REFERENCES

Giblett, M.A. (1932) The structure of wind over level country. *Meteorological Office Geophysical Memoir No. 54,* HMSO, London

Houghton, D. M. (1984) *Wind strategy.* Fernhurst Books

Ishida, H. (1989) Spectra of surface wind speed and air temperature over the ocean in the mesoscale frequency range in JASIN – 1978. *Boundary-Layer Meteorol.,* **47,** pp. 71–84

LeMone, M. A. (1973) The structure and dynamics of horizontal roll vortices in the planetary boundary layer. *J. Atmos. Sci.,* **30,** pp. 1077–1091

Meteorological Office (1975) *Handbook of weather forecasting*

Singleton, F. (1981) *Weather forecasting for sailors.* Teach Yourself Books

Watts, A. (1967) The real wind and the yacht. *Weather,* **22,** pp. 23–29

PRINCIPAL COMPONENT ANALYSIS: A BEGINNER'S GUIDE — I. Introduction and application

By IAN T. JOLLIFFE

Institute of Mathematics, University of Kent, Canterbury

PRINCIPAL Component Analysis (PCA) is a widely used technique in meteorology and climatology. Many papers which apply the method assume that the reader is familiar with its objectives, how to interpret its results, and what it can and cannot do. This may often not be the case, especially since there is some confusion over terminology and notation, and a wide variety of uses, some of which are far less valid than others. Indeed, it sometimes appears that the writers of climatological papers, as well as potential readers, do not fully understand the technique. In this paper I attempt to explain in simple terms what PCA really does, and the manner in which it is most frequently used in meteorology and climatology. Different terminologies are explained. A companion paper will discuss some common myths and pitfalls in implementing PCA, as well as connections with related or competing techniques and a number of extensions of the basic technique and its applications.

TOO MANY VARIABLES – REDUCING DIMENSIONALITY

When confronted with a very large dataset, a natural instinct is to try to reduce its size, whilst minimising any loss of information, in order to better understand and interpret the structure of the data. A typical dataset can be viewed as n observations measured on p variables. Thus if maximum temperature was measured daily for a year at 50 different recording stations we would have $n = 365$ observations on $p = 50$ variables. This is the most common format for meteorological or climatological data (*i.e.* one meteorological variable measured on n occasions at p sites), and most of what follows assumes this type of data. However, it should be noted that other types of data can occur, *e.g.* p meteorological variables measured at n stations on a single occasion, or p meteorological variables measured at a single station on n occasions.

Often the p variables are highly correlated; this will certainly be true if the p variables correspond to p stations and some of the p stations are geographically very close. High correlation implies that the 'true' dimension of the dataset is less than p, *i.e.* we can choose m variables, where m may be substantially less than p, which convey virtually all the information in the original p variables. There are essentially two possible strategies for finding such a set of m variables. The first, and perhaps the most obvious, is to choose a

subset of our original p variables. For instance, in our example, one might choose 10 of our original 50 stations which we believe to be representative of the complete set of 50. For example, in the *Weather Log*, the geographical variation of a month's weather is summarised using only a couple of dozen from the far larger number of potential stations.

An alternative strategy is to build new variables from the original ones, so that each of our selection of m variables is typically different from any of the original p, but is constructed from them. This approach has less intuitive appeal than choosing a subset, but has the advantage that, for the same amount of information loss, we can achieve a greater reduction in dimensionality.

PCA is the simplest of these variable-building techniques. Its simplicity lies in its restriction to *linear* functions of the original variables.

PRINCIPAL COMPONENT ANALYSIS

Denote the p variables by $x_1, x_2, \ldots, x_p$. For example, x_1 might be the maximum temperature at Station 1, and $x_2, x_3, \ldots, x_p$ are similarly maximum temperatures at Stations 2,3, $\ldots, p$. A linear function of the p variables will be of the form $z = \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_p x_p$, where $\alpha_1, \alpha_2, \ldots, \alpha_p$ are constants. As we change $\alpha_1, \alpha_2, \ldots, \alpha_p$ we get different linear functions, and we can calculate the variance of any such linear function. The first principal component (PC) is that linear function which has the maximum possible variance, the second PC is the linear function with maximum possible variance subject to being uncorrelated with the first PC, the third PC is the linear function which maximises variance subject to being uncorrelated with the first and second PCs, and so on. Altogether we could construct p PCs but this gives no reduction in dimensionality. PCA provides the optimal m-dimensional representation of the data for each $m = 1, 2, \ldots, p-1$, for various different definitions of optimality (Jolliffe 1986). In particular, at each stage the sum of the variances of the PCs is as large as possible. In other words, with PCA we have, for each $m = 1, 2, \ldots p-1$, the m linear functions of $x_1, x_2, \ldots, x_p$, which account for the maximum possible proportion of the original variation. Before going much further we need to introduce and explain some terminology associated with PCA, but first we look at a simple artificial example.

A TWO-DIMENSIONAL EXAMPLE

Figure 1 shows a scatter plot of 50 points in two dimensions corresponding to 50 measurements on a pair of variables x_1, x_2. For example, x_1, x_2 respectively might represent soil temperature and air temperature on 50 days at a particular site. Suppose that we wish to reduce the dimensionality of this dataset. Replacing the pair of variables by x_1 or x_2 alone will not be very successful; there is a considerable amount of variation in each variable, although rather more in x_2 than x_1, so omitting one of them throws away a non-trivial proportion of the original variability. However, it will be noticed that the points are scattered fairly closely about a straight line. This implies that there is a linear function of x_1, x_2 which explains a substantially larger amount of variation than either x_1 or x_2 alone. The linear function which maximises this variance is the first PC, z_1, and Fig. 2 plots the same 50 observations with respect to z_1, and z_2, the second PC. It can now be seen that nearly all the variation in the data can be accounted for by the single dimension defined by z_1.

Two comments are worth making at this stage. The first is that Fig. 2 is identical to Fig. 1 except that the axes have been rotated, so that the z_1 axis goes through the middle of the points (and the second (z_2) axis is constrained to be at right angles to the first). For $p > 2$ variables the effect of performing a PCA is to similarly rotate the axes with respect to which the observations are measured.

The second comment is to note the similarity of what we have done in our two-dimensional example to fitting a regression line through the 50 points. The crucial difference is that in a regression of x_2 on x_1 we fit the line which minimises sums of squared distances in the *vertical* direction, whereas in PCA the fitted line minimises the sum of squared distances *perpendicular to the line*.

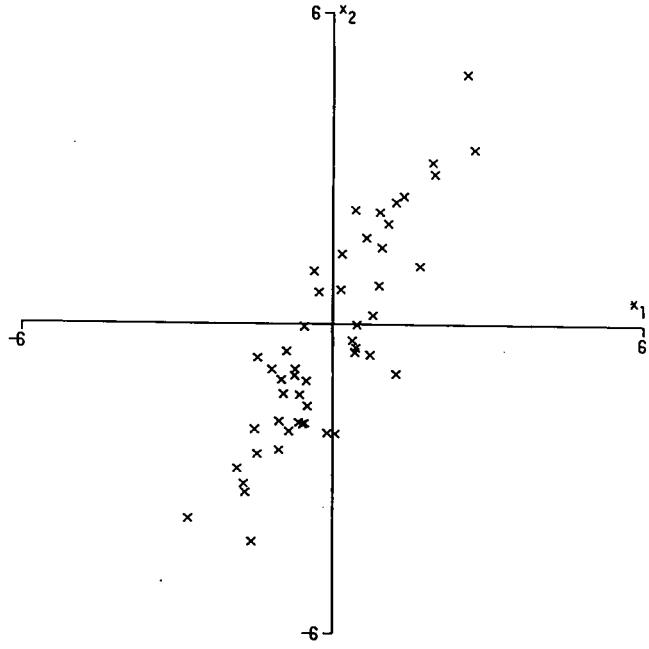

Fig. 1 Plot of 50 observations on two variables, x_1, x_2

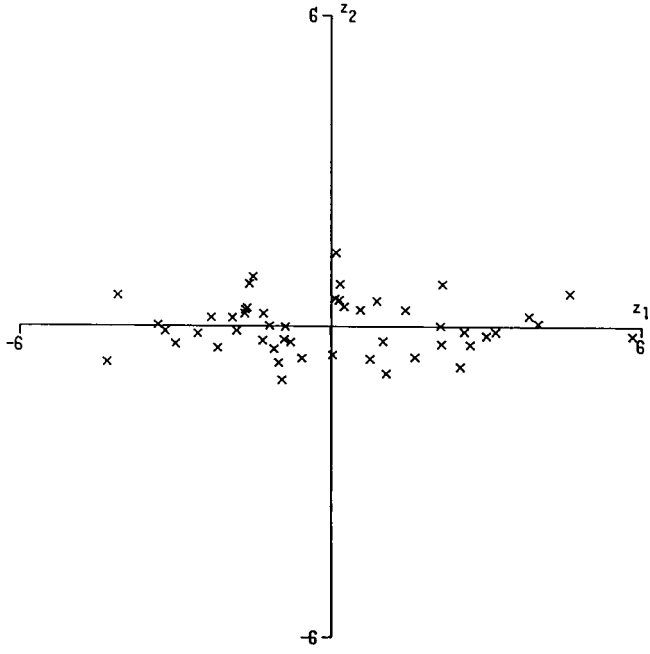

Fig. 2 Plot of the 50 observations from Fig. 1 with respect to their principal components, z_1, z_2

A BRIEF HISTORICAL NOTE

PCA was originally defined in a statistical context by Pearson (1901) via an extension of the geometric argument just presented. If we plot n observations as points in p dimensional space, we can define PCs by successively finding a line, plane, hyperplanes of dimension 1, 2, 3, . . . from which sums of squared perpendicular distances to the points are minimised. The more usual definition in terms of successive maximisation of variance came 30 years later – Hotelling (1933). As we shall see in the next section, to actually determine the PCs we need non-trivial mathematics and, for all but very small values of p, substantial computational effort. Thus, it was not until the advent of electronic computers in the late 1940s that the method could be implemented on realistically sized problems. The first uses in meteorology and climatology date from this period. Preisendorfer (1988) gives a review of early examples, with Wadsworth *et al.* (1948) as the first that he cites. Preisendorfer also gives some references which pre-date Pearson (1901), but which look at the ideas of PCA in an abstract, and not data-analytic way.

The first book to appear on PCA was by Daultrey (1976) which is a slim volume (51 pages) aimed at geographers. Two recent books by Jolliffe (1986) and Preisendorfer (1988) give a much broader coverage of the technique. Preisendorfer's book concentrates on PCA within the context of meteorology and oceanography, and gives a comprehensive theoretical account. Surprisingly, it is much less practically oriented than Jolliffe's text which is aimed at a wider audience.

The most recent book to be added to the literature is Dunteman (1989). Like Daultrey (1976), it is a short book (96 pages) written for a mathematically unsophisticated readership. Its target audience consists mainly of social scientists but, apart from its examples, it would be suitable for readers from other disciplines such as climatology.

The alert reader will have noticed that two of the cited books use (like this paper) the expression 'Principal Component Analysis'; the other two add an 's' to Component. Both forms are in widespread use, and mean exactly the same thing.

COMPUTATION OF PRINCIPAL COMPONENTS — EIGENVALUES AND EIGENVECTORS

The actual computation of PCs will typically be done not by hand, but by a standard computer package. However, in order to interpret the output it is desirable to know something of the way in which the computations are done. Unfortunately this takes us into the realms of matrix algebra, and in particular the concept of eigenvalues and eigenvectors.

It might be possible to use and interpret PCA without knowing anything about eigenvalues and eigenvectors were it not for the fact that the terms are freely used in papers describing applications, and in the output of computer packages, so it is useful to know what they mean. We present the basic ideas here, with a little more information in Appendix I. The output for PCA from a computer package will typically give columns of numbers labelled 1st eigenvector, 2nd eigenvector, etc. Recall that PCs are linear functions of $x_1, x_2, \ldots, x_p$. Suppose then that the first PC is

$$z_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1p}x_p.$$

More generally, suppose that the kth PC is

$$z_k = a_{k1}x_1 + a_{k2}x_2 + \ldots + a_{kp}x_p,$$

for $k = 1, 2, \ldots, p$.

The first eigenvector is simply the set of coefficients $a_{11}, a_{12}, \ldots, a_{1p}$ appearing in the first PC. Similarly, subsequent eigenvectors consist of coefficients of $x_1, x_2, \ldots, x_p$ in each successive PC. The term 'eigenvalue' is also encountered in computer output. The first eigenvalue is the variance of the first PC, and therefore a measure of its importance in explaining variation. Second, third, and subsequent eigenvalues are similarly the variances of the second, third and subsequent PCs. Having said that the terms 'eigenvector/eigenvalue' are commonly encountered, the reader should be warned that other terminology is

sometimes used instead, for example latent vectors/latent roots, characteristic vectors/characteristic roots, proper vectors/proper values. We have referred to the αs as *coefficients* above; other commonly used terminology includes *loadings* or *weights*, but they are also sometimes referred to as the principal components – this is clearly *wrong*. z_k, and not its corresponding eigenvector, is the kth principal component. The z_ks may simply be called the PCs, or the PC *scores*, or the *amplitudes*.

COVARIANCES OR CORRELATIONS?

Next we need to discuss some variations on the basic definition of PCs. First, we have assumed that we are working with the variables as given, and that our analysis is therefore based on the so-called covariance matrix for our variables. You will find in practice that PCs are often found from *correlations* between $x_1, x_2, \ldots, x_p$. What is effectively being done in such an analysis is to standardise $x_1, x_2, \ldots, x_p$, by dividing each by its standard deviation, and then finding linear functions of these standardised variables which successively maximise variance. Using standardised variables has the effect of giving all variables equal weight, whereas the original variables may have vastly different variances. In this latter situation, the high-variance variables will dominate the first few PCs, which is often undesirable, although sometimes it can be exactly what is wanted. A second argument for using standardised variables, and hence correlations, is that the variables may be measured in different units. For example, some may be temperatures, others atmospheric pressures, others rainfall amounts, and so on. In this case, the relative sizes of the variances and covariances depend crucially, and arbitrarily, on the units used to measure the various different elements. Standardisation of variables is an obvious strategy for overcoming this arbitrariness.

A second variation of PCA which is more apparent than real, in that the basic interpretation of a PC is unaffected, concerns the subject of normalisation. This is discussed in Appendix II.

AN EXAMPLE

It is an unfortunate necessity that we have reached the final section of this paper before presenting a real example. Most of the terminology and discussion presented so far is needed to interpret the example. The data consist of mean-sea-level pressure for 120 half-months in January/February from 1951 to 1980, measured at (or rather interpolated to) a 20 $\times$ 15 grid of points over the area of Europe and the North Atlantic displayed in Figs. 3 to 5. Thus the number of observations is 120 and the number of variables is 300($= 20 \times 15$) and the PCs discussed below are based on the correlation matrix for the 300 variables. Using correlations means that we are giving all 300 stations equal weight in our analysis – see above. A great deal of meteorological and climatological data have a similar form, in which variables correspond to different grid points or observing stations, and for such data the PC loadings can be represented conveniently in the form of maps such as Fig. 3. For each PC we have a loading, a numerical value, for each station or grid point. We could leave these loadings as a column of numbers (which is what most computer packages will provide as output), but if we wish to interpret the PCs (*i.e.* which geographical areas are important in which PCs) it will be easier to do so if we display them at the appropriate geographical location on a map, and even better if we then draw contours through them, as in Figs. 3 to 5.

So how do we interpret these figures? In Fig. 3, virtually the whole region has positive loadings, so the first PC is a weighted average of pressure at all grid points with larger weights being given to points near to the centre of the region. The fact that the first PC takes this form means that the major source of variation in the dataset is between, on the one hand, half-months when the (weighted) average pressure over the whole region was high and, on the other, half-months when this average pressure was low. This single first dimension accounts for 40.9 per cent of the total variation in the (standardised) data. The reason for the weights (coefficients) being unequal is simply that points near the edges of the region have smaller average correlation with all other points than do those in the centre.

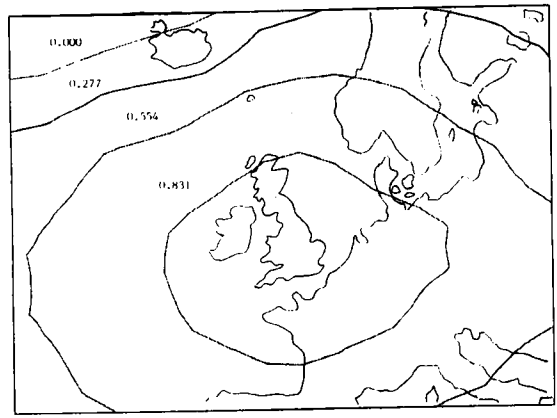

Fig. 3 Mean half-monthly sea-level pressure – first principal component

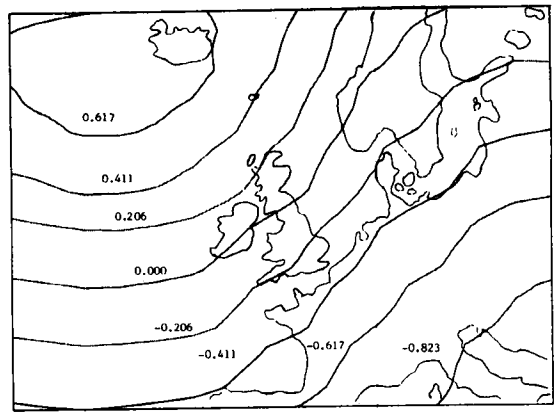

Fig. 4 Mean half-monthly sea-level pressure – second principal component

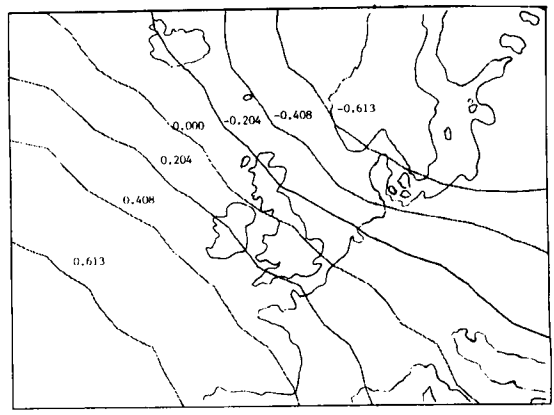

Fig. 5 Mean half-monthly sea-level pressure – third principal component

Indeed, it has been argued that the patterns shown in figures such as Figs. 3 to 5 can be predicted to some extent by the shape of the geographical area examined, without any knowledge of the data. This point will be addressed in the companion paper, as will a number of extensions to the PCA methodology, together with various pitfalls that await the unwary.

The loadings in Figs. 3 to 5 have been plotted here with the loadings normalised to have

$$\sum_{i=1}^{p} \alpha_{ki}^2 = \lambda_k \text{ (see Appendix II)},$$

so they represent the correlations between the PC and the original variables. The peculiar values associated with each contour are due to an idiosyncracy of the plotting package, but they serve to emphasise the point that if we change the normalisation constraint the figure does not change; only the labels on the contours will be different.

Turning to Fig. 4, we see that there are positive coefficients in the north-west of the map, negative coefficients in the south-east. What this means is that *after removing the first dominant source of variation* the next major source of variation is a contrast between those half-months in which there was high pressure in the north-west and low pressure in the south-east, and the half-months for which the opposite pressure pattern holds. This second component accounts for a further 22.6 per cent of the original variables.

Similarly in Fig. 5, after removing the first two major sources of variation, the next most important contrast is between half-months with high pressure in the south-west, low pressure in the north-east, and half-months with the opposite tendency. The third component accounts for an additional 22.2 per cent of the total variation. Thus a total of 86 per cent of the variation in 300 variables is accounted for in just three dimensions, illustrating the great dimension-reducing potential of PCA.

ACKNOWLEDGEMENTS

I am very grateful to Miss J. M. Potts for producing Figs. 3 to 5, and to Springer-Verlag for permission to reproduce Figs. 1 and 2. The Editor made many useful suggestions for improving the presentation of the paper.

REFERENCES

Daultrey, S. (1976) *Principal Components Analysis*. Geo Abstracts, Norwich

Dunteman, G. H. (1989) *Principal Components Analysis*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-069, Sage, Beverly Hills

Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. *J. Educ. Psych.*, **24**, pp. 417–441

Jolliffe, I. T. (1986) *Principal Component Analysis*, Springer-Verlag, New York

Pearson, K. (1901) On lines and planes of closest fit to systems of points in space. *Philos. Mag. Ser.6*, **2**, pp. 559–572

Preisendorfer, R. W. (1988) *Principal Component Analysis in meteorology and oceanography*. Elsevier, Amsterdam.

Richman, M. B. (1986) Rotation of principal components. *J. Climatol.*, **6**, pp. 293–335

Wadsworth, G. P., Bryan, J. G. and Gordon, C. H. (1948) *Short range and extended forecasting by statistical methods*. U S Air Force, Air Weather Service Technical Report No. 105–38, Washington D C

APPENDIX I

A more compact way of writing $z = \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_p x_p$ is $z = \boldsymbol{\alpha}^T \mathbf{x}$, where $\boldsymbol{\alpha}$, $\mathbf{x}$ are vectors consisting of the αs and xs respectively, and T denotes transpose. The way in which PCs are defined, either algebraically or geometrically, means that we are looking for linear functions $\boldsymbol{\alpha}^T \mathbf{x}$ of $\mathbf{x}$ which optimise some criterion (maximise variance, or equivalently minimise sums of squared perpendicular distances) subject to constraints. It is fairly straightforward, mathematically speaking, to find $\boldsymbol{\alpha}$ which optimises the criterion, subject to constraints (see, for example, Jolliffe 1986). It turns out that we must find eigenvalues

and eigenvectors of the $(p \times p)$ covariance matrix, S. The matrix S has as its (i, j)th element the covariance between x_i and x_j, and the scalar λ_k and the vector $\boldsymbol{\alpha}_k$ are respectively an eigenvalue and corresponding eigenvector of S if they satisfy the equation

$$\mathbf{S}\,\boldsymbol{\alpha}_k = \lambda_k\,\boldsymbol{\alpha}_k.$$

Solving this equation for λ_k reduces to finding the p roots of a polynomial equation. If all of the roots are distinct, then there are p different, orthogonal $\boldsymbol{\alpha}_k$, $k = 1, 2, \ldots, p$, uniquely defined (apart from the choice of normalisation constraint, discussed below), corresponding to the p eigenvalues which are conventionally labelled in descending order $\lambda_1 > \lambda_2 > \ldots > \lambda_p$. The case of exactly equal λs is unusual and will not be discussed further in the present paper.

APPENDIX II – NORMALISATION CONSTRAINTS

To uniquely define the PCs, a normalisation constraint needs to be imposed on the αs – otherwise we can increase the variance of z_k without bound, simply by multiplying all of the αs in z_k by the same constant (*e.g.* $10z_k$ has a variance 100 times as large as z_k). The constraint which leads to the conventional set of PCs is

$$\sum_{i=1}^{p} \alpha_{ki}^2 = 1,$$

but once we have found the αs we can change these constraints if we wish. The basic interpretation of each PC is unchanged, since the *relative* loadings of each variable are unchanged.

There are two competing normalisations which may be encountered. By far the most common is

$$\sum_{i=1}^{p} \alpha_{ki}^2 = \lambda_k,$$

where λ_k is the variance of z_k. Compared with the original normalisation, this typically has the effect of increasing the loadings in the first PCs, whilst decreasing those of the later PCs. The great attraction of this normalisation is that, if we are dealing with a correlation matrix, then α_{kj} is the correlation between x_j and the kth PC. For this reason, PCs are presented with this normalisation by several standard computer packages such as BMDP and SPSS. Some authors, notably Richman (1986) reserve the term 'principal components' for z_k with this competing normalisation. He then refers to the corresponding quantities with the original normalisation

$$\sum_{i=1}^{p} \alpha_{ki}^2 = 1$$

as 'empirical orthogonal functions' or EOFs, a terminology which is quite widely used in meteorology. Although the 'correlation interpretation' of the alternative normalisation makes it attractive, the variances of the resulting PCs are distorted, so that if individual observations are plotted with respect to the first two components, the plot will give an exaggerated impression of the variation of the first component compared with the second. Similar distortions will occur for other plots, so a gain in one type of interpretation is matched by a loss in another.

A third possible normalisation is

$$\sum_{i=1}^{p} \alpha_{ki}^2 = 1/\lambda_k$$

which gives $\text{var}(z_k) = 1$, for all $k = 1, 2, \ldots, p$. For some purposes, such as the detection of outlying observations (see Jolliffe 1986) equalising the contribution of each component in this way has advantages. As with other normalisations, however, it is really only when we go beyond the PCs as a simple descriptive tool that the normalisation chosen matters much. As noted earlier in this section, the basic interpretation of a PC is not changed by changing its normalisation.