1
Introduction

The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first *few* retain most of the variation present in *all* of the original variables.

The present introductory chapter is in two parts. In the first, PCA is defined, and what has become the standard derivation of PCs, in terms of eigenvectors of a covariance matrix, is presented. The second part gives a brief historical review of the development of PCA.

1.1 Definition and Derivation of Principal Components

Suppose that $\mathbf{x}$ is a vector of p random variables, and that the variances of the p random variables and the structure of the covariances or correlations between the p variables are of interest. Unless p is small, or the structure is very simple, it will often not be very helpful to simply look at the p variances and all of the $\frac{1}{2}p(p-1)$ correlations or covariances. An alternative approach is to look for a few ($\ll p$) derived variables that preserve most of the information given by these variances and correlations or covariances.

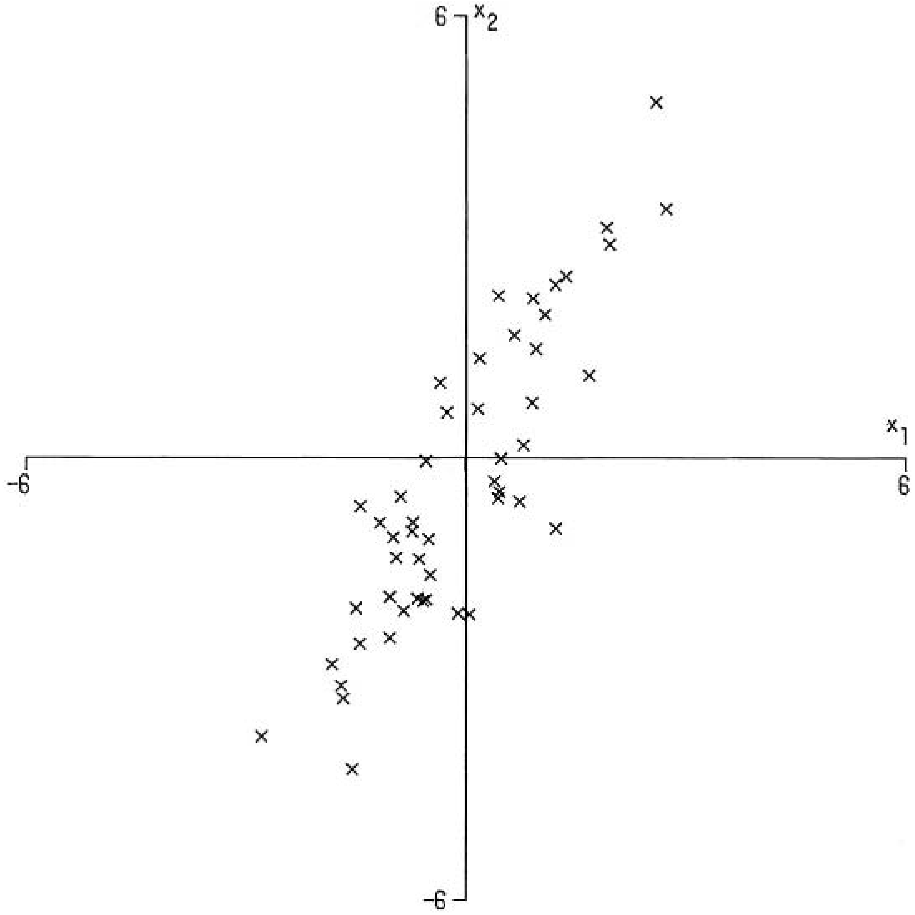

Figure 1.1. Plot of 50 observations on two variables x_1,x_2.

Although PCA does not ignore covariances and correlations, it concentrates on variances. The first step is to look for a linear function $\boldsymbol{\alpha}_1'\mathbf{x}$ of the elements of $\mathbf{x}$ having maximum variance, where $\boldsymbol{\alpha}_1$ is a vector of p constants α_{11}, $\alpha_{12}, \ldots, \alpha_{1p}$, and $'$ denotes transpose, so that

$$\boldsymbol{\alpha}_1'\mathbf{x} = \alpha_{11}x_1 + \alpha_{12}x_2 + \cdots + \alpha_{1p}x_p = \sum_{j=1}^{p} \alpha_{1j}x_j.$$

Next, look for a linear function $\boldsymbol{\alpha}_2'\mathbf{x}$, uncorrelated with $\boldsymbol{\alpha}_1'\mathbf{x}$ having maximum variance, and so on, so that at the kth stage a linear function $\boldsymbol{\alpha}_k'\mathbf{x}$ is found that has maximum variance subject to being uncorrelated with $\boldsymbol{\alpha}_1'\mathbf{x}$, $\boldsymbol{\alpha}_2'\mathbf{x}, \ldots, \boldsymbol{\alpha}_{k-1}'\mathbf{x}$. The kth derived variable, $\boldsymbol{\alpha}_k'\mathbf{x}$ is the kth PC. Up to p PCs could be found, but it is hoped, in general, that most of the variation in $\mathbf{x}$ will be accounted for by m PCs, where $m \ll p$. The reduction in complexity achieved by transforming the original variables to PCs will be demonstrated in many examples later in the book, but it will be useful here to consider first the unrealistic, but simple, case where $p = 2$. The advantage of $p = 2$ is, of course, that the data can be plotted exactly in two dimensions.

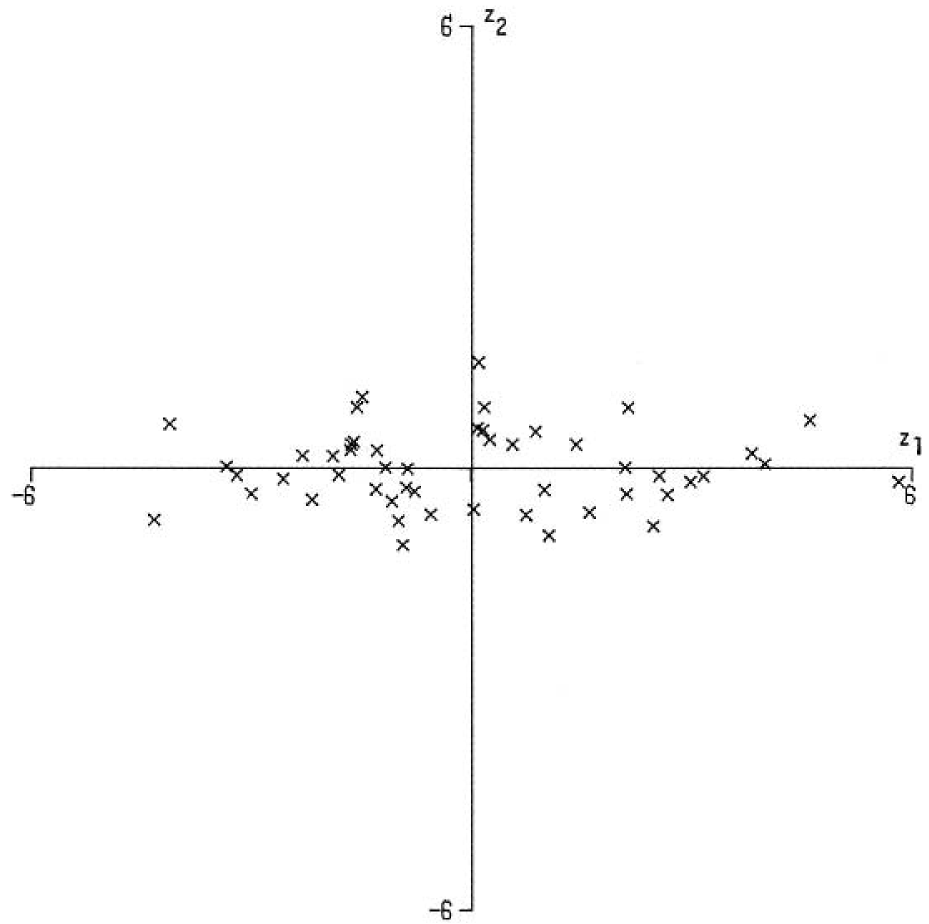

Figure 1.2. Plot of the 50 observations from Figure 1.1 with respect to their PCs z_1, z_2.

Figure 1.1 gives a plot of 50 observations on two highly correlated variables x_1, x_2 . There is considerable variation in both variables, though rather more in the direction of x_2 than x_1. If we transform to PCs z_1, z_2, we obtain the plot given in Figure 1.2.

It is clear that there is greater variation in the direction of z_1 than in either of the original variables, but very little variation in the direction of z_2. More generally, if a set of p (> 2) variables has substantial correlations among them, then the first few PCs will account for most of the variation in the original variables. Conversely, the last few PCs identify directions in which there is very little variation; that is, they identify near-constant linear relationships among the original variables.

As a taster of the many examples to come later in the book, Figure 1.3 provides a plot of the values of the first two principal components in a 7-variable example. The data presented here consist of seven anatomical measurements on 28 students, 11 women and 17 men. This data set and similar ones for other groups of students are discussed in more detail in Sections 4.1 and 5.1. The important thing to note here is that the first two PCs account for 80 percent of the total variation in the data set, so that the

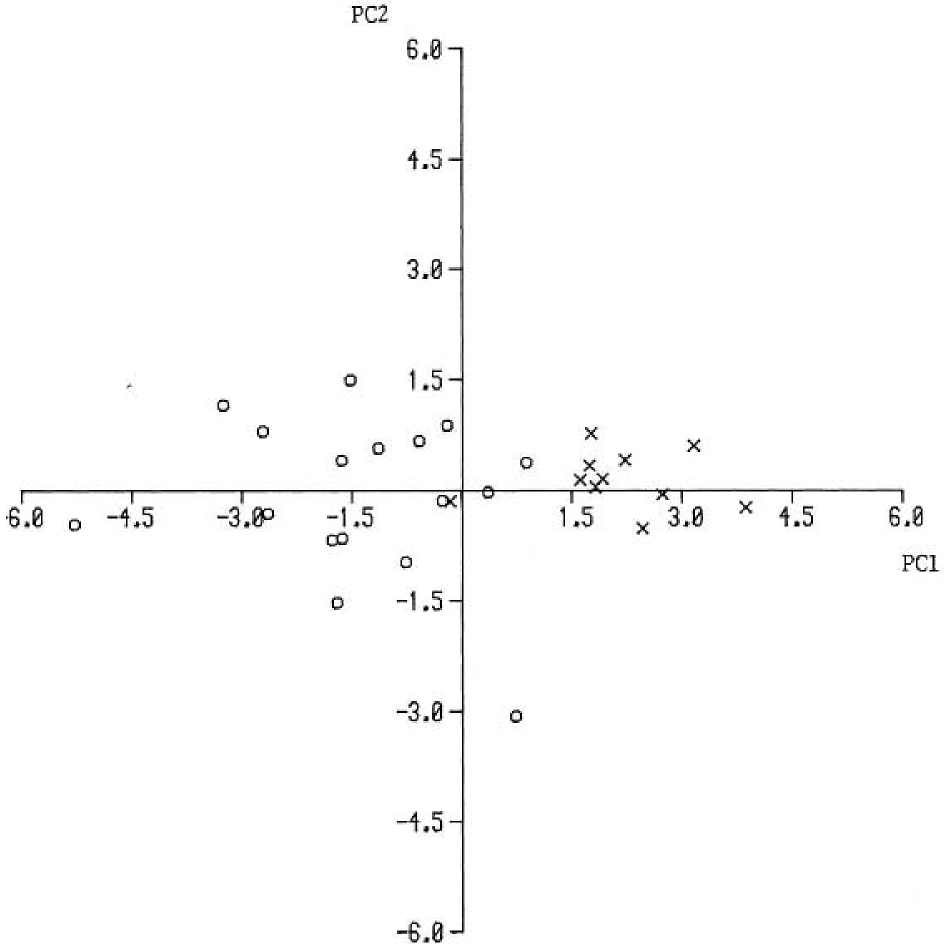

Figure 1.3. Student anatomical measurements: plots of 28 students with respect to their first two PCs. × denotes women; ○ denotes men.

2-dimensional picture of the data given in Figure 1.3 is a reasonably faithful representation of the positions of the 28 observations in 7-dimensional space. It is also clear from the figure that the first PC, which, as we shall see later, can be interpreted as a measure of the overall size of each student, does a good job of separating the women and men in the sample.

Having defined PCs, we need to know how to find them. Consider, for the moment, the case where the vector of random variables $\mathbf{x}$ has a known covariance matrix $\boldsymbol{\Sigma}$. This is the matrix whose (i, j)th element is the (known) covariance between the ith and jth elements of $\mathbf{x}$ when $i \neq j$, and the variance of the jth element of $\mathbf{x}$ when $i = j$. The more realistic case, where $\boldsymbol{\Sigma}$ is unknown, follows by replacing $\boldsymbol{\Sigma}$ by a sample covariance matrix $\mathbf{S}$ (see Chapter 3). It turns out that for $k = 1, 2, \cdots, p$, the kth PC is given by $z_k = \boldsymbol{\alpha}'_k\mathbf{x}$ where $\boldsymbol{\alpha}_k$ is an eigenvector of $\boldsymbol{\Sigma}$ corresponding to its kth largest eigenvalue λ_k. Furthermore, if $\boldsymbol{\alpha}_k$ is chosen to have unit length ($\boldsymbol{\alpha}'_k\boldsymbol{\alpha}_k = 1$), then $\text{var}(z_k) = \lambda_k$, where $\text{var}(z_k)$ denotes the variance of z_k.

The following derivation of these results is the standard one given in many multivariate textbooks; it may be skipped by readers who mainly are interested in the applications of PCA. Such readers could also skip much of Chapters 2 and 3 and concentrate their attention on later chapters, although Sections 2.3, 2.4, 3.3, 3.4, 3.8, and to a lesser extent 3.5, are likely to be of interest to most readers.

To derive the form of the PCs, consider first $\boldsymbol{\alpha}_1'\mathbf{x}$; the vector $\boldsymbol{\alpha}_1$ maximizes $\mathrm{var}[\boldsymbol{\alpha}_1'\mathbf{x}] = \boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1$. It is clear that, as it stands, the maximum will not be achieved for finite $\boldsymbol{\alpha}_1$ so a normalization constraint must be imposed. The constraint used in the derivation is $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 1$, that is, the sum of squares of elements of $\boldsymbol{\alpha}_1$ equals 1. Other constraints, for example $\mathrm{Max}_j|\alpha_{1j}| = 1$, may more useful in other circumstances, and can easily be substituted later on. However, the use of constraints other than $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 =$ *constant* in the derivation leads to a more difficult optimization problem, and it will produce a set of derived variables different from the PCs.

To maximize $\boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1$ subject to $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 1$, the standard approach is to use the technique of Lagrange multipliers. Maximize

$$\boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 - \lambda(\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 - 1),$$

where λ is a Lagrange multiplier. Differentiation with respect to $\boldsymbol{\alpha}_1$ gives

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 - \lambda\boldsymbol{\alpha}_1 = \mathbf{0},$$

or

$$(\boldsymbol{\Sigma} - \lambda\mathbf{I}_p)\boldsymbol{\alpha}_1 = \mathbf{0},$$

where $\mathbf{I}_p$ is the $(p \times p)$ identity matrix. Thus, λ is an eigenvalue of $\boldsymbol{\Sigma}$ and $\boldsymbol{\alpha}_1$ is the corresponding eigenvector. To decide which of the p eigenvectors gives $\boldsymbol{\alpha}_1'\mathbf{x}$ with maximum variance, note that the quantity to be maximized is

$$\boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_1'\lambda\boldsymbol{\alpha}_1 = \lambda\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = \lambda,$$

so λ must be as large as possible. Thus, $\boldsymbol{\alpha}_1$ is the eigenvector corresponding to the largest eigenvalue of $\boldsymbol{\Sigma}$, and $\mathrm{var}(\boldsymbol{\alpha}_1'\mathbf{x}) = \boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \lambda_1$, the largest eigenvalue.

In general, the kth PC of $\mathbf{x}$ is $\boldsymbol{\alpha}_k'\mathbf{x}$ and $\mathrm{var}(\boldsymbol{\alpha}_k'\mathbf{x}) = \lambda_k$, where λ_k is the kth largest eigenvalue of $\boldsymbol{\Sigma}$, and $\boldsymbol{\alpha}_k$ is the corresponding eigenvector. This will now be proved for $k = 2$; the proof for $k \geq 3$ is slightly more complicated, but very similar.

The second PC, $\boldsymbol{\alpha}_2'\mathbf{x}$, maximizes $\boldsymbol{\alpha}_2'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2$ subject to being uncorrelated with $\boldsymbol{\alpha}_1'\mathbf{x}$, or equivalently subject to $\mathrm{cov}[\boldsymbol{\alpha}_1'\mathbf{x}, \boldsymbol{\alpha}_2'\mathbf{x}] = 0$, where $\mathrm{cov}(x,\, y)$ denotes the covariance between the random variables x and y . But

$$\mathrm{cov}\ [\boldsymbol{\alpha}_1'\mathbf{x}, \boldsymbol{\alpha}_2'\mathbf{x}] = \boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_2'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2'\lambda_1\boldsymbol{\alpha}_1' = \lambda_1\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_1 = \lambda_1\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2.$$

Thus, any of the equations

$$\begin{aligned} \boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 &= 0, \quad & \boldsymbol{\alpha}_2'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 &= 0, \\ \boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2 &= 0, \quad & \boldsymbol{\alpha}_2'\boldsymbol{\alpha}_1 &= 0 \end{aligned}$$

could be used to specify zero correlation between $\boldsymbol{\alpha}_1'\mathbf{x}$ and $\boldsymbol{\alpha}_2'\mathbf{x}$. Choosing the last of these (an arbitrary choice), and noting that a normalization constraint is again necessary, the quantity to be maximized is

$$\boldsymbol{\alpha}_2'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda(\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_2 - 1) - \phi\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_1,$$

where λ, ϕ are Lagrange multipliers. Differentiation with respect to $\boldsymbol{\alpha}_2$ gives

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda\boldsymbol{\alpha}_2 - \phi\boldsymbol{\alpha}_1 = \mathbf{0}$$

and multiplication of this equation on the left by $\boldsymbol{\alpha}_1'$ gives

$$\boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2 - \phi\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 0,$$

which, since the first two terms are zero and $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 1$, reduces to $\phi = 0$. Therefore, $\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda\boldsymbol{\alpha}_2 = \mathbf{0}$, or equivalently $(\boldsymbol{\Sigma} - \lambda\mathbf{I}_p)\boldsymbol{\alpha}_2 = \mathbf{0}$, so λ is once more an eigenvalue of $\boldsymbol{\Sigma}$, and $\boldsymbol{\alpha}_2$ the corresponding eigenvector.

Again, $\lambda = \boldsymbol{\alpha}_2'\boldsymbol{\Sigma}\boldsymbol{\alpha}_2$, so λ is to be as large as possible. Assuming that $\boldsymbol{\Sigma}$ does not have repeated eigenvalues, a complication that is discussed in Section 2.4, λ cannot equal λ_1. If it did, it follows that $\boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_1$, violating the constraint $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2 = 0$. Hence λ is the second largest eigenvalue of $\boldsymbol{\Sigma}$, and $\boldsymbol{\alpha}_2$ is the corresponding eigenvector.

As stated above, it can be shown that for the third, fourth, ..., pth PCs, the vectors of coefficients $\boldsymbol{\alpha}_3, \boldsymbol{\alpha}_4, \ldots, \boldsymbol{\alpha}_p$ are the eigenvectors of $\boldsymbol{\Sigma}$ corresponding to $\lambda_3, \lambda_4, \ldots, \lambda_p$, the third and fourth largest, ..., and the smallest eigenvalue, respectively. Furthermore,

$$\text{var}[\boldsymbol{\alpha}_k'\mathbf{x}] = \lambda_k \qquad \text{for } k = 1, 2, \ldots, p.$$

This derivation of the PC coefficients and variances as eigenvectors and eigenvalues of a covariance matrix is standard, but Flury (1988, Section 2.2) and Diamantaras and Kung (1996, Chapter 3) give alternative derivations that do not involve differentiation.

It should be noted that sometimes the vectors $\boldsymbol{\alpha}_k$ are referred to as 'principal components.' This usage, though sometimes defended (see Dawkins (1990), Kuhfeld (1990) for some discussion), is confusing. It is preferable to reserve the term 'principal components' for the derived variables $\boldsymbol{\alpha}_k'\mathbf{x}$, and refer to $\boldsymbol{\alpha}_k$ as the vector of coefficients or loadings for the kth PC. Some authors distinguish between the terms 'loadings' and 'coefficients,' depending on the normalization constraint used, but they will be used interchangeably in this book.

1.2 A Brief History of Principal Component Analysis

The origins of statistical techniques are often difficult to trace. Preisendorfer and Mobley (1988) note that Beltrami (1873) and Jordan (1874)