Chapter 7

Principal component analysis

7.1 Introduction

In Chapter 2 principal component analysis (PCA) and correspondence analysis were introduced to show how easy it is, using `R`, to undertake such analyses. It is (almost) as simple as calculating a mean. To remind the reader, given a suitable $n \times p$ table of data, $\mathbf{Y}$ entered as `Y` in `R`,, a basic PCA can be carried out using `prcomp(Y)`.

Calculation of a mean should not be carried out without thought. It is useless if the data are seriously multi-modal; can be compromised if there are obvious outliers in the data; and is not sufficient as a single summary measure of location if the data are highly skewed. Preliminary data inspection is called for, and PCA is no exception. Principal component analysis is conceptually, as well as computationally, simple. As usually applied, it takes an $n \times p$ data set and reduces it to a 'picture' that allows patterns in the data to be investigated using conventional two- and three-dimensional plots. There are practicalities to be addressed in applications and interpretation, and this chapter discusses and illustrates these.

Assuming $n > p$, the table of data is p-dimensional and not susceptible to conventional methods of data exploration if $p > 3$. It is possible to define the *distance* between the rows of data cases (Section 7.3); one way of thinking about PCA is that it is designed to approximate these distances in two- or three-dimensional space using the first two or three principal components (PCs). Inevitably information is lost in the approximation so some means of measuring its quality is desirable.

Sections 7.2 to 7.4 are centered on examples, used to introduce the ideas that underpin PCA, implementation and interpretation. The algebra that underpins PCA is covered in Appendix D. It is helpful to have some acquaintance with the associated terminology that finds its way into software output. It is perfectly

possible to apply PCA usefully without needing to master the algebra. There is a focus in Section 7.2 on data standardization and transformation, and Section 7.3 interpolates a brief discussion of the important concept of distance before the example of Section 7.4.

The earliest uses of PCA in archaeology date back to the 1960s. This coincided with the start of a period when the method of *factor analysis* was in vogue. The two methods were, and still are, confused. Chapter 8 attempts to explain what these differences are and why they can be considered fundamental. The opportunity is taken, in Section 8.4, to recount, briefly, some of the history of the use of the methods in archaeology, along with critical comment on more recent advocacy of factor analysis.

7.2 Example 1 – Roman glass compositions

Standardization and log-transformation

The data consist of 105 specimens of Romano-British waste glass, measured with respect to nine major and minor oxides, excavated from sites at Leicester and Mancetter in the UK (Tables B.7 and B.8)[1]. One question is whether or not the glass from the two sites is chemically distinct. The analysis will be used as a peg on which to hang a discussion of *data transformation* in PCA.

The initial data table (or matrix), $\mathbf{X}$, has a typical entry x_{ij}, with $\bar{x}_j$ and s_j denoting the estimated mean and standard deviation of variable (column) j, $j = (1, 2, \ldots, p)$. Invariably, PCA is carried out after some transformation of the original data to a data matrix $\mathbf{Y}$.

The simplest form of transformation is *centering* where y_{ij} is defined as

$$y_{ij} = (x_{ij} - \bar{x}_j).$$

This is the first stage in packages where PCA is available, but usually analysis goes beyond this for two reasons. One is that, if variables are measured in different units, the use of centered data alone is inappropriate as the variables are not comparable. The second reason, illustrated in Figures 7.2 and 7.3, is that even if the variables are in the same units a PCA of centered data will be dominated by the variables with the larger variances, so potential information provided by other variables is lost. Thus, unless variables have similar variances to begin with, some further transformation beyond centering is called for.

[1]The data were collected by Dr. Caroline Jackson of Sheffield University, UK, as part of an unpublished PhD thesis. Some analyses are undertaken in Baxter (1994a).

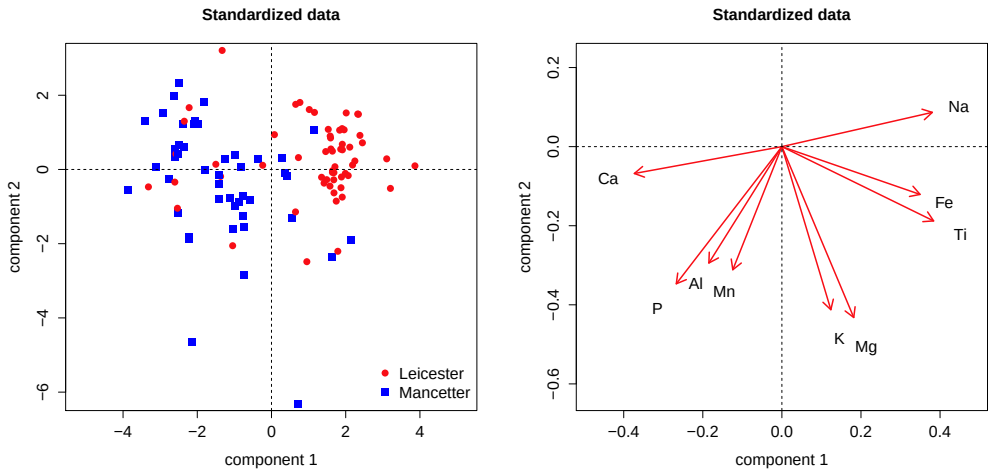

Figure 7.1: *PCA score and variable plots for the standardized Romano-British glass data of Tables B.7 and B.8.*

The most common form of data pre-treatment, and the default in many software packages though not R, is to *standardize* the data as

$$y_{ij} = (x_{ij} - \bar{x}_j)/s_j$$

producing variables with a mean of 0 and standard deviation (and variance) of 1. This gives each variable equal 'importance' so that each has an 'equal chance' of influencing the PCA. A biplot for the PCA of standardized data is shown in Figure 7.1 presented as adjacent score and variable plots.

This can be contrasted with the output obtained using log-transformed (to base 10) data (Figure 7.2)

$$y_{ij} = \log x_{ij}$$

which is the other approach to have found widespread use. This is applied before any subsequent centering and standardization [2].

It is quite common to standardize the data after log-transformation, though this then often produces results very similar to standardization of the original data (Baxter, 1995). Values recorded exactly as zero cannot be log-transformed and the problem is usually resolved by adding a small value to the data (see Section 8.3.2 for an example).

[2]Terminology in the literature is confusing. Standardization is sometimes called *normalization*, implying that the transformed variables have a normal distribution. They will not, unless the untransformed data begin with a normal distribution. To confuse matters further log-transformation is sometimes referred to as standardization. It is also sometimes used with the hope that it will induce normality.

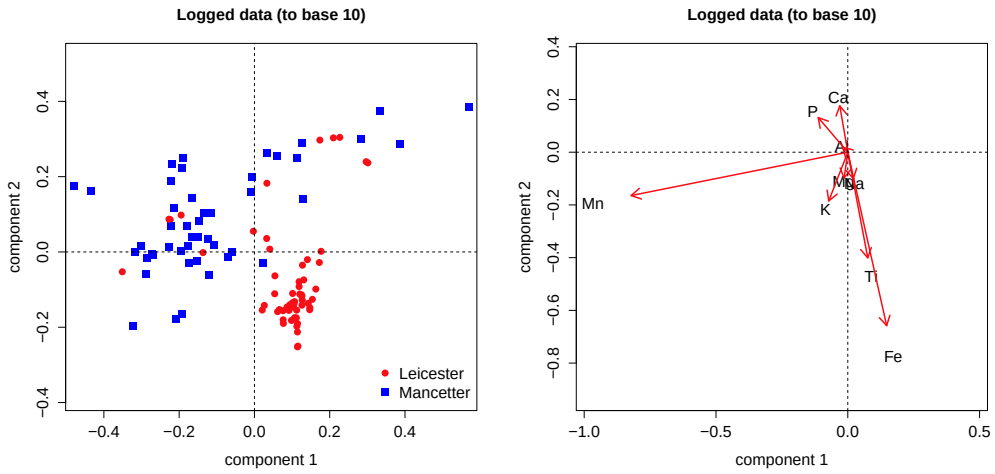

Figure 7.2: *PCA score and variable plots for the unstandardized log-transformed Romano-British glass data of Tables B.7 and B.8.*

Some of the differences that can arise between the use of standardized and unstandardized log-transformed data are illustrated in Figures 7.1 and 7.2. The two sites largely separate out, with the grouping for Leicester the more compact. There are about 10 cases that plot more closely with the Mancetter data. The score plot using log-transformed data suggests a possible sub-division among the Mancetter data.

The variable plots are more obviously distinct. The variable markers for the standardized data lie (very roughly) round a circle, equidistant from the origin. This reflects the equal weighting induced by standardization, though it is still possible for some variables to have little effect on the PCA, or to play a lesser role (e.g., Al, Fe, Mn). If row (cases) and column (variable) plots are compared it can be seen for the row plot that the Leicester data lie largely to the right and the Mancetter data to the left. For the variables, (Fe, Na, Ti) plot to the right, while (Al, Ca, Mn, P) plot to the left. It can be inferred that the Mancetter glasses are richer in the latter group and poorer in the former group, relative to Leicester. This is readily checked, where it can be seen that the variable means for the two sites support this inference (Table 7.1).

Site	Al	Fe	Mg	Ca	Na	K	Ti	P	Mn
Leicester	2.38	0.70	0.55	6.59	18.20	0.71	0.10	0.12	0.27
Mancetter	2.47	0.48	0.53	7.19	17.20	0.72	0.08	0.14	0.41

Table 7.1: *Variable means (%) for the Romano-British glass and the two sites.*

The variable plot for the log-transformed data is rather different, being dominated by Mn and Fe. This implies that these two variables have much the largest variances on the log-transformed scale, and this too is readily checked. The variances for Mn and Fe are 0.0338 and 0.0175 respectively, with the next largest that for Ti of 0.007. The pattern in the row plot for the log-transformed data is thus dominated by the effect of Mn and Fe, in contrast to that for standardized data. The row plots, while showing differences, are not incompatible with each other, pointing to the fact that different transformations can give rise to similar patterns in the data, even though different variables may be responsible. Similarly, it is possible for row plots to differ, but admit equally valid archaeological interpretations. The message is that in exploratory work the examination of different transformations is worthwhile.

The interpretation of biplots was discussed briefly in Chapter 2, and some elaboration is provided here. Further discussion is provided in Section 9.2 after their use in correspondence analysis has been introduced. What is understood by the term 'biplot' varies – I err on the side of a more informal usage, allowing the term to embrace the joint presentation of row and column plots, rather than superimposing them, for example. It is easy enough to imagine the two plots being superimposed, provided a common origin is indicated and correct aspect ratios are used, though issues of axis scaling arise (see Section 9.2).

With the caveat that the PCA should be of reasonable quality, variable markers that lie opposite to each other on the plot should have a negative correlation (e.g., Ca and Na in the plot for standardized data); variables at right angles should show weak correlation (e.g., K, Na); variables plotting close to each other, with an acute angle at the origin, should exhibit strong positive correlation (e.g., Fe, Ti). This can be seen to be broadly the case from Table 7.2, which shows correlations to one significant digit, with the ordering based on reading clockwise from Na on the variable plot for standardized data. It can be seen from Figure 7.2 that the Mancetter glass has a comparatively higher concentration of Mn than Leicester where the concentration of Fe stands out. As already noted the coincidence of interpretation of the score plots can be attributed to different subsets of variables and is not uncommon in applications that contrast standardized with unstandardized log-transformed data. It is not inevitable; situations where the latter approach highlights variables with a low absolute presence leading to no useful interpretation are also not uncommon.

Other forms of transformation have been proposed but little used. Baxter (1995) suggested rank-transformation as a possibility, and 'standardizing' variables to the range [0,1] has occasionally been seen. More needs to be said about log-ratio transformations.

	Na	Fe	Ti	Mg	K	Mn	Al	P	Ca
Na	1	0.5	0.6	0.3	0.2	-0.2	-0.4	-0.6	-0.8
Fe	0.5	1	0.8	0.4	0.2	-0.2	-0.1	-0.3	-0.6
Ti	0.6	0.8	1	0.5	0.3	-0.2	-0.1	-0.3	-0.7
Mg	0.3	0.4	0.5	1	0.4	0.1	0	0.1	-0.1
K	0.2	0.2	0.3	0.4	1	0.2	0.1	0.1	-0.2
Mn	-0.2	-0.2	-0.2	0.1	0.2	1	0.2	0.4	0.1
Al	-0.4	-0.1	-0.1	0	0.1	0.2	1	0.3	0.4
P	-0.6	-0.3	-0.3	0.1	0.1	0.4	0.3	1	0.5
Ca	-0.8	-0.6	-0.7	-0.1	-0.2	0.1	0.4	0.5	1

Table 7.2: *Correlations between variables for the Romano-British glass data.*

Log-ratio transformation

Data of the kind used here are sub-compositional. They would be fully compositional, adding to 100%, if all the naturally occurring elements were measured. Only a subset are ever measured and hence the data are *sub-compositional*. It is possible to convert such data to fully compositional form, either by rescaling to 100% or defining a 'residual' as the sum of the measured data subtracted from 100%. In the present case the 'residual' will largely coincide with the silica content of the glass (not measured by the instrumentation used).

The use of ratio transformations has been debated in archaeology from time to time and was explored more generally in the seminal text of Aitchison (1986). He advocated the use of log-ratio transformation, a symmetric version being

$$y_{ij} = log(x_{ij}/g(\mathbf{x}_i))$$

where $g(\mathbf{x}_i)$ is the geometric mean of row i, defined as the $1/p$th root of the product of the elements of row i. An argument for this is that the raw compositional data are positive and constrained to lie in a $(p-1)$-dimensional space, for which the more common methods of analysis, including the use of standardization, are inappropriate. The log-ratio transformation removes the constraints on the data, allowing standard methods to be applied. This is illustrated in Figure 7.3, where the original composition is augmented by the constructed 'silica' (Si) variable.

The analysis is practically indistinguishable from that of log-transformed data. This is precisely the point. Aitchison's advocacy of the methodology he developed for analyzing compositional data is theoretically compelling. Baxter (1989) used the methodology for analyzing glass compositional data and was initially enthusiatic, but later exploration suggested that, regardless of theory, the more usual methods of analysis often produced equivalent or more satisfactory and archaeologically interpretable results. This is because log-ratio transformed data are not

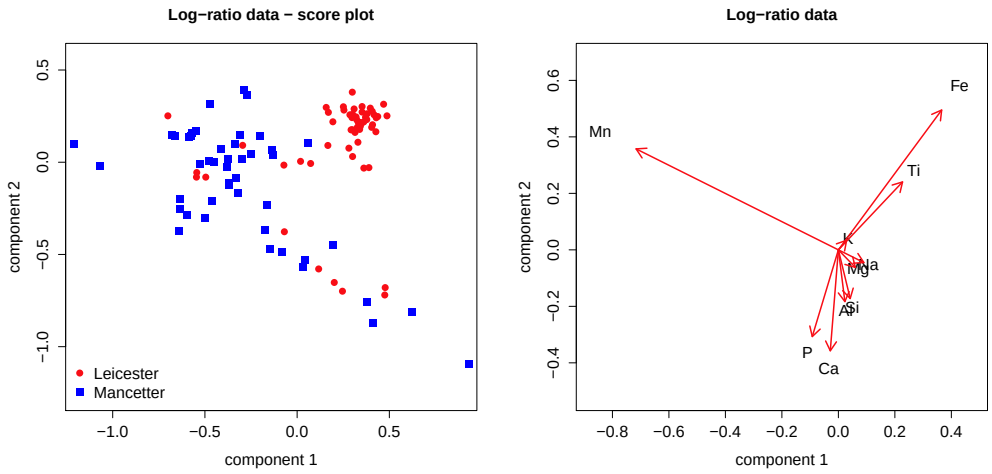

Figure 7.3: *PCA score and variable plots for the unstandardized log-ratio-transformed Romano-British glass data of Tables B.7 and B.8.*

usually standardized in subsequent analysis. This means that variables with a low absolute presence and high relative variance are emphasized, at the expense of variables with a greater presence that may be more important for understanding the production processes that produced the glass.

The dominance of the variables with low absolute presence is what is illustrated in Figure 7.3. It happens to make sense in that interpretable results, in the form of as good a separation between sites as can be reasonably be expected, is attained. This is also, in a sense, 'accidental' since, as noted, variables with smaller typical values can dominate log-ratio (and log-transformed) analyses to no good effect. Usually this can be 'corrected' for by adopting pragmatic measures, omitting such variables from an analysis for example, but this often then leads to outcomes similar to the simpler (if not necessarily 'theoretically correct') methods that are prevalent in the literature. Log-ratio analysis has not really caught on in archaeological applications but anyone with compositional data to analyze should be aware of the issues involved; a detailed account of what these are is provided in Baxter and Freestone (2006)[3].

[3]A lot of work has been done on developing a rigorous mathematical framework for compositional data analysis (Pawlowsky-Glahn and Buccianti, 2011; Pawlowsky-Glahn *et al.*, 2015). Advocates of log-ratio methodology can be dismissive of analytical approaches that do not conform to their theoretical *dicta*. My practice, when confronted with compositional data, is to examine a range of analyses including the use of log-ratios, so my opinions are driven by practical experience (relating primarily to multivariate data analysis) rather than theoretical agonizing. An `R` package, `compositions`, is available along with a book devoted to its use (van den Boogaart and Tolosana-Delgado, 2013) for those wishing to explore the ideas and application.

7.3 The idea of distance

A fundamental idea behind PCA is that, using all the components (PCs), the distance between cases on the scale of the data used is reproduced. A subset of the first few PCs allows distances to be approximated in a low-dimensional space that can be more readily interrogated using standard graphical methods. The distances approximated depend on the standardization/tranformation used. As will be seen in later chapters, different definitions of distance underpin, and distinguish between, different methods of multivariate analysis. The (mathematically) simplest application of the ideas arises in the context of PCA, and a general discussion is provided here.

Given two rows of a data matrix – call them $\mathbf{y}_i$ and $\mathbf{y}_k$ – it is possible to measure exactly the distance between them, d_{ik}, in p dimensions. Several multivariate methods work by defining new variables, in which the rows are $\mathbf{z}_i$ and $\mathbf{z}_k$. For the first $r < p$ columns the distance between the rows can be defined, but will only approximate the true distances. This begs the questions of how to define the new variables and how the quality of approximation is judged. These will be dealt with in the context of the example in Section 7.4.

The point to stress is that distance, d_{ik}, is not a uniquely defined concept. Any proposed measure of distance qualifies as such if it satisfies a particular set of mathematical rules. Different measures of distance are appropriate for different kinds of data and problem specification. Of the 'standard' methods of multivariate analysis PCA is the easiest to understand since it is based on Euclidean distance which we are familiar with from everyday experience.

We can judge distances between points, and measure them exactly if needed. At the scale we normally operate on this is *Euclidean* distance. It can be defined mathematically in a way that generalizes to p-dimensions[4]. In p-dimensional space a point is defined by a set of values for the variables $(Y_1\ Y_2\ \ldots\ Y_p)$. After PCA the distances between cases are approximated by the distances between new variables, $(Z_1\ Z_2\ \ldots\ Z_r)$, where r is usually 2 or 3.

[4]The definition, of squared Euclidean distance, is

$$d_{ik}^2 = \sum_{j=1}^{p} (y_{ij} - y_{kj})^2$$

with the square-root of this giving d_{ik}.

7.4 Example 2 – Stone axe morphology

In this section ideas previously introduced are elaborated on, along with the practicalities of application and interpretation. The data used, from O'Hare (1990), are dimensional meaurements on 11 variables for 181 Neolithic polished stone axes from southern Italy, classified into three types according to their butt shape – pointed, rounded or square. The data are a subset of a larger sample of 209 axes, two small groups of intermediate butt types having been omitted for the purposes of our analyses. The full data set, with a definition of the variables, is given in Tables B.9 to B.11 in Appendix B.

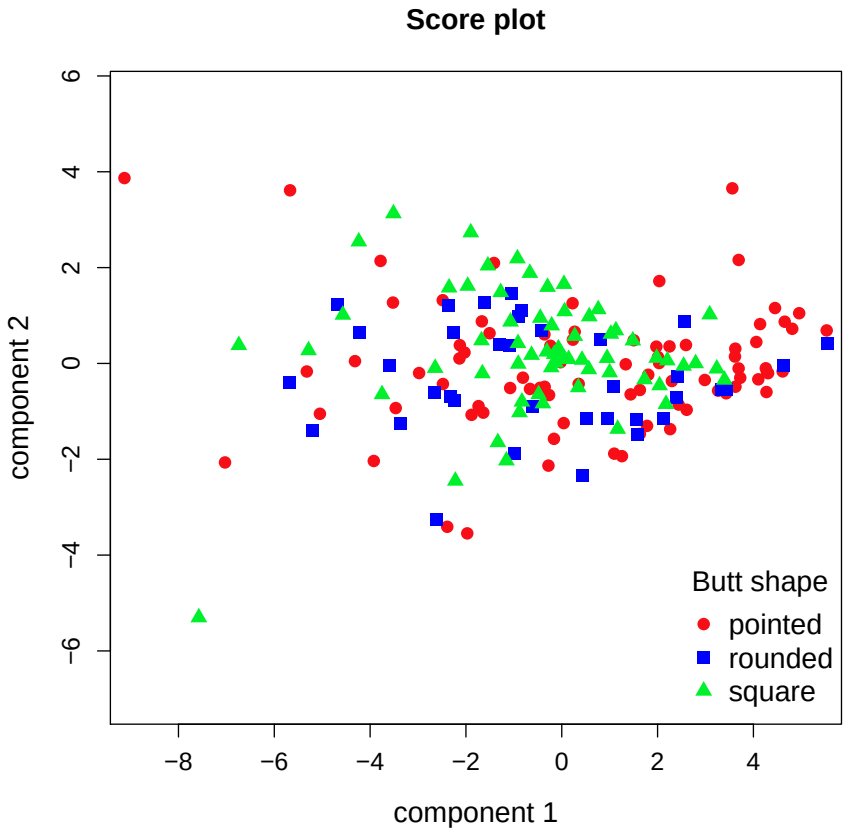

Figure 7.4: *A score plot for components 1 and 2 from a PCA of the standardized stone axe data labeled by butt shape.*

One of the research questions was whether the dimensional data revealed patterns that could be associated with butt type. One of the analytical tools used was PCA and this is emulated here, using butt type for labeling and interpretive

purposes only. An example is provided first, before spelling out some of the detail. Figure 7.4 plots scores on the first two components from a PCA of the standardized data. There is no obvious grouping in the data and no outliers to cause undue concern.

7.4.1 Definition and properties of principal components

For a more detailed account of the notation introduced immediately below see Appendix D. The $n \times p$ data matrix $\mathbf{Y}$ has typical element y_{ij}, for $j = 1, \ldots p$ and $i = 1, \ldots, n$; principal components (PCs) are *defined* to be linear combinations of the form

$$Z_j = a_{j1}Y_1 + a_{j2}Y_2 + \ldots + a_{jp}Y_p$$

for component j.

The a_{ji} are *coefficients* to be determined[5]. Given the a_{ji} principal component *scores*, z_{ij}, held in the matrix $\mathbf{Z}$ with the same dimensions as $\mathbf{Y}$, can be calculated.

Principal components are *defined* by the following criteria.

(a) The components are uncorrelated.

(b) The first component, Z_1, has maximum variance; subject to the lack of correlation Z_2 has the second largest variance; and so on. The variances are called *eigenvalues* in some software and will be denoted by λ_i.

(c) There is a complication in that the variances are unbounded unless a constraint is imposed on the coefficients. Often this has the form

$$a_{j1}^2 + a_{j2}^2 + \ldots + a_{jp}^2 = 1 \tag{7.1}$$

but

$$a_{j1}^2 + a_{j2}^2 + \ldots + a_{jp}^2 = \lambda_j. \tag{7.2}$$

is also used. Constraint (7.1) is usual in software where PCA and factor analysis are clearly distinguished; constraint (7.2) is usual in implementations of factor analysis[6].

Given these conditions we let the mathematics do the work of obtaining the 'solution' to the problem (of determining the a_{ji}) (Appendix D). Computational aspects are embedded in `R` functions such as `prcomp`, the fine detail of which the average user may remain blissfully unaware.

[5] Also sometimes called *loadings*, a term more commonly used in factor analysis (Chapter 8).

[6] In the widely-used `SPSS` package PCA is treated as a special case of factor analysis. This has led to a lot of confusion among users and will be discussed further in Chapter 8.

The idea behind (a) is that it can be easier to work with uncorrelated variables. It should be emphasized that all that is involved is a *mathematical transformation* of the data. There is no requirement that the components be interpretable other than as a linear combination of the original variables, though they often will be. That is, assignment of a 'meaning' to components is not a fundamental issue. This is a source of confusion between PCA and factor analysis where the definition of 'meaningful' factors is a central concern (see Chapter 8).

The idea behind (b) is that the components with the larger variances are likely to be *structure carrying* in the sense that plots based on them will reveal patterns in the data, if they exist. This idea is empirically rather than theoretically based but it frequently 'works'.

The correlation diagram in Figure 6.6 shows that there are generally high correlations among the variables. This leads to the expectation that PCA will be an effective dimension-reduction method. Furthermore, all the correlations are positive, which leads to the expectation that the coefficients for the first PC will be of the same sign and similar order of magnitude[7]. This has an interpretation as a '*size*' component, literally in the present example. Components with a mixture of signs among the coefficients can be interpreted as '*shape*' components, further interpretability concerning aspects of shape depending on the context.

The importance of a variable in defining a component depends on the value of $|a_{ji}|$. Where this, or its square, is close to 1 (if constraint (7.1) is used) the component is effectively the same as variable i. Various strategies exist for aiding interpretation, for example by ignoring coefficients for which $|a_{ji}|$ is less than some predetermined value (Section 8.3). It can help if a coefficient is either 'large' or 'close' to zero[8]. This can be achieved more formally, and mathematically, by subjecting components to *rotation*, more common in applications of factor analysis than PCA and discussed further in Sections 8.1 and D.3.2.

7.4.2 Interrogating PCA output

To interrogate numerical information the PCA of standardized data is first undertaken using

```
axe.pca <- prcomp(stoneaxe135.data, scale = T)
```

where `stoneaxe135.data` is the subset of Tables B.9 to B.11 containing the three butt types under investigation. Component scores are held in `axe.pca$x` and

[7]The signs of the PCs are arbitrary so if, for example, all signs are positive or all negative the interpretation is unaffected.

[8]This is not essential; 'size' components are readily interpretable without satisfying this criterion.

coefficients in `axe.pca$rotation`. The former are used as the basis for the plot of Figure 7.4. The latter can be viewed to a sensible number of significant digits using `round(axe.pca$rotation, 1)` with the result shown in Table 7.3.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
L1	-0.3	0.2	0.2	-0.1	0.4	0.4	-0.6	0	-0.3	0	0
L2	-0.2	0.4	0.5	-0.4	-0.3	0.2	0.4	0.1	0	-0.1	0
B1	-0.3	-0.3	0	-0.2	0.1	0.1	0.2	-0.1	0	0.4	-0.7
B2	-0.3	-0.2	0	-0.4	-0.5	-0.4	-0.5	-0.1	0.1	-0.1	0
B3	-0.3	-0.3	-0.1	-0.2	0.1	0.1	0.2	0	0	0.4	0.7
WC	-0.3	-0.4	-0.1	-0.1	0.2	0.1	0.3	0.2	0	-0.7	0
DC	-0.2	-0.3	0.6	0.7	-0.2	0	0	0	0	0.1	0
TH	-0.3	0.3	-0.3	0.2	-0.1	0.1	0.1	-0.8	0	-0.2	0
L3	-0.3	0.2	0.2	0	0.6	-0.6	0.1	0	0.2	0	0
T1	-0.3	0.3	-0.3	0.2	-0.2	-0.2	0.2	0.4	-0.6	0.1	0
T2	-0.3	0.2	-0.3	0.2	-0.1	0.3	-0.1	0.4	0.7	0.1	0

Table 7.3: *PC coefficients, rounded to one decimal place, from the PCA of the stone axe data.*

As expected, the coefficients for the first component have the same sign and similar magnitude and can be interpreted as a size component – essentially it averages the standardized measurements of all the variables. The second component is a shape component that contrasts the length and thickness variables with the breadth and cutting-edge variables. The pattern is displayed in a readily appreciated form in the variable plot for the first two components, in the left-hand plot of Figure 7.5.

Size may or may not be of intrinsic interest; for the axe data the focus in O'Hare (1990) was on typology as revealed by shape, so size is of less interest and the plot based on the second and third components in Figure 7.5 is of potentially greater interest. It shows three distinct clusters based on length, thickness and breadth variables, the last of these also associated with the width of the cutting-edge. Depth of cutting-edge (`DC`) is isolated from the other variables, and also dominates the fourth PC. The graphs tell the same story as those in O'Hare (1990).

Output concerning the 'importance' of the PCs can be investigated in several ways. Commonly in software packages the variances (eigenvalues), both individual and cumulative, are presented. In `R`, if `prcomp` is used, the standard deviations, as opposed to variances, are stored in `axe.pca$sd`. These can be manipulated to produce Table 7.4, emulating what is to be seen in other software.

Section 6.1 of Jolliffe (2002) discusses a large number of criteria that have been used for selection; the most commonly used, and the only ones considered here,

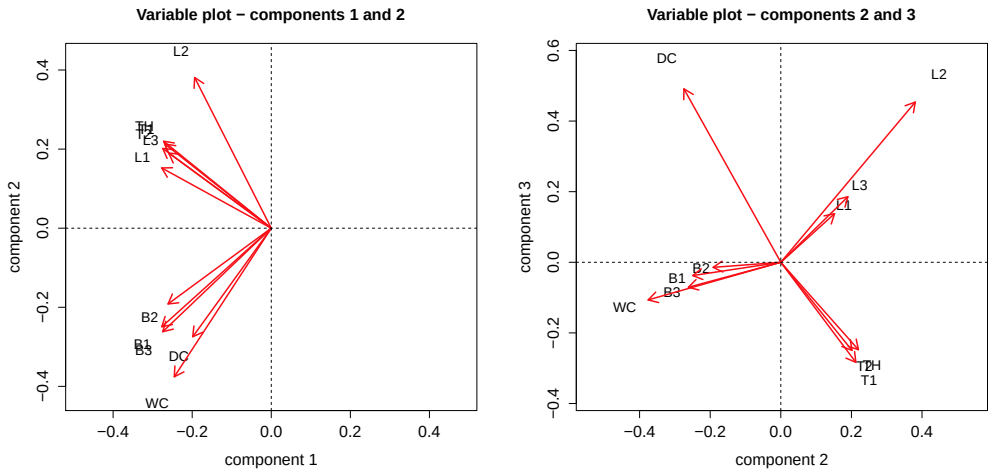

Figure 7.5: *Coefficient plots for components 1 and 2, and components 2 and 3, from a PCA of the stone axe data.*

Component	1	2	3	4	5	6	7	8	9	10	11
St.Dev.	2.8	1.3	0.7	0.7	0.6	0.4	0.2	0.2	0.1	0.1	0
Variance (%)	71.0	14.6	4.7	4.4	2.8	1.3	0.6	0.2	0.2	0.1	0
Cumulative (%)	71.0	85.6	90.3	94.7	97.5	98.9	99.4	99.7	99.9	100	100

Table 7.4: *Standard deviations, variances (%) and cumulative variances (%) for the PCs for the stone axe data.*

are described as *ad-hoc*. Jolliffe (2002: 112) comments that they are 'intuitively plausible' and 'work in practice'. One simple rule is to require that the cumulative percentage of variance accounted for by the components exceeds some threshold. Thus, 80% would lead to a choice of two components here. Choice of the threshold is arbitrary; 70% is sometimes mentioned as a reasonable lower limit, but I have seen examples (mostly archaeometric) where the first two components only account for 50–60% of the variance, but are useful. Another common criterion is to require the variances (when standardized data are used) to exceed some value such as 1 (Kaiser's rule) or 0.7. Both lead to a choice of two components here.

The information in Table 7.4 can be represented in a *scree plot*, the `R` version of which is shown in the left-hand plot of Figure 7.6. The idea is to identify the point at which the plot 'levels off' or, as it is sometimes expressed, an 'elbow' is evident. Once again two components are suggested. Such plots can be quite difficult to interpret; roughly, it is not unusual for them to exhibit something looking like 'exponential decay' so an elbow is not apparent.

Given that the size component is of limited interest in O'Hare (1990) it might be ignored and the scree plot rescaled, as in the right-hand plot in Figure 7.6. It

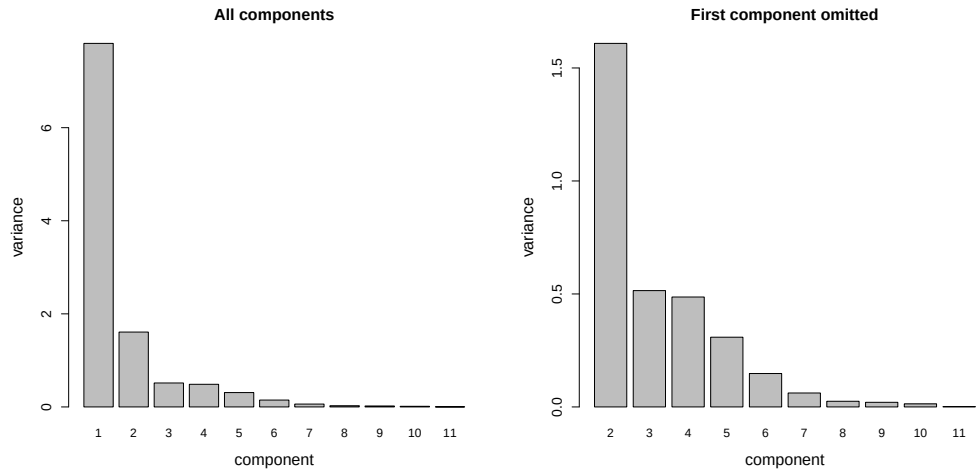

Figure 7.6: *Scree plots for the PCA of the stone axe data, with and without the first component.*

can now be seen that the second component is clearly dominant, with the next two or three contributing information.

Some of the rules just described, as Jolliffe notes, derive from studies in factor analysis, where the 'correct' choice of the number of factors can be critical. The concept of a 'correct' choice in PCA, if it has a meaning, has possibly been exaggerated in importance. My reason for thinking this is that, without committing oneself to a choice, an inspection of pairs plots of all the PCs that seem useful is perfectly possible – those that are useful being subject to more detailed scrutiny. It can happen that, for example, plots involving the fourth component reveal useful information whereas those based on the third component do not. Also, analysis often proceeds iteratively, with outliers removed and analysis repeated, for instance. Under these circumstances attempting to select a 'correct' number of components seems pointless.

Figure 7.7 is a pairs plot based on PCs 2, 3 and 4, using the version available in the `car` package. The plots are best viewed in color but, however approached, can be 'messy'. Concentrating on the plot for the second and third components, there is no discernible separation of types. Pointed axes are perhaps more evident on the periphery of the plot, but are all over the place, and are the most numerous class. Similar observations can be made about the other plots.

Given the obvious overlap of butt-types, confidence ellipses or convex hulls (Figures 6.3 and 6.5) will be of no value for separating types. Experimenting with contouring is an idea since denser regions at the centers of the scatter for each type might separate, but this did not happen and the plots are not shown. The

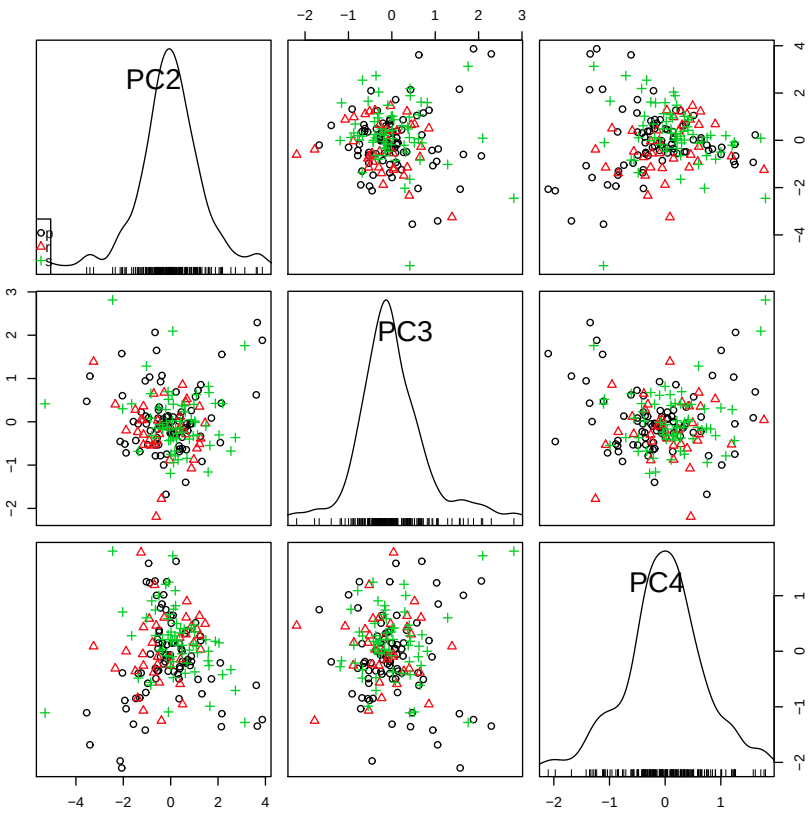

Figure 7.7: *Pairs plot for components 2-4 from a PCA of the stone axe data.*

main outcome of this analysis has been to show that there are clear patterns of correlation in the variables, but this is not reflected in any structure in the score plots related to butt type. Analysis continues in Section 8.3 where the idea of rotation of components is illustrated.

7.5 R notes

Figures 7.1 to 7.6

Some of the presentational arguments and legend have been omitted. Data for the oxides are held in `oxides` and `site` respectively. The colors (`Col`) and plotting characters (`Sym`) for the site were created separately and the relevant code is not shown. The following basic code is for Figure 7.1.

```
library(MASS)
```

```
pc0 <- prcomp(oxides, scale = T)
x1 <- pc0$x[ , 1]; x2 <- pc0$x[ , 2]
y1 <- pc0$rotation[ , 1]; y2 <- pc0$rotation[ , 2]

# Score plot
eqscplot(x1, x2, main = "Standardized data")
abline(h = 0); abline(v = 0)

# Coefficient plot
eqscplot(y1, y2, type = "n")
text(y1, y2, names(oxides))
arrows (0, 0, y1 * .85, y2 * .85, code = 2, length = .15)
abline(h = 0); abline(v = 0)
```

For Figure 7.1 define `pc0 <- prcomp(log10(oxides), scale = F)`; for Figure 7.3 replace `log10(oxides)` with `LR` where `LR` is the (centered) log-ratio transformation that can be obtained from the `clr` function in the `Hotelling` package

```
library(Hotelling)
    Si <- 100 - apply(oxides, 1, sum)
    oxidesSi <- as.data.frame(cbind(oxides, Si))
    LR <- clr(oxidesSi)
```

where `Si` is the 'residual', which can be equated with silica, that is used to augment the data set so that it is fully compositional.

Other than the data used, Figures 7.4 and 7.5 introduce nothing new. The scree plot to the left of Figure 7.6 might be obtained using the `screeplot` function, `screeplot(PCA, xlab = "component")`, where `PCA` is the object obtained on using `prcomp` in the first stage of analysis. This draws on the `barplot` function, which was used directly here in order to obtain the plot to the right of Figure 7.6 with `barplot(PCA$sd^2, names.arg = 1:11)` producing the plot to the left and `barplot(PCA$sd[2:11]^2, names.arg = 1:11)` that to the right.