1} & kr_{p2} & \dots & 1 \end{pmatrix}.$$

Only the *relative* sizes of the correlations are relevant in determining the eigenvectors of the correlation matrix.

To illustrate this, consider the following two correlation matrices

$$\mathbf{R}_1 = \begin{pmatrix} 1 & 0.9 & 0.5 \\ 0.9 & 1 & 0.1 \\ 0.5 & 0.1 & 1 \end{pmatrix}, \quad \mathbf{R}_2 = \begin{pmatrix} 1 & 0.0009 & 0.0005 \\ 0.0009 & 1 & 0.0001 \\ 0.0005 & 0.0001 & 1 \end{pmatrix}$$

The components are the same for both $\mathbf{R}_1$ and $\mathbf{R}_2$:

$$\begin{aligned} \widetilde{y}_1 &= 0.69\widetilde{x}_1 + 0.61\widetilde{x}_2 + 0.38\widetilde{x}_3 \\ \widetilde{y}_2 &= -0.05\widetilde{x}_1 - 0.48\widetilde{x}_2 + 0.88\widetilde{x}_3 \\ \widetilde{y}_3 &= -0.72\widetilde{x}_1 + 0.62\widetilde{x}_2 + 0.30\widetilde{x}_3 \end{aligned}$$

where $\widetilde{x}_j$ is the standardised variable $\widetilde{x}_j = x_j/S_j$. The eigenvalues are as follows

$$\mathbf{R}_1 : 2.074, 0.915, 0.011,$$

$$\mathbf{R}_2 : 1.001, 0.9997, 0.9990$$

Then, $\widetilde{y}_1$ accounts for 69.1% of the variance in $\mathbf{R}_1$ and for 33.4% in $\mathbf{R}_2$, while $\widetilde{y}_3$ accounts for 0.351% of the variance in $\mathbf{R}_1$ and 33.3% in $\mathbf{R}_2$. The first two components of $\mathbf{R}_1$ account for 99.6% of the variance, but the first two components of $\mathbf{R}_2$ account for only 66.7% of the variance. Since the three eigenvalues of $\mathbf{R}_2$ are essentially equal, none of the three components from $\mathbf{R}_2$ can be neglected.

4. One advantage of using the correlation matrix $\mathbf{R}$ might be to make *interpretation* easier. Suppose we set

$$\mathbf{C} = (\sqrt{\lambda_1}\mathbf{a}_1, \dots \sqrt{\lambda_p}\mathbf{a}_p) = (\mathbf{c}_1, \dots, \mathbf{c}_p).$$

   This re-scales the eigenvectors $\mathbf{a}_j$, to give corresponding eigenvectors $\mathbf{c}_j$, scaled so that $\mathbf{c}_j'\mathbf{c}_j = \lambda_j$. It can be shown that the resulting components c_{ij} of $\mathbf{c}_j$ are just the sample correlations between the $\widetilde{x}_i$ and the component $\widetilde{y}_j$. The c_{ij} are all *comparable*, and to interpret $\widetilde{y}_j$, we just look at the sizes of the c_{ij}, $i = 1, \dots, p$, interpreting these as correlations.

   Proof
   Let,

$$\mathbf{C} = (\sqrt{\lambda_1}\mathbf{a}_1, \dots \sqrt{\lambda_p}\mathbf{a}_p) = (\mathbf{c}_1, \dots, \mathbf{c}_p)$$

   Then, $\mathbf{C} = \mathbf{A}\ \mathrm{diag}(\sqrt{\lambda_i})$.

   Now, $\mathbf{y} = A^T\mathbf{x} \Rightarrow \mathbf{x} = A\mathbf{y}$ so that $x_i = \sum_{k=1}^p a_{ik}y_k$, (c/f $y_i = \sum_{k=1}^p a_{ki}x_k$).

$$\begin{aligned}\mathrm{Cov}(x_i, y_j) &= \mathrm{Cov}(\sum_{k=1}^p a_{ik}y_k, y_j)\\ &= a_{ij}\mathrm{Var}(y_j)\\ &= a_{ij}\lambda_j\end{aligned}$$

$$\Rightarrow \mathrm{Correlation}(x_i, y_j) = \frac{a_{ij}\lambda_j}{1\sqrt{\lambda_j}} = \sqrt{\lambda_j}a_{ij} = c_{ij}$$

□

5. A common situation is for the matrix $\mathbf{R}$ to consist of all positive elements. Since using $\mathbf{R}$ constrains all variables to have equal variances, it is not surprising that in this case the first component generally gives almost equal weights to the (scaled) variables. Even fairly dissimilar $\mathbf{R}$ matrices are likely to have very similar first components.

Clearly the choice of $\mathbf{S}$ versus $\mathbf{R}$ is not straightforward and should depend on the particular application involved. The above properties of principal components derived from $\mathbf{R}$ indicate the constraints which this imposes. However any sample covariance matrix $\mathbf{S}$ with diagonal elements of very different orders of magnitude is unlikely to give useful components. Changing the units of measurement can sometimes help, but the best choice might be to use $\mathbf{R}$.