29 Principal component analysis applied to harbour porpoise fatty acid data

Jolliffe, I.T., Learmonth, J.A., Pierce, G.J., Santos, M.B., Trendafilov, N., Zuur, A.F., Ieno, E.N. and Smith, G.M.

29.1 Introduction

In this chapter we apply principal component analysis (PCA) to data on blubber fatty acid composition in harbour porpoises. Various decisions need to be made when using PCA. As well as showing the usefulness of the dimension reduction achieved by PCA for these data, the implications of these decisions will also be illustrated. The next two sections of the chapter describe the data and the statistical technique respectively. A data exploration section is followed by the main section (Section 29.5), which describes and discusses the results of PCA for the data. Interpretation of principal components can sometimes be difficult and several methods have been suggested for simplifying interpretation. These will be discussed in Section 29.6, and one of them will be illustrated on the fatty acid data. The chapter is completed by a short discussion section.

29.2 The data

The harbour porpoise (Figure 29.1) is probably the most abundant cetacean in British waters (Hammond et al. 2002). However, harbour porpoises, as most cetaceans, are subject to various threats and pressures throughout their range, and increased concern for the status of harbour porpoises has led to the need for more information on the species, including their diet.

Diet is an important aspect of the ecology of marine mammals. Changes in prey type or availability have the potential to affect the distribution, body condition, susceptibility to disease, exposure to contaminants, reproductive success and, ultimately, survival of most marine mammals, including harbour porpoises.

Figure 29.1. Harbour porpoises off Scotland.

Traditional methods used to estimate diets in marine mammals, such as stomach contents analysis, are often limited and estimates can be biased. Fatty acids in predator body tissues have led to the use of fatty acid analysis as a method for understanding the diet of marine mammals. Fatty acids have the potential to act as tracers of diet, with the fatty acid composition of tissues reflecting the average diet ingested over a period of days or months. The influence of prey fatty acid signatures on predator fatty acid profiles has been clearly shown in captive feeding experiments of fish, squid and seals (Kirsch et al. 1998, 2000; Stowasser et al. 2006). Fatty acids in the blubber of different species of free-living marine mammals have also been shown to reflect their diet (for example, Brown et al. 1999; Hooker et al. 2001; Iverson et al. 1995, 1997; Smith et al. 1997; Walton et al. 2000).

However, in addition to diet, other factors also have the potential to influence the fatty acid composition of blubber. Therefore, prior to using fatty acid analysis of blubber samples from stranded porpoises to examine diet, it is important to determine what factors, other than diet, may influence the fatty acid profile. These factors may include decomposition state and blubber thickness.

Blubber samples from 89 harbour porpoises stranded around the Scottish coast between 2001 and 2003 were used for fatty acid analysis, and they form the basis of the present data set. All tissue samples were removed from the left side in front of the dorsal fin during the postmortem examinations by the vets at the Scottish Agricultural College, Inverness. During the postmortem examinations, data were collected on sex, total body length, weight, girth, etc. but these (explanatory) variables were not used in the current analysis because, based on stomach contents analysis, porpoise diet is known to vary in relation to body size (Santos et al. 2004).

Lipids were extracted from the inner blubber layer, and individual fatty acids were identified. The normalised area percentage was calculated for 31 fatty acids: 12:0, 14:0, 14:1n-5, 15:0, 16:0, 16:1n-7, 16:2n-6, 16:3n-6, 16:4n-3, 18:0, 18:1n-9,

18:1n-7, 18:2n-6, 18:3n-6, 18:3n-3, 18:4n-3, 20:0, 20:1n-11, 20:1n-9, 20:2n-6, 20:4n-6, 20:3n-3, 20:4n-3, 20:5n-3, 22:0, 22:1n-11, 22:1n-9, 21:5n-3, 22:5n-3, 22:6n-3 and 24:1n-9. Full details of the lipid extraction method are given in Learmonth (2006).

29.3 Principal component analysis

Principal component analysis was described in Chapter 12. It is a dimension reducing technique, which replaces the original variables $Y_1, Y_2, \ldots, Y_N$ by a smaller number of linear combinations of those variables, while keeping as much as possible of the variation in the original variables.

Let Y_{ij} be the value of variable j ($j = 1,..,N$) for observation i ($i = 1,..,M$). Then the value of the k^{th} principal component (PC) for the i^{th} observation is given by

$$Z_{ik} = c_{k1} Y_{i1} + c_{k2} Y_{i2} + \ldots + c_{kN} Y_{iN}$$

For the current data, j denotes the j^{th} fatty acid, i is the i^{th} porpoise, $N = 31$ and $M = 89$. The numbers c_{1j} ($j = 1,..,N$) are chosen so that the first PC accounts for as much of the variance in the original variables as possible, the numbers c_{2j} ($j = 1,..,N$) are chosen so that the second PC accounts for as much of the variation as possible in the original variables, subject to the constraint that the second PC is uncorrelated with the first. Third, fourth, ... PCs can be similarly constructed. In theory as many as N PCs can be constructed but, in practice, if the first few account for most of the variation in the original variables, then often only those PCs are of interest.

Various decisions need to be made in conducting a PCA. The first decision is whether to base the PCA on a covariance matrix or correlation matrix. Both analyses will be presented and discussed below. It is fair to say that in most circumstances a correlation-based approach is more appropriate, but there are occasions when a covariance-based analysis is suitable; see Jolliffe (2002, Section 3.3) for more discussion.

Another decision is how many PCs are needed to adequately represent the variation in the original data. Some methods for making this decision are discussed in Chapter 12 and will be illustrated in this chapter. For more details and other methods, see Jolliffe (2002, Chapter 6).

A third decision concerns the values of the c_{kj}. It is their relative values for a particular PC that determines the nature of that component, but for detailed interpretation of a PC it is necessary to decide on a particular 'normalization' of the c_{kj}. The two main normalizations are

$$\sum_{j=1}^{N} c_{kj}^2 = 1 \quad \text{or} \quad \sum_{j=1}^{N} c_{kj}^2 = \text{var}(Z_k)$$

where $\text{var}(Z_k)$ is the variance of the k^{th} PC. Each of these normalizations will be illustrated and explained below.

29.4 Data exploration

Boxplot and Cleveland dotplots (Chapter 4) were generated for the 31 fatty acids to identify any extreme values and to determine whether the data required transformation. The boxplots (Figure 29.2) and dotplots (not shown here) indicate a few extreme values, for example, fatty acid 18:1n-7. However, transformation of the data was not required for the fatty acids variables.

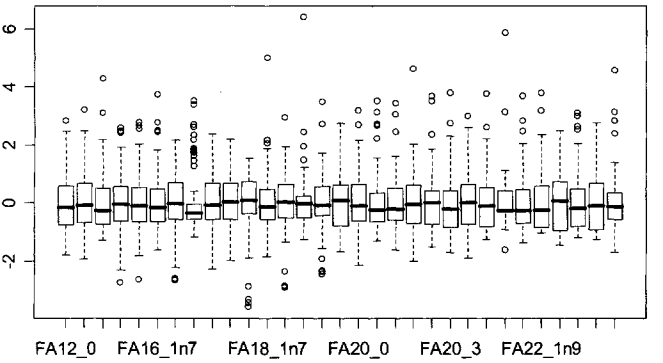

Figure 29.2. Boxplot of 31 fatty acids from the inner blubber layer of 89 harbour porpoises. The data were normalized so that all boxplots are on the same scale. Not all 31 labels are plotted.

29.5 Principal component analysis results

Covariance-based PCA

The first PC accounts for 71% of the total variation in the original variables and the second PC for a further 17%, so that a total of 88% of the variation in the original 31 dimensions can be represented by just the two dimensions defined by the first two PCs. Furthermore, the first two PCs look easy to interpret. Although they are linear combinations of all 31 variables, most of the constants c_{1j} and c_{2j} (the loadings) are small. The first PC is $Z_1 = 0.81Y_6 + 0.41Y_{30} + \ldots$, where none of the other 29 terms in the linear combination has a loading greater than 0.21, and only six of them exceed 0.10. Similarly the second PC is $Z_2 = 0.78Y_{11} - 0.49Y_{30} + \ldots$, with no other loadings greater than 0.21 and only four greater than 0.10.

However, the massive reduction in dimensionality reduction and easy interpretation of the first two PCs is not as impressive as it might seem. It is caused by large differences in the variances of the original 31 variables. These range from 83.8 for Y_6 (16:1n-7) to 0.00154 for Y_{17} (20:0). When there are big discrepancies between variances, a covariance-based PCA will often tell you little more than which variables have the largest variances. It is no co-incidence that the three variables that dominate the first two PCs are those with the three largest variances. A correlation-based PCA is far more appropriate for these data.

Correlation-based PCA

Table 29.1 provides the variances of, and cumulative proportion of variance accounted for by the first 10 PCs, and Figure 29.3 plots the variances (also known as eigenvalues; see Chapter 12).

Table 29.1. Eigenvalues and eigenvalues expressed as a cumulative proportion.

Axis	Eigenvalue	Cumulative Proportion	Axis	Eigenvalue	Cumulative Proportion
1	13.04	0.42	6	1.28	0.76
2	4.00	0.55	7	0.98	0.79
3	2.23	0.62	8	0.96	0.82
4	1.75	0.68	9	0.82	0.85
5	1.32	0.72	10	0.74	0.87

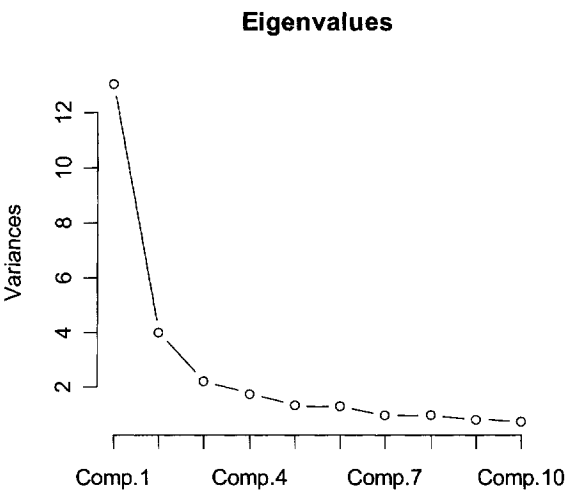

Figure 29.3. First 10 eigenvalues (variances) obtained by the PCA based on the correlation matrix.

There are various proposed rules for deciding how many PCs are needed to adequately represent the variation in the original variables; see Jolliffe (2002, Chapter 6). Some of the simpler ones are described in Chapter 12 of the current text. One method is to plot eigenvalues (variances) against PC number as in Figure 29.3, giving a so-called scree plot, and look for an 'elbow' in the plot. Deciding the location of the 'elbow' is subjective. Here, component 3 probably gives the most plausible 'elbow', implying that only two PCs need to be retained, but other readers looking at the plot might disagree.

A second rule is to retain the minimum number of components needed to account for a certain (high) percentage of the total variation. A common choice for the threshold is 80%, and it can be seen from Table 29.1 that it needs 8 PCs to achieve this. A third popular rule for correlation-based PCA is to retain all PCs whose variances exceed 1, the average variance of PCs for a correlation-based analysis. This rule implies that six PCs are needed.

The three rules are contradictory. The suggestion made after visual examination of the scree plot is almost certainly too small. Two PCs only account for 55% of the total variation, which is distinctly unimpressive. The six PCs suggested by looking at sizes of individual eigenvalues account for 76% of the variation, and this might be acceptably close to 80%.

It is clear that six or more PCs are needed if a substantial part of the original variation is not to be lost. Having said that, what follows concentrates on describing and interpreting the first two components.

The columns of Table 29.2 labelled PC1 and PC2 give the loadings for the first two PCs. The columns labelled SC1, SC2 will be explained in Section 29.6.

From Table 29.2, the first PC is

$$Z_1 = -0.22Y_1 - 0.24Y_2 - 0.23Y_3 - \ldots + 0.25Y_{30} + 0.07Y_{31}$$

where the normalization $\sum_{j=1}^{N} c_{kj}^2 = 1$ is used for the loadings, and var(Z_1) = 13.04. If the normalization $\sum_{j=1}^{N} c_{kj}^2 = \text{var}(Z_k)$ had been used instead, the first PC would be

$$Z_1 = -0.83Y_1 - 0.88Y_2 - 0.85Y_3 - \ldots + 0.93Y_{30} + 0.24Y_{31}$$

This PC is obtained from the first one simply by multiplying it by a constant, the standard deviation of the original Z_1, so that a plot of the values of the observations for a given PC (the PC scores) looks exactly the same for the two normalizations, apart from relabelling the axis to reflect the change of scale. Both normalizations are in common use, and you may see either, depending on which computer software produces your PCs.

The first normalization is fundamental to the derivation of PCs, but the second has the advantage that, for a correlation-based PCA, the loadings are equal to the correlation between the component and each variable. Hence, PC1 for the current data has correlation –0.83 with Y_1, correlation –0.88 with Y_2, and so on.

A third normalization is also sometimes implicitly used. When plotting PC scores, they are sometimes normalized to have unit variance for each PC. This

implies that the loadings in the first normalization are *divided* by the standard deviation of the PC (to get the second normalization the loadings in the first were *multiplied* by this standard deviation).

Table 29.2. Loadings for the first two PCs obtained by the PCA based on the correlation matrix (denoted PC1, PC2), and for the first two SCoTLASS components (denoted SC1, SC2) described in Section 29.6.

Variable	PC1	PC2	SC1	SC2	Variable	PC1	PC2	SC1	SC2
Y_1 12:0	−0.23	0.13	−0.15	0.00	Y_{17} 20:0	0.18	0.16	0.00	−0.35
Y_2 14:0	−0.24	0.13	−0.36	0.00	Y_{18} 20:1n-11	0.15	0.22	0.00	−0.46
Y_3 14:1n-5	−0.23	−0.17	−0.17	0.05	Y_{19} 20:1n-9	0.18	−0.18	0.00	0.01
Y_4 15:0	−0.16	0.27	0.00	0.00	Y_{20} 20:2n-6	0.20	−0.14	0.00	−0.12
Y_5 16:0	−0.05	−0.08	0.00	0.00	Y_{21} 20:4n-6	0.23	−0.16	0.29	0.00
Y_6 16:1n-7	−0.24	−0.21	−0.13	0.22	Y_{22} 20:3n-3	0.16	−0.09	0.00	−0.03
Y_7 16:2n-6	−0.08	0.28	0.00	0.00	Y_{23} 20:4n-3	0.25	0.08	0.23	0.00
Y_8 16:3n-6	−0.01	0.05	0.00	0.00	Y_{24} 20:5n-3	0.24	−0.13	0.37	0.00
Y_9 16:4n-3	0.16	0.13	0.00	−0.17	Y_{25} 22:0	0.11	0.06	0.00	−0.05
Y_{10} 18:0	0.24	0.02	0.26	0.00	Y_{26} 22:1n-11	0.12	0.30	0.00	−0.44
Y_{11} 18:1n-9	0.03	0.28	0.00	0.00	Y_{27} 22:1n-9	0.13	0.07	0.00	−0.14
Y_{12} 18:1n-7	0.16	−0.20	0.00	0.00	Y_{28} 21:5n-3	0.25	−0.01	0.29	0.00
Y_{13} 18:2n-6	0.07	0.34	0.00	−0.22	Y_{29} 22:5n-3	0.24	−0.09	0.39	0.00
Y_{14} 18:3n-6	0.16	−0.07	0.00	0.00	Y_{30} 22:6n-3	0.25	−0.08	0.47	0.08
Y_{15} 18:3n-3	0.09	0.33	0.00	−0.30	Y_{31} 24:1n-9	0.07	0.23	0.00	−0.03
Y_{16} 18:4n-3	0.21	0.16	0.00	−0.45					

Turning to the interpretation of the PCs in this example, the loadings of most of the variables are positive, but those for the first 8, in particular Y_1 – Y_4, Y_6, are negative. This implies that the main source of variation in the data is between those porpoises that have greater than average values for the first group of variables coupled with smaller than average values for the remaining variables and those porpoises with the opposite features. Looking at the second PC, the largest positive loadings are for Y_4, Y_7, Y_{11}, Y_{13}, Y_{15}, Y_{18}, Y_{26}, Y_{31}, whereas the largest negative loadings are for Y_6, Y_{12}. This means that the main source of variation that is uncorrelated with the first PC contrasts those porpoises with greater than average values for the first group of variables, and smaller than average values for the second group, with porpoises having the opposite features.

Using knowledge of the nature of the variables, PC1 appears to order fatty acids in relation to carbon chain length, with the shortest chain fatty acids having the most negative correlations with PC1, and vice versa. It may also be noted that short-chain fatty acids are biosynthesised in marine mammals, whereas most long-

chain fatty acids are derived from the diet. There appears to be no simple interpretation of PC2, however. Interpretation could be attempted for all the PCs that are retained, but this is by no means straightforward for these data. Further discussion of interpretation will be given in Section 29.6.

Biplots

A complementary way of displaying results from a PCA is by means of a biplot. Such plots are discussed in Chapter 12. Figure 29.4 gives the so-called correlation biplot for the data. From this plot we can glean information about relationships between the variables, about relationships between porpoises, and the values of different variables for different porpoises.

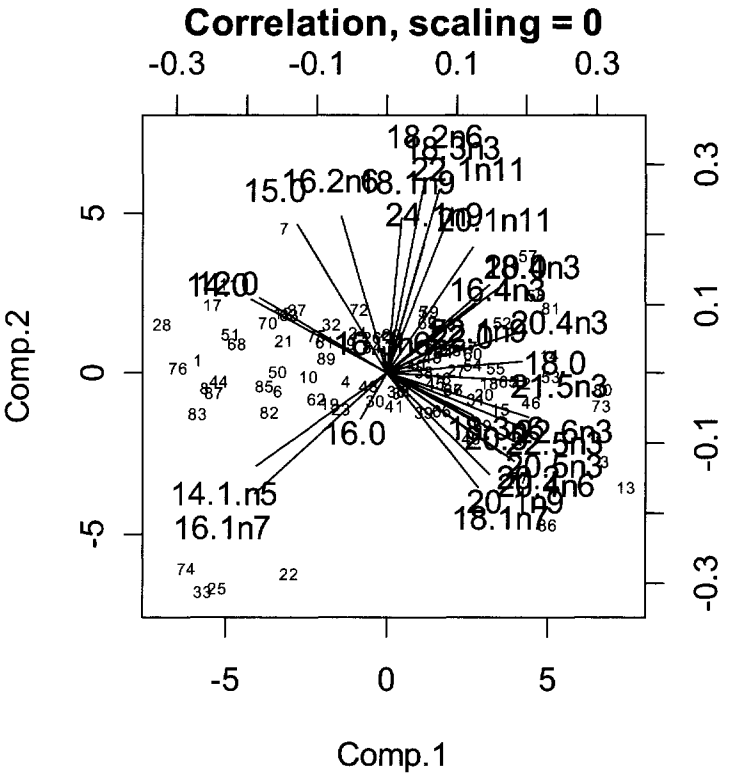

Figure 29.4. PCA biplot. Comp.1 and Comp.2 are the first two principal components.

The angle between lines corresponding to different variables gives an indication of the correlation between the variables (see Chapter 12). Lines pointing in similar directions correspond to variables that are positively correlated, lines pointing in opposite directions to negatively correlated variables, and lines roughly at right angles correspond to almost uncorrelated variables. For example, the lines corresponding to variables Y_3 (14:1n-5) and Y_6 (16:1n-7) point in almost the same direction, so the variables are likely to be highly correlated. Similarly, the pair of variables Y_1 (12:0), Y_2 (14:0) seem likely to be highly correlated with each other, but almost uncorrelated with Y_3, Y_6. The reason for saying 'likely to be' rather than 'are' highly correlated is that the two-dimensional plot only accounts for 55% of total variation in the data, the amount accounted for by the first two PCs. If the first two PCs, and hence the two-dimensional plot, had accounted for more, 90% say, of the variation, the relationship between the angles and the correlations would be more definite. The grouping of fatty acids in the PCA biplot appears to reflect chain length, with all eight of the fatty acids with negative values for PC1 having a carbon chain length of 16 or less and all fatty acids with positive values for PC2, except one Y_9 (16:4n-3), having a carbon chain length of 18 of more.

Turning to the observations, those porpoises that appear close together on the plot are likely to have similar values for the 31 variables. There are two caveats attached to this statement. The first is to note again that the plot displays only 55% of the variation in the data; the second is that the distances approximated on the plot are Mahalanobis distances, not Euclidean distances (see Chapter 12). Despite these caveats, there is still useful information to be gleaned from the plot. For example, porpoises 22, 25, 33, 74 (which had all died due to neonatal death) seem to form a group similar to each other, but separate from the rest, whereas porpoise 7, which had the largest body length out of all the porpoises sampled, seems to be rather different from any of the others.

Finally, the reason for the 'bi' in biplot is that the loadings and scores can be looked at simultaneously. If a line (L_1) is drawn from an observation to the line corresponding to a variable (L_2) so that it intersects L_2 at right angles, the position of the intersection provides information about the value of the variable for that observation. If the intersection with L_1 is a long way from the origin, it suggests that the value of the variable is larger than average for the observation, whereas if the intersection is close to the origin, the variable probably has a fairly average value for that observation. If the line L_2 needs to be extended through the origin in order for L_1 to intersect it at right angles, then the value of the variable is likely to be smaller than average for the observation. For example, porpoise 7 is likely to have larger than average values for variables Y_1, Y_2, Y_4, Y_7, near average values for Y_3, Y_5, Y_6, and below average values for Y_{12}, Y_{19}, Y_{20}, Y_{21} among others. Again the word 'likely' is inserted because the plot represents only 55% of total variation. In fact, examining the original data, porpoise 7 is in the upper quartile for variables Y_4, Y_1, Y_2, in the lower quartile for Y_{19}, Y_{20}, Y_{21}, and in the middle half of values for Y_5, in line with predictions from the biplot. For the other variables mentioned, the predictions are less good, but would be much better if the biplot had represented 90%, say, of the total variation.

29.6 Simpler alternatives to PCA

The linear combinations of variables defined by the PCs are optimal in the sense that they successively account for as much as possible of the total variation in the data. However, they can be difficult to interpret. For example, the loadings of the first two PCs given in Table 29.1 have a wide range of values, which makes them difficult to interpret. In component 2, variables Y_{13} and Y_{15} with loadings 0.34 and 0.33 make a large contribution to the component whereas Y_8 and Y_{10} with loadings 0.05, 0.02 have trivial contributions. But what about Y_1, Y_2? Are their loadings (both 0.13) large enough to consider when interpreting what the component represents?

Although some caution is needed in taking the size of loadings as a definite indication of the importance of variables in a component (Cadima and Jolliffe 1995), it would nevertheless be much easier to interpret a component that had most of its loadings unambiguously large or small with few intermediate values. Several techniques have been developed that make components simpler in this sense. Of course, to achieve this it is necessary to sacrifice something. Simpler alternatives to principal components will typically account for less of the total variation than the PCs and/or they may lose the property of being uncorrelated.

A review of some simpler alternatives to PCA is given in Chapter 11 of Jolliffe (2002), but this is an active area of research and new methods are still in development. The existing methods can be divided into four broad categories:

- Rotation. This is probably the best-known and most popular strategy, and it is borrowed from the related technique of factor analysis (Jolliffe 2002, Chapter 7). The idea is that a decision is made on how many components to retain, perhaps 6 for the current data. This defines a six-dimensional subspace of the 31-dimensional space spanned by the complete data set. Rotation of the axes is then carried out in the six-dimensional space in a way that optimises some 'simplicity' criterion.
- Do a PCA, then simplify the PCs in some way, often by severe rounding of the loadings.
- Find linear combinations that optimise some criterion that simultaneously searches for large variance retention and simplicity.
- Find linear combinations of the variables that successively maximize variance but are subject to additional constraints designed to achieve greater simplicity.

Here, just one technique from the final category will be briefly described and illustrated. As with some other techniques from this category, it can also be formulated as a method in the third category.

The idea for the technique is borrowed from multiple regression, where collinearity between variables can cause difficulties in interpreting a regression equation (Chapter 5). Tibshirani (1996) suggested a technique called the LASSO (Least Absolute Shrinkage and Selection Operator) that addresses this problem. It adds an extra constraint to the usual least squares method of fitting a regression equation.

Jolliffe et al. (2003) adapted the LASSO to the PC context and called the technique SCoTLASS (Simple Components LASSO). As noted in Section 3, PCA finds linear combinations Z_1, Z_2, ... of the variables that successively maximize Var(Z_k) subject to

$$\sum_{j=1}^{N} c_{kj}^2 = 1$$

and subject also to Z_k being uncorrelated with previous Zs. Here c_{kj} is the loading of the j^{th} variable for the k^{th} component. In SCoTLASS an extra constraint is added, namely

$$\sum_{j=1}^{N} |c_{kj}| \leq t,$$

where t is some threshold, which can lie in the range $1 \leq t \leq \sqrt{N}$, and N is the number of variables. For $t = \sqrt{N}$ the method simply gives the standard principal components, whereas for $t = 1$ it chooses the original variables according to the magnitude of their variances. As t decreases within the range, the components found by SCoTLASS become increasingly simple, with more and more loadings driven towards zero by the extra constraint. At the same time the variance accounted for by the first few components decreases compared with that accounted for in PCA. This behaviour is illustrated in Figure 29.5, which plots the amount of variance retained by the first six SCoTLASS components and the number of zero loadings (to two decimal places) in those components, as t varies.

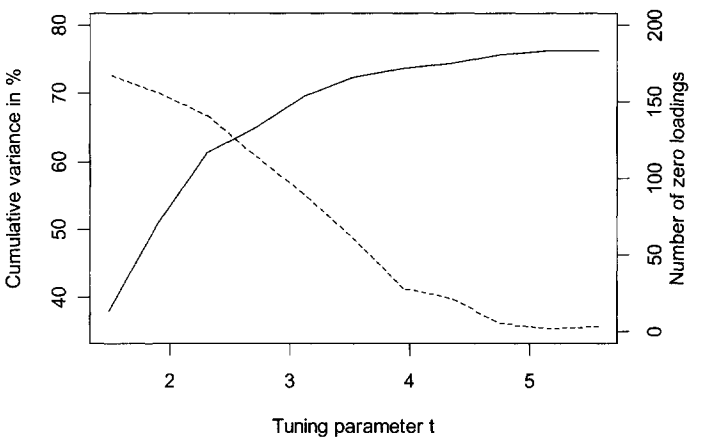

Figure 29.5. For SCoTLASS, the cumulative variance accounted for by the first six components (solid line) and the number of zero loadings in those six components (dotted line) as the tuning parameter varies from 1 to 5.57 (the latter gives PCA).

For example, for PCA (corresponding to t = $\sqrt{N}$ = 5.57) the first six components account for 76% of total variation but there are only three zero loadings. As t decreases to 4.35, 2.72, 1.91, the variance accounted for by six components drops to 74%, 65%, and 51%, respectively. At the same time, the number of zeros increases to 22, 114 and 156. The choice of t is subjective; from a plot such as Figure 29.5, the user can decide at what point the gain in simplicity is more than offset by the loss of variance.

A possible choice here is $t = 2.72$, and for this value, the loadings produced by SCoTLASS were included in Table 29.2. It can be seen that the numerous zeros in the loadings make the components potentially much simpler to interpret than the PCs. Component 1 is a contrast among variables Y_1, Y_2, Y_3, Y_6 (fatty acids 12:0‡, 14:0^{+}, 14:1n-5‡ and 16:1n-7^{+}, which are ‡primarily from biosynthesis or $^{+}$can be from both biosynthesis and diet, based on Iverson et al. 2004) on the one hand and Y_{10}, Y_{21}, Y_{23}, Y_{24}, Y_{28}, Y_{29}, Y_{30} (fatty acids 18:0^{+}, 20:4n-6*, 20:4n-3*, 20:5n-3*, 21:5n-3*, 22:5n-3^{+} and 22:6n-3*, which are *primarily from direct dietary intake or $^{+}$can be from both biosynthesis and diet, based on Iverson et al. 2004) on the other, with the remaining 20 variables apparently unimportant. The 'cost' of this simplification is a substantial reduction in variance accounted for by the first component, down from 42% to 20%. This is not as bad as it might seem. The reduction in variance for the first six components, from 76% to 65%, is much less, as some of the later SCoTLASS components account for more variance than the corresponding PCs. For example, for the second component, the amount of variance accounted for is 15%, more than the 13% associated with the second PC.

Turning to interpretation of the second component, there is a more straightforward interpretation than for the second PC. Here the second component is mainly measuring the group of variables Y_9, Y_{13}, Y_{15}, Y_{16}, Y_{17}, Y_{18}, Y_{20}, Y_{26}, Y_{27} (fatty acids 16:4n-3*, 18:2n-6*, 18:3n-3*, 18:4n-3*, 20:0^{+}, 20:1n-11*, 20:2n-6*, 22:1n-11* and 22:1n-9*, which are *primarily from direct dietary intake or $^{+}$can be from both biosynthesis and diet, based on Iverson et al. 2004), but contrasted with Y_6 (fatty acid 16:1n-7, which can be from both biosynthesis and diet). There are also a few variables with smaller, but non-zero, loadings whose contributions are equivocal. Fifteen of the 31 variables have zero loadings.

29.7 Discussion

In this chapter, PCA has been illustrated on an interesting data set, and various decisions that are needed in order to implement PCA have been discussed. An alternative to PCA that makes interpretation of the derived variables simpler has also been described and illustrated. More specifically, we looked at the relationship between fatty acids in the inner blubber layer of harbour porpoises stranded around Scotland. This study was used as a preliminary analysis of the fatty acid data prior to using fatty acid analysis as a method to examine diet.

A considerable reduction of dimensionality was achieved for the data set whilst still accounting for a large proportion of the original variation. The reduction was

less for correlation-based PCA than for covariance-based PCA, but the latter was clearly inappropriate for these data.

Biological interpretations were found based on some leading PCs and their simpler alternatives. The biplot associated with the PCA also provided useful information, even though the two-dimensional approximation to the data that it provides was not especially good for these data. There was some clear grouping of the fatty acids, which was consistent with known differences in the probable sources of these fatty acids. Long-chain mono- and polyunsaturated fatty acids that were primarily from direct dietary intake were generally separated from fatty acids that are primarily from biosynthesis. Of the fatty acids that can be from both the diet and the biosynthesis, fatty acids with a carbon chain length of 16 or less were grouped with fatty acids that were primarily from biosynthesis.

The four porpoises that had died due to neonatal death were clearly separated in the principal component analysis biplot. This probably reflects the fact that, in neonates, the blubber fatty acids have been obtained through maternal transfer and foetal synthesis prior to birth, as well as the transfer of fatty acids in milk. A study on the fatty acid profile of harbour porpoises from the mid-Atlantic coast of America revealed differences between maternal and foetal blubber, suggesting selective transfer of fatty acids to the foetus (Koopman 2001). Similar observations have also been made in seals, for example, by Iverson et al. (1997).

Although PCA is fundamentally a very simple idea, it is an extremely powerful tool, and new applications and modifications are still being developed. The idea of constructing alternatives to PCA that retain its main objectives while trying to simplify the results is just one manifestation of this. New developments are scattered through the literature of many subject areas. As well as mainstream statistics and ecology, computer science, data mining, genetics and psychology are among the areas of active research in such topics.

Finally, one particular modification of PCA should be mentioned. The data analysed in this chapter are 'compositional'; that is the sum of the variables is the same (100%) for every porpoise. Special techniques are available for such data (Jolliffe 2002, Section 13.3), although with as many as 31 variables, it is unlikely that conclusions would be changed much by using these methods.

Summarising, the data set studied proved to be unsuitable for covariance-based PCA, but correlation-based PCA revealed informative patterns. The results of this analysis suggest that the relationships between different fatty acids in the inner blubber layer of harbour porpoises, namely the tentative identification of several groups of fatty acids, could reflect differences in their origin, i.e., diet or biosynthesis.

Acknowledgements

This project would not have been possible without the samples, data, help and advice of Bob Reid and Tony Patterson at the SAC in Inverness and the help of various staff members at the FRS Marine Laboratory, Aberdeen. Many thanks to Sarah Canning for the photo of the harbour porpoises.