PHILOSOPHICAL TRANSACTIONS A

rsta.royalsocietypublishing.org

Review

Cite this article: Jolliffe IT, Cadima J. 2016 Principal component analysis: a review and recent developments. *Phil. Trans. R. Soc. A* **374**: 20150202.
http://dx.doi.org/10.1098/rsta.2015.0202

Accepted: 19 January 2016

One contribution of 13 to a theme issue 'Adaptive data analysis: theory and applications'.

Subject Areas:
statistics, atmospheric science, palaeontology

Keywords:
dimension reduction, eigenvectors, multivariate analysis, palaeontology

Author for correspondence:
Jorge Cadima
e-mail: jcadima@isa.ulisboa.pt

Principal component analysis: a review and recent developments

Ian T. Jolliffe[1] and Jorge Cadima[2,3]

[1]College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK
[2]Secção de Matemática (DCEB), Instituto Superior de Agronomia, Universidade de Lisboa, Tapada da Ajuda, Lisboa 1340-017, Portugal
[3]Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL), Lisboa, Portugal

Large datasets are increasingly common and are often difficult to interpret. Principal component analysis (PCA) is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables that successively maximize variance. Finding such new variables, the principal components, reduces to solving an eigenvalue/eigenvector problem, and the new variables are defined by the dataset at hand, not *a priori*, hence making PCA an adaptive data analysis technique. It is adaptive in another sense too, since variants of the technique have been developed that are tailored to various different data types and structures. This article will begin by introducing the basic ideas of PCA, discussing what it can and cannot do. It will then describe some variants of PCA and their application.

1. Introduction

Large datasets are increasingly widespread in many disciplines. In order to interpret such datasets, methods are required to drastically reduce their dimensionality in an interpretable way, such that most of the information in the data is preserved. Many techniques have been developed for this purpose, but principal component analysis (PCA) is one of the oldest and most widely used. Its idea is simple—reduce the dimensionality of a dataset, while preserving as much 'variability' (i.e. statistical information) as possible.

THE ROYAL SOCIETY PUBLISHING

Although it is used, and has sometimes been reinvented, in many different disciplines it is, at heart, a statistical technique and hence much of its development has been by statisticians.

This means that 'preserving as much variability as possible' translates into finding new variables that are linear functions of those in the original dataset, that successively maximize variance and that are uncorrelated with each other. Finding such new variables, the principal components (PCs), reduces to solving an eigenvalue/eigenvector problem. The earliest literature on PCA dates from Pearson [1] and Hotelling [2], but it was not until electronic computers became widely available decades later that it was computationally feasible to use it on datasets that were not trivially small. Since then its use has burgeoned and a large number of variants have been developed in many different disciplines. Substantial books have been written on the subject [3,4] and there are even whole books on variants of PCA for special types of data [5,6]. In §2, the formal definition of PCA will be given, in a standard context, together with a derivation showing that it can be obtained as the solution to an eigenproblem or, alternatively, from the singular value decomposition (SVD) of the (centred) data matrix. PCA can be based on either the covariance matrix or the correlation matrix. The choice between these analyses will be discussed. In either case, the new variables (the PCs) depend on the dataset, rather than being pre-defined basis functions, and so are adaptive in the broad sense. The main uses of PCA are descriptive, rather than inferential; an example will illustrate this.

Although for inferential purposes a multivariate normal (Gaussian) distribution of the dataset is usually assumed, PCA as a descriptive tool needs no distributional assumptions and, as such, is very much an adaptive exploratory method which can be used on numerical data of various types. Indeed, many adaptations of the basic methodology for different data types and structures have been developed, two of which will be described in §3a,d. Some techniques give simplified versions of PCs, in order to aid interpretation. Two of these are briefly described in §3b, which also includes an example of PCA, together with a simplified version, in atmospheric science, illustrating the adaptive potential of PCA in a specific context. Section 3c discusses one of the extensions of PCA that has been most active in recent years, namely robust PCA (RPCA). The explosion in very large datasets in areas such as image analysis or the analysis of Web data has brought about important methodological advances in data analysis which often find their roots in PCA. Each of §3a–d gives references to recent work. Some concluding remarks, emphasizing the breadth of application of PCA and its numerous adaptations, are made in §4.

2. The basic method

(a) Principal component analysis as an exploratory tool for data analysis

The standard context for PCA as an exploratory data analysis tool involves a dataset with observations on p *numerical* variables, for each of n entities or individuals. These data values define p n-dimensional vectors $\mathbf{x}_1, \ldots, \mathbf{x}_p$ or, equivalently, an $n \times p$ data matrix $\mathbf{X}$, whose jth column is the vector $\mathbf{x}_j$ of observations on the jth variable. We seek a linear combination of the columns of matrix $\mathbf{X}$ with maximum variance. Such linear combinations are given by $\sum_{j=1}^{p} a_j \mathbf{x}_j = \mathbf{Xa}$, where $\mathbf{a}$ is a vector of constants $a_1, a_2, \ldots, a_p$. The variance of any such linear combination is given by $\mathrm{var}(\mathbf{Xa}) = \mathbf{a}'\mathbf{Sa}$, where $\mathbf{S}$ is the sample covariance matrix associated with the dataset and $'$ denotes transpose. Hence, identifying the linear combination with maximum variance is equivalent to obtaining a p-dimensional vector $\mathbf{a}$ which maximizes the quadratic form $\mathbf{a}'\mathbf{Sa}$. For this problem to have a well-defined solution, an additional restriction must be imposed and the most common restriction involves working with unit-norm vectors, i.e. requiring $\mathbf{a}'\mathbf{a} = 1$. The problem is equivalent to maximizing $\mathbf{a}'\mathbf{Sa} - \lambda(\mathbf{a}'\mathbf{a} - 1)$, where λ is a Lagrange multiplier. Differentiating with respect to the vector $\mathbf{a}$, and equating to the null vector, produces the equation

$$\mathbf{Sa} - \lambda\mathbf{a} = \mathbf{0} \Longleftrightarrow \mathbf{Sa} = \lambda\mathbf{a}. \tag{2.1}$$

Thus, $\mathbf{a}$ must be a (unit-norm) eigenvector, and λ the corresponding eigenvalue, of the covariance matrix $\mathbf{S}$. In particular, we are interested in the *largest* eigenvalue, λ_1 (and corresponding eigenvector $\mathbf{a}_1$), since the eigenvalues are the variances of the linear combinations defined by the corresponding eigenvector $\mathbf{a}$: $\mathrm{var}(\mathbf{Xa}) = \mathbf{a}'\mathbf{Sa} = \lambda\mathbf{a}'\mathbf{a} = \lambda$. Equation (2.1) remains valid if the eigenvectors are multiplied by -1, and so the signs of all loadings (and scores) are arbitrary and only their relative magnitudes and sign patterns are meaningful.

Any $p \times p$ real symmetric matrix, such as a covariance matrix $\mathbf{S}$, has exactly p real eigenvalues, λ_k ($k = 1, \ldots, p$), and their corresponding eigenvectors can be defined to form an orthonormal set of vectors, i.e. $\mathbf{a}_k'\mathbf{a}_{k'} = 1$ if $k = k'$ and zero otherwise. A Lagrange multipliers approach, with the added restrictions of orthogonality of different coefficient vectors, can also be used to show that the full set of eigenvectors of $\mathbf{S}$ are the solutions to the problem of obtaining up to p new linear combinations $\mathbf{Xa}_k = \sum_{j=1}^{p} a_{jk}\mathbf{x}_j$, which successively maximize variance, subject to *uncorrelatedness* with previous linear combinations [4]. Uncorrelatedness results from the fact that the covariance between two such linear combinations, $\mathbf{Xa}_k$ and $\mathbf{Xa}_{k'}$, is given by $\mathbf{a}_{k'}'\mathbf{Sa}_k = \lambda_k\mathbf{a}_{k'}'\mathbf{a}_k = 0$ if $k' \neq k$.

It is these linear combinations $\mathbf{Xa}_k$ that are called the *principal components* of the dataset, although some authors confusingly also use the term 'principal components' when referring to the eigenvectors $\mathbf{a}_k$. In standard PCA terminology, the elements of the eigenvectors $\mathbf{a}_k$ are commonly called the *PC loadings*, whereas the elements of the linear combinations $\mathbf{Xa}_k$ are called the *PC scores*, as they are the values that each individual would score on a given PC.

It is common, in the standard approach, to define PCs as the linear combinations of the *centred* variables $\mathbf{x}_j^*$, with generic element $x_{ij}^* = x_{ij} - \bar{x}_j$, where $\bar{x}_j$ denotes the mean value of the observations on variable j. This convention does not change the solution (other than centring), since the covariance matrix of a set of centred or uncentred variables is the same, but it has the advantage of providing a direct connection to an alternative, more geometric approach to PCA.

Denoting by $\mathbf{X}^*$ the $n \times p$ matrix whose columns are the centred variables $\mathbf{x}_j^*$, we have

$$(n-1)\mathbf{S} = \mathbf{X}^{*'}\mathbf{X}^*. \tag{2.2}$$

Equation (2.2) links up the eigendecomposition of the covariance matrix $\mathbf{S}$ with the *singular value decomposition* of the column-centred data matrix $\mathbf{X}^*$. Any arbitrary matrix $\mathbf{Y}$ of dimension $n \times p$ and rank r (necessarily, $r \leq \min\{n, p\}$) can be written (e.g. [4]) as

$$\mathbf{Y} = \mathbf{ULA}', \tag{2.3}$$

where $\mathbf{U}, \mathbf{A}$ are $n \times r$ and $p \times r$ matrices with orthonormal columns ($\mathbf{U}'\mathbf{U} = \mathbf{I}_r = \mathbf{A}'\mathbf{A}$, with $\mathbf{I}_r$ the $r \times r$ identity matrix) and $\mathbf{L}$ is an $r \times r$ diagonal matrix. The columns of $\mathbf{A}$ are called the right singular vectors of $\mathbf{Y}$ and are the eigenvectors of the $p \times p$ matrix $\mathbf{Y}'\mathbf{Y}$ associated with its non-zero eigenvalues. The columns of $\mathbf{U}$ are called the left singular vectors of $\mathbf{Y}$ and are the eigenvectors of the $n \times n$ matrix $\mathbf{YY}'$ that correspond to its non-zero eigenvalues. The diagonal elements of matrix $\mathbf{L}$ are called the singular values of $\mathbf{Y}$ and are the non-negative square roots of the (common) non-zero eigenvalues of both matrix $\mathbf{Y}'\mathbf{Y}$ and matrix $\mathbf{YY}'$. We assume that the diagonal elements of $\mathbf{L}$ are in decreasing order, and this uniquely defines the order of the columns of $\mathbf{U}$ and $\mathbf{A}$ (except for the case of equal singular values [4]). Hence, taking $\mathbf{Y} = \mathbf{X}^*$, the right singular vectors of the column-centred data matrix $\mathbf{X}^*$ are the vectors $\mathbf{a}_k$ of PC loadings. Due to the orthogonality of the columns of $\mathbf{A}$, the columns of the matrix product $\mathbf{X}^*\mathbf{A} = \mathbf{ULA}'\mathbf{A} = \mathbf{UL}$ are the PCs of $\mathbf{X}^*$. The variances of these PCs are given by the squares of the singular values of $\mathbf{X}^*$, divided by $n-1$. Equivalently, and given (2.2) and the above properties,

$$(n-1)\mathbf{S} = \mathbf{X}^{*'}\mathbf{X}^* = \left(\mathbf{ULA}'\right)'\left(\mathbf{ULA}'\right) = \mathbf{ALU}'\mathbf{ULA}' = \mathbf{AL}^2\mathbf{A}', \tag{2.4}$$

where $\mathbf{L}^2$ is the diagonal matrix with the squared singular values (i.e. the eigenvalues of $(n-1)\mathbf{S}$). Equation (2.4) gives the *spectral decomposition*, or *eigendecomposition*, of matrix $(n-1)\mathbf{S}$. Hence, PCA is equivalent to an SVD of the column-centred data matrix $\mathbf{X}^*$.

The properties of an SVD imply interesting geometric interpretations of a PCA. Given any rank r matrix $\mathbf{Y}$ of size $n \times p$, the matrix $\mathbf{Y}_q$ of the same size, but of rank $q < r$, whose elements minimize

the sum of squared differences with corresponding elements of $\mathbf{Y}$ is given [7] by

$$\mathbf{Y}_q = \mathbf{U}_q \mathbf{L}_q \mathbf{A}_q', \tag{2.5}$$

where $\mathbf{L}_q$ is the $q \times q$ diagonal matrix with the first (largest) q diagonal elements of $\mathbf{L}$ and $\mathbf{U}_q$, $\mathbf{A}_q$ are the $n \times q$ and $p \times q$ matrices obtained by retaining the q corresponding columns in $\mathbf{U}$ and $\mathbf{A}$.

In our context, the n rows of a rank r column-centred data matrix $\mathbf{X}^*$ define a scatterplot of n points in an r-dimensional subspace of $\mathbb{R}^p$, with the origin as the centre of gravity of the scatterplot. The above result implies that the 'best' n-point approximation to this scatterplot, in a q-dimensional subspace, is given by the rows of $\mathbf{X}_q^*$, defined as in equation (2.5), where 'best' means that the sum of squared distances between corresponding points in each scatterplot is minimized, as in the original approach by Pearson [1]. The system of q axes in this representation is given by the first q PCs and defines a *principal subspace*. Hence, PCA is at heart a dimensionality-reduction method, whereby a set of p original variables can be replaced by an optimal set of q derived variables, the PCs. When $q = 2$ or $q = 3$, a graphical approximation of the n-point scatterplot is possible and is frequently used for an initial visual representation of the full dataset. It is important to note that this result is incremental (hence adaptive) in its dimensions, in the sense that the best subspace of dimension $q + 1$ is obtained by adding a further column of coordinates to those that defined the best q-dimensional solution.

The quality of any q-dimensional approximation can be measured by the variability associated with the set of retained PCs. In fact, the sum of variances of the p original variables is the trace (sum of diagonal elements) of the covariance matrix $\mathbf{S}$. Using simple matrix theory results it is straightforward to show that this value is also the sum of the variances of all p PCs. Hence, the standard measure of quality of a given PC is the *proportion of total variance* that it accounts for,

$$\pi_j = \frac{\lambda_j}{\sum_{j=1}^{p} \lambda_j} = \frac{\lambda_j}{\mathrm{tr}(\mathbf{S})}, \tag{2.6}$$

where $\mathrm{tr}(\mathbf{S})$ denotes the trace of $\mathbf{S}$. The incremental nature of PCs also means that we can speak of a proportion of total variance explained by a set $\mathcal{S}$ of PCs (usually, but not necessarily, the first q PCs), which is often expressed as a *percentage* of total variance accounted for: $\sum_{j \in \mathcal{S}} \pi_j \times 100\%$.

It is common practice to use some predefined percentage of total variance explained to decide how many PCs should be retained (70% of total variability is a common, if subjective, cut-off point), although the requirements of graphical representation often lead to the use of just the first two or three PCs. Even in such situations, the percentage of total variance accounted for is a fundamental tool to assess the quality of these low-dimensional graphical representations of the dataset. The emphasis in PCA is almost always on the first few PCs, but there are circumstances in which the last few may be of interest, such as in outlier detection [4] or some applications of image analysis (see §3c).

PCs can also be introduced as the optimal solutions to numerous other problems. Optimality criteria for PCA are discussed in detail in numerous sources (see [4,8,9], among others). McCabe [10] uses some of these criteria to select optimal *subsets* of the original variables, which he calls *principal variables*. This is a different, computationally more complex, problem [11].

(b) Example: fossil teeth data

PCA has been applied and found useful in very many disciplines. The two examples explored here and in §3b are very different in nature. The first examines a dataset consisting of nine measurements on 88 fossil teeth from the early mammalian insectivore Kuehneotherium, while the second, in §3b, is from atmospheric science.

Kuehneotherium is one of the earliest mammals and remains have been found during quarrying of limestone in South Wales, UK [12]. The bones and teeth were washed into fissures in the rock, about 200 million years ago, and all the lower molar teeth used in this analysis are from a single fissure. However, it looked possible that there were teeth from more than one species of Kuehneotherium in the sample.

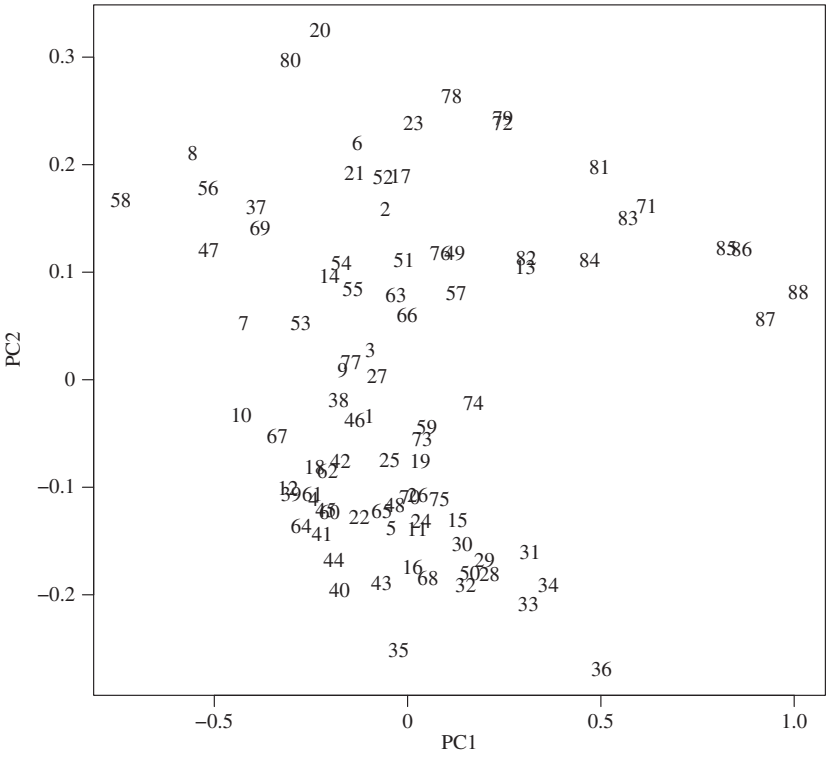

Figure 1. The two-dimensional principal subspace for the fossil teeth data. The coordinates in either or both PCs may switch signs when different software is used.

Of the nine variables, three measure aspects of the length of a tooth, while the other six are measurements related to height and width. A PCA was performed using the `prcomp` command of the R statistical software [13]. The first two PCs account for 78.8% and 16.7%, respectively, of the total variation in the dataset, so the two-dimensional scatter-plot of the 88 teeth given by figure 1 is a very good approximation to the original scatter-plot in nine-dimensional space. It is, by definition, the best variance-preserving two-dimensional plot of the data, representing over 95% of total variation. All of the loadings in the first PC have the same sign, so it is a weighted average of all variables, representing 'overall size'. In figure 1, large teeth are on the left and small teeth on the right. The second PC has negative loadings for the three length variables and positive loadings for the other six variables, representing an aspect of the 'shape' of teeth. Fossils near the top of figure 1 have smaller lengths, relative to their heights and widths, than those towards the bottom. The relatively compact cluster of points in the bottom half of figure 1 is thought to correspond to a species of Kuehneotherium, while the broader group at the top cannot be assigned to Kuehneotherium, but to some related, but as yet unidentified, animal.

(c) Some key issues

(i) Covariance and correlation matrix principal component analysis

So far, PCs have been presented as linear combinations of the (centred) original variables. However, the properties of PCA have some undesirable features when these variables have different units of measurement. While there is nothing inherently wrong, from a strictly mathematical point of view, with linear combinations of variables with different units of measurement (their use is widespread in, for instance, linear regression), the fact that PCA is

defined by a criterion (variance) that depends on units of measurement implies that PCs based on the covariance matrix **S** will change if the units of measurement on one or more of the variables change (unless *all p* variables undergo a *common* change of scale, in which case the new covariance matrix is merely a scalar multiple of the old one, hence with the same eigenvectors and the same proportion of total variance explained by each PC). To overcome this undesirable feature, it is common practice to begin by standardizing the variables. Each data value x_{ij} is both centred and divided by the standard deviation s_j of the n observations of variable j,

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}. \tag{2.7}$$

Thus, the initial data matrix **X** is replaced with the standardized data matrix **Z**, whose jth column is vector $\mathbf{z}_j$ with the n standardized observations of variable j (2.7). Standardization is useful because most changes of scale are linear transformations of the data, which share the same set of standardized data values.

Since the covariance matrix of a standardized dataset is merely the correlation matrix **R** of the original dataset, a PCA on the standardized data is also known as a correlation matrix PCA. The eigenvectors $\mathbf{a}_k$ of the correlation matrix **R** define the uncorrelated maximum-variance linear combinations $\mathbf{Z}\mathbf{a}_k = \sum_{j=1}^{p} a_{jk}\mathbf{z}_j$ *of the standardized variables* $\mathbf{z}_1, \ldots, \mathbf{z}_p$. Such *correlation matrix PCs* are not the same as, nor are they directly related to, the *covariance matrix PCs* defined previously. Also, the percentage variance accounted for by each PC will differ and, quite frequently, more correlation matrix PCs than covariance matrix PCs are needed to account for the same percentage of total variance. The trace of a correlation matrix **R** is merely the number p of variables used in the analysis, hence the proportion of total variance accounted for by any correlation matrix PC is just the variance of that PC divided by p. The SVD approach is also valid in this context. Since $(n-1)\mathbf{R} = \mathbf{Z}'\mathbf{Z}$, an SVD of the standardized data matrix **Z** amounts to a correlation matrix PCA of the dataset, along the lines described after equation (2.2).

Correlation matrix PCs are invariant to *linear* changes in units of measurement and are therefore the appropriate choice for datasets where different changes of scale are conceivable for each variable. Some statistical software assumes by default that a PCA means a correlation matrix PCA and, in some cases, the normalization used for the vectors of loadings $\mathbf{a}_k$ of correlation matrix PCs is not the standard $\mathbf{a}_k'\mathbf{a}_k = 1$. In a correlation matrix PCA, the coefficient of correlation between the jth variable and the kth PC is given by (see [4])

$$r_{\mathrm{var}_j,\mathrm{PC}_k} = \sqrt{\lambda_k}\, a_{jk}. \tag{2.8}$$

Thus, if the normalization $\tilde{\mathbf{a}}_k'\tilde{\mathbf{a}}_k = \lambda_k$ is used instead of $\mathbf{a}_k'\mathbf{a} = 1$, the coefficients of the new loading vectors $\tilde{\mathbf{a}}_k$ are the correlations between each original variable and the kth PC.

In the fossil teeth data of §2b, all nine measurements are in the same units, so a covariance matrix PCA makes sense. A correlation matrix PCA produces similar results, since the variances of the original variable do not differ very much. The first two correlation matrix PCs account for 93.7% of total variance. For other datasets, differences can be more substantial.

(ii) Biplots

One of the most informative graphical representations of a multivariate dataset is via a *biplot* [14], which is fundamentally connected to the SVD of a relevant data matrix, and therefore to PCA. A rank q approximation $\mathbf{X}_q^*$ of the full column-centred data matrix $\mathbf{X}^*$, defined by (2.5), is written as $\mathbf{X}_q^* = \mathbf{G}\mathbf{H}'$, where $\mathbf{G} = \mathbf{U}_q$ and $\mathbf{H} = \mathbf{A}_q\mathbf{L}_q$ (although other options are possible, see [4]). The n rows $\mathbf{g}_i$ of matrix **G** define graphical markers for each individual, which are usually represented by points. The p rows $\mathbf{h}_j$ of matrix **H** define markers for each variable and are usually represented by vectors. The properties of the biplot are best discussed assuming that $q = p$, although the biplot is

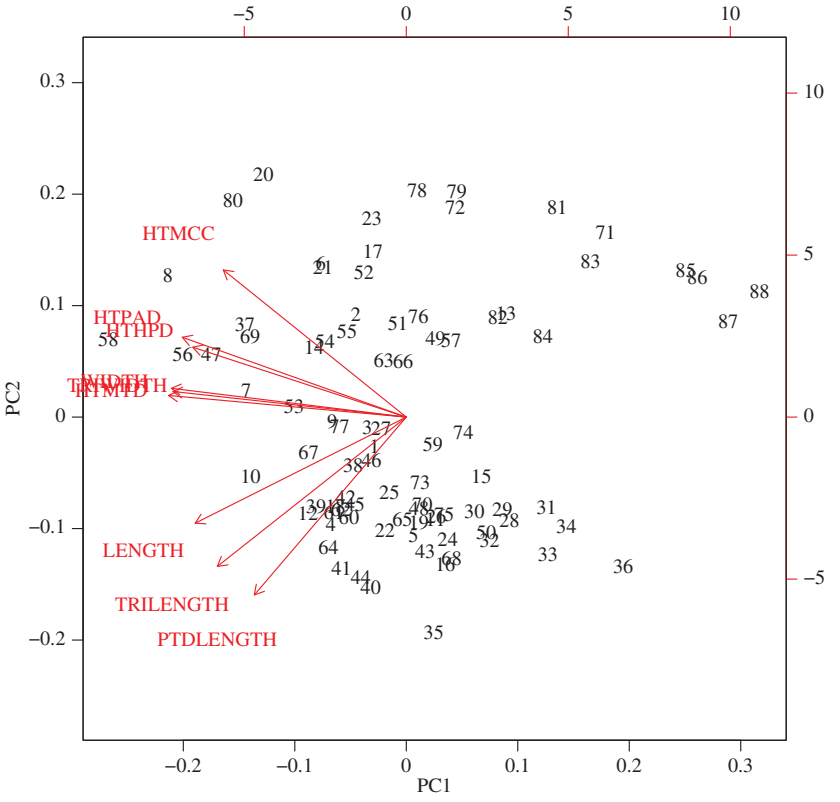

Figure 2. Biplot for the fossil teeth data (correlation matrix PCA), obtained using R's `biplot` command. (Online version in colour.)

defined on a low-rank approximation (usually $q = 2$), enabling a graphical representation of the markers. When $q = p$ the biplot has the following properties:

- The cosine of the angle between any two vectors representing variables is the coefficient of correlation between those variables; this is a direct result of the fact that the matrix of inner products between those markers is $\mathbf{HH}' = \mathbf{AL}^2\mathbf{A}' = (n-1)\mathbf{S}$ (2.4), so that inner products between vectors are proportional to covariances (variances for a common vector).
- Similarly, the cosine of the angle between any vector representing a variable and the axis representing a given PC is the coefficient of correlation between those two variables.
- The inner product between the markers for individual i and variable j gives the (centred) value of individual i on variable j. This is a direct result of the fact that $\mathbf{GH}' = \mathbf{X}^*$. The practical implication of this result is that orthogonally projecting the point representing individual i onto the vector representing variable j recovers the (centred) value $x_{ij} - \bar{x}_j$.
- The Euclidean distance between the markers for individuals i and i' is proportional to the Mahalanobis distance between them (see [4] for more details).

As stated above, these results are only exact if all $q = p$ dimensions are used. For $q < p$, the results are merely approximate and the overall quality of such approximations can be measured by the percentage of variance explained by the q largest variance PCs, which were used to build the marker matrices $\mathbf{G}$ and $\mathbf{H}$.

Figure 2 gives the biplot for the correlation matrix PCA of the fossil teeth data of §2b. The variable markers are displayed as arrows and the tooth markers as numbers. The group of three

nearly horizontal and very tightly knit variable markers for two width variables and one height variable, `WIDTH`, `HTMDT` and `TRIWIDTH`, suggests a group of highly correlated variables, which are also strongly correlated with the first PC (represented by the horizontal axis). The very high proportion of variability explained by the two-dimensional principal subspace provides solid grounds for these conclusions. In fact, the smallest of the three true coefficients of correlation between these three variables is 0.944 (`HTMDT` and `TRIWIDTH`), and the smallest magnitude correlation between PC1 and any of these variables is 0.960 (`TRIWIDTH`). The sign difference in PC2 loadings between the three length variables (towards the bottom left of the plot) and the other variables is clearly visible. Projecting the marker for individual 58 onto the positive directions of all variable markers suggests that fossil tooth 58 (on the left of the biplot) is a large tooth. Inspection of the data matrix confirms that it is the largest individual on six of the nine variables, and close to largest on the remaining three. Likewise, individuals 85–88 (on the right) are small-sized teeth. Individuals whose markers are close to the origin have values close to the mean for all variables.

(iii) Centrings

As was seen in §2, PCA amounts to an SVD of a column-centred data matrix. In some applications [15], centring the columns of the data matrix may be considered inappropriate. In such situations, it may be preferred to avoid any pre-processing of the data and to subject the uncentred data matrix to an SVD or, equivalently, to carry out the eigendecomposition of the matrix of non-centred second moments, $\mathbf{T}$, whose eigenvectors define linear combinations of the uncentred variables. This is often referred to as an *uncentred PCA* and there has been an unfortunate tendency in some fields to equate the name SVD only with this uncentred version of PCA.

Uncentred PCs are linear combinations of the uncentred variables which successively maximize non-central second moments, subject to having their crossed non-central second moments equal to zero. Except when the vector of column means $\bar{\mathbf{x}}$ (i.e. the centre of gravity of the original n-point scatterplot in p-dimensional space) is near zero (in which case centred and uncentred moments are similar), it is not immediately intuitive that there should be similarities between both variants of PCA. Cadima & Jolliffe [15] have explored the relations between the standard (column-centred) PCA and uncentred PCA and found them to be closer than might be expected, in particular when the size of vector $\bar{\mathbf{x}}$ is large. It is often the case that there are great similarities between many eigenvectors and (absolute) eigenvalues of the covariance matrix $\mathbf{S}$ and the corresponding matrix of non-centred second moments, $\mathbf{T}$.

In some applications, row centrings, or both row- and column-centring (known as double-centring) of the data matrix, have been considered appropriate. The SVDs of such matrices give rise to *row-centred* and *doubly centred PCA*, respectively.

(iv) When $n < p$

Datasets where there are fewer observed entities than variables ($n < p$) are becoming increasingly frequent, thanks to the growing ease of observing variables, together with the high costs of repeating observations in some contexts (such as microarrays [16]). For example, [17] has an example in genomics in which $n = 59$ and $p = 21\,225$.

In general, the rank of an $n \times p$ data matrix is $r \leq \min\{n, p\}$. If the data matrix has been column-centred, it is $r \leq \min\{n - 1, p\}$. When $n < p$, it is the number of observed individuals, rather than the number of variables, that usually determines the matrix rank. The rank of the column-centred data matrix $\mathbf{X}^*$ (or its standardized counterpart $\mathbf{Z}$) must equal the rank of the covariance (or correlation) matrix. The practical implication of this is that there are only r non-zero eigenvalues; hence r PCs explain all the variability in the dataset. Nothing prevents the use of PCA in such contexts, although some software, as is the case with R's `princomp` (but not the `prcomp`) command, may balk at such datasets. PCs can be determined as usual, by either an SVD of the (centred) data matrix or the eigenvectors/values of the covariance (or correlation) matrix.

Recent research (e.g. [18,19]) has examined how well underlying 'population' PCs are estimated by the sample PCs in the case where $n \ll p$, and it is shown that in some circumstances there is little resemblance between sample and population PCs. However, the results are typically based on a model for the data which has a very small number of structured PCs, and very many noise dimensions, and which has some links with recent work in RPCA (see §3c).

3. Adaptations of principal component analysis

The basic idea of PCA, leading to low-dimensional representations of large datasets in an adaptive and insightful way, is simple. However, the subsections in §2 have shown a number of subtleties that add some complexity. Going further, there are many ways to adapt PCA to achieve modified goals or to analyse data of different types. Because PCA is used in a large number of areas, research into modifications and adaptations is spread over literatures from many disciplines. Four such adaptations, chosen fairly arbitrarily from the many that exist, namely functional PCA, modifications of PCA to simplify interpretations, RPCA and symbolic data PCA, are described in the following subsections. Other adaptations are briefly mentioned in §4.

(a) Functional principal component analysis

In some applications, such as chemical spectroscopy, observations are functional in nature, changing with some continuous variable which, for simplicity, we assume is time. The dataset is then a collection of n functions $x_i(t)$.

How to incorporate such functional features in the analysis is the goal of functional data analysis [20]. Early work on functional PCA (e.g. [21]) performed a standard PCA on an $n \times p$ data matrix obtained by sampling n curves $x_i(t)$ at each of p points in time (t_j, with $j = 1, \ldots, p$), so that the element in row i, column j, of the data matrix is $x_i(t_j)$. The resulting p-dimensional vectors of loadings from a PCA of this data matrix are then viewed as sampled principal functions, which can be smoothed to recover functional form and can be interpeted as principal sources of variability in the observed curves [20]. The above approach does not make explicit use of the functional nature of the n observations $x_i(t)$. To do so requires adapting concepts. In the standard setting, we consider linear combinations of p vectors, which produce new vectors. Each element of the new vectors is the result of an inner product of row i of the data matrix, $(x_{i1}, x_{i2}, \ldots, x_{ip})$, with a p-dimensional vector of weights, $\mathbf{a} = (a_1, \ldots, a_p)$: $\sum_{j=1}^{p} a_j x_{ij}$. If rows of the data matrix become functions, a functional inner product must be used instead, between a 'loadings function', $a(t)$, and the ith functional observation, $x_i(t)$. The standard functional inner product is an integral of the form $\int a(t)x_i(t)\,\mathrm{d}t$, on some appropriate compact interval. Likewise, the analogue of the $p \times p$ covariance matrix $\mathbf{S}$ is a bivariate function $S(s,t)$ which, for any two given time instants s and t, returns the respective covariance, defined as

$$S(s,t) = \frac{1}{n-1}\sum_{i=1}^{n}[x_i(s) - \bar{x}(s)][x_i(t) - \bar{x}(t)] = \frac{1}{n-1}\sum_{i=1}^{n} x_i^*(s)x_i^*(t), \tag{3.1}$$

where $\bar{x}(t) = (1/n)\sum_{i=1}^{n} x_i(t)$ is the mean function and $x_i^*(t) = x_i(t) - \bar{x}(t)$ is the ith centred function.

The analogue of the eigen-equation (2.1) involves an *integral transform*, which reflects the functional nature of $S(s,t)$ and of inner products

$$\int S(s,t)a(t)\,\mathrm{d}t = \lambda a(s). \tag{3.2}$$

The *eigenfunctions* $a(t)$ which are the analytic solutions of equation (3.2) cannot, in general, be determined. Ramsay & Silverman [20] discuss approximate solutions based on numerical integration. An alternative approach, which they explore in greater detail, involves the assumption that the curves $x_i(t)$ can be written as linear combinations of a set of G *basis functions*

$\phi_1(t), \ldots, \phi_G(t)$, so that, for any data function i,

$$x_i(t) = \sum_{j=1}^{G} c_{ij}\phi_j(t). \tag{3.3}$$

These basis functions can be chosen to reflect characteristics that are considered relevant in describing the observed functions. Thus, Fourier series functions may be chosen to describe periodic traits and splines for more general trends (B-splines are recommended). Other basis functions that have been used and can be considered are wavelets, exponential, power or polynomial bases. In theory, other bases, adapted to specific properties of a given set of observed functions, may be considered, although the computational problems that arise from any such choice must be kept in mind. The advantage of the basis function approach lies in the simplification of the expressions given previously. Denoting the n-dimensional vector of functions $x_i(t)$ as $\mathbf{x}(t)$, the G-dimensional vector of basis functions as $\boldsymbol{\phi}(t)$ and the $n \times G$ matrix of coefficients c_{ij} as $\mathbf{C}$, the n data functions in equation (3.3) can be written as a single equation $\mathbf{x}(t) = \mathbf{C}\boldsymbol{\phi}(t)$. The eigenfunction $a(t)$ can also be written in terms of the basis functions, with $a(t) = \boldsymbol{\phi}(t)'\mathbf{b}$ for some G-dimensional vector of coefficients $\mathbf{b} = (b_1, \ldots, b_G)$. Assuming furthermore that $\mathbf{x}(t)$ and $\boldsymbol{\phi}(t)$ are centred, the covariance function at time (s, t) becomes

$$S(s,t) = \frac{1}{n-1}\mathbf{x}(t)'\mathbf{x}(t) = \frac{1}{n-1}\boldsymbol{\phi}(s)'\mathbf{C}'\mathbf{C}\boldsymbol{\phi}(t)$$

and eigen-equation (3.2) becomes, after some algebraic manipulation (see [4,20] for details),

$$\frac{1}{n-1}\boldsymbol{\phi}(s)'\mathbf{C}'\mathbf{C}\mathbf{W}\mathbf{b} = \lambda\boldsymbol{\phi}(s)'\mathbf{b},$$

where $\mathbf{W}$ is the $G \times G$ matrix of inner products $\int \phi_j(t)\phi_{j'}(t)\,\mathrm{d}t$ between the basis functions. Since this equation must hold for all values of s, it reduces to

$$\frac{1}{n-1}\mathbf{C}'\mathbf{C}\mathbf{W}\mathbf{b} = \lambda\mathbf{b}. \tag{3.4}$$

If the basis functions are orthonormal, $\mathbf{W}$ is the $G \times G$ identity matrix and we end up with a standard eigenvalue problem which provides the solutions $a(t) = \boldsymbol{\phi}(t)'\mathbf{b}$ to equation (3.2).

Ramsay & Silverman [20] further explore methods in which data functions $x_i(t)$ are viewed as solutions to differential equations, an approach which they call *principal differential analysis*, in order to highlight its close connections with PCA.

Research on functional PCA has continued apace since the publication of Ramsay and Silverman's comprehensive text. Often this research is parallel to, or extends, similar ideas for data of non-functional form. For example, deciding how many PCs to retain is an important topic. A large number of suggestions have been made for doing so [4] and many selection criteria are based on intuitive or descriptive ideas, such as the obvious 'proportion of total variance'. Other approaches are based on models for PCs. The problem of 'how many functional PCs?' is addressed in [22] using a model-based approach and criteria based on information theory.

As with other statistical techniques, it is possible that a few outlying observations may have a disproportionate effect on the results of a PCA. Numerous suggestions have been made for making PCA more robust to the presence of outliers for the usual data structure (see [4] and also §3c). One suggestion, using so-called S-estimators, is extended to functional PCA in [23].

Sometimes, as well as correlations between the p variables, there is a dependence structure between the n observations. A 'dynamic' version of functional PCA is proposed in [24], which is relevant when there are correlations *between* the observed curves, as well as the obvious correlation within the curves. It is based on an idea first suggested by Brillinger [25] for vector time series and uses frequency domain analysis.

(b) Simplified principal components

PCA gives the best possible representation of a p-dimensional dataset in q dimensions ($q < p$) in the sense of maximizing variance in q dimensions. A disadvantage is, however, that the new variables that it defines are usually linear functions of all p original variables. Although it was possible to interpret the first two PCs in the fossil teeth example, it is often the case for larger p that many variables have non-trivial coefficients in the first few components, making the components difficult to interpret. A number of adaptations of PCA have been suggested that try to make interpretation of the q dimensions simpler, while minimizing the loss of variance due to not using the PCs themselves. There is a trade-off between interpretability and variance. Two such classes of adaptations are briefly described here.

Rotation. The idea of rotating PCs is borrowed from factor analysis [26] (a different method, which is sometimes confused with PCA—see [4] for a fuller discussion). Suppose, as before, that $\mathbf{A}_q$ is the $p \times q$ matrix, whose columns are the loadings of the first q PCs. Then $\mathbf{XA}_q$ is the $n \times q$ matrix whose columns are the scores on the first q PCs for the n observations. Now let $\mathbf{T}$ be an orthogonal ($q \times q$) matrix. Multiplication of $\mathbf{A}_q$ by $\mathbf{T}$ performs an orthogonal rotation of the axes within the space spanned by the first q PCs, so that $\mathbf{B}_q = \mathbf{A}_q\mathbf{T}$ is a $p \times q$ matrix whose columns are loadings of q rotated PCs. The matrix $\mathbf{XB}_q$ is an $n \times q$ matrix containing the corresponding rotated PC scores. Any orthogonal matrix $\mathbf{T}$ could be used to rotate the components, but if it is desirable to make the rotated components easy to interpret, then $\mathbf{T}$ is chosen to optimize some simplicity criterion. A number of such criteria have been suggested, including some that include non-orthogonal (oblique) rotation [26]. The most popular is perhaps the varimax criterion in which an orthogonal matrix $\mathbf{T}$ is chosen to maximize $Q = \sum_{k=1}^{q}[\sum_{j=1}^{p} b_{jk}^4 - (1/p)\left(\sum_{j=1}^{p} b_{jk}^2\right)^2]$, where b_{jk} is the (j,k)th element of $\mathbf{B}_q$.

Rotation can considerably simplify interpretation and, when viewed with respect to the q-dimensional space that is rotated, no variance is lost, as the sum of variances of the q rotated components is the same as for the unrotated components. What is lost is the *successive* maximization of the unrotated PCs, so that the total variance of the q components is more evenly distributed between components after rotation.

Drawbacks of rotation include the need to choose from the plethora of possible rotation criteria, though this choice often makes less difference than the choice of how many components to rotate. The rotated components can look quite different if q is increased by 1, whereas the successively defined nature of unrotated PCs means that this does not happen.

Adding a constraint. Another approach to simplification of PCs is to impose a constraint on the loadings of the new variables. Again, there are a number of variants of this approach, one of which adapts the LASSO (least absolute shrinkage and selection operator) approach from linear regression [27]. In this approach, called SCoTLASS (simplified component technique–LASSO), components are found which successively solve the same optimization problem as PCA, but with the additional constraint $\sum_{j=1}^{p} |a_{jk}| \leq \tau$, where τ is a tuning parameter. For $\tau > \sqrt{p}$, the constraint has no effect and PCs are obtained, but as τ decreases more and more loadings are driven to zero, thus simplifying interpretion. These simplified components necessarily account for less variance than the corresponding number of PCs, and usually several values of τ are tried to determine a good trade-off between added simplicity and loss of variance.

A difference between the rotation and constraint approaches is that the latter has the advantage for interpretation of driving some loadings in the linear functions exactly to zero, whereas rotation usually does not. Adaptations of PCA in which many coefficients are exactly zero are generally known as sparse versions of PCA, and there has been a substantial amount of research on such PCs in recent years. A good review of such work can be found in Hastie *et al.* [28] (see also §3c).

A related technique to SCoTLASS adds a penalty function to the variance criterion maximized, so that the optimization problem becomes to successively find $\mathbf{a}_k$, $k = 1, 2, \ldots, p$, that maximize $\mathbf{a}_k'\mathbf{S}\mathbf{a}_k + \psi \sum_{j=1}^{p} |a_{jk}|$, subject to $\mathbf{a}_k'\mathbf{a}_k = 1$, where ψ is a tuning parameter [29]. One of the present

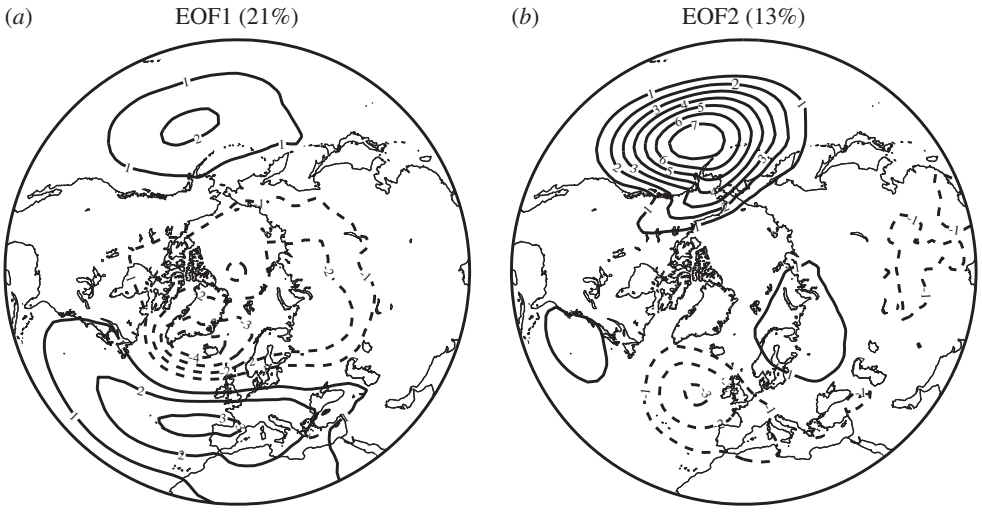

Figure 3. (*a*,*b*) The first two correlation-based EOFs for the SLP data account for 21% and 13% of total variation. (Adapted from [36].)

authors has recently reviewed a paper in which it is demonstrated that these apparently equivalent constraint and penalty approaches actually have quite distinct properties.

The original SCoTLASS optimization problem is non-convex and is also not solvable by simple iterative algorithms, although it is possible to re-express SCoTLASS as an equivalent, though still non-convex, optimization problem for which simple algorithms can be used [30]. Another approach, due to d'Aspremont *et al.* [31], reformulates SCoTLASS in a more complex manner, but then drops one of the constraints in this new formulation in order to make the problem convex.

Achieving sparsity is important for large p and particularly when $n \ll p$. A number of authors have investigated versions of sparse PCA for this situation using models for the data in which the vast majority of the variables are completely unstructured noise [18,19,32]. These papers and others suggest and investigate the properties of algorithms for estimating sparse PCs when data are generated from their models. Lee *et al.* [17] use a different type of model, this time a random effects model for PC loadings, to derive an alternative penalty function to that used by SCoTLASS, giving another sparse PCA method. Additionally incorporating shrinkage of eigenvalues leads to yet another method, deemed super-sparse PCA in [17]. Comparisons are given between their methods, SCoTLASS and the elastic net [28] for simulated data and a large genomic example.

(i) Example: sea-level pressure data

One discipline in which PCA has been widely used is atmospheric science. It was first suggested in that field by Obukhov [33] and Lorenz [34] and, uniquely to that discipline, it is usually known as empirical orthogonal function (EOF) analysis. The book by Preisendorfer & Mobley [35] discusses many aspects of PCA in the context of meteorology and oceanography.

The format of the data in atmospheric science is different from that of most other disciplines. This example is taken from [36]. The data consist of measurements of winter (December, January and February) monthly mean sea-level pressure (SLP) over the Northern Hemisphere north of 20° N. The dataset is available on a 2.5° × 2.5° regular grid and spans the period from January 1948 to December 2000. Some preprocessing is done to adjust for the annual cycle and the different areas covered by grid squares at different latitudes. In many atmospheric science examples, the variables are measurements at grid points, and the loadings, known as EOFs, are displayed as smooth spatial patterns, as in figure 3 for the first two correlation-based EOFs for the SLP data [36]. There are 1008 variables (grid-points) in this dataset, and the first two PCs account for 21% and 13% of the variation in these 1008 variables. Figure 3 gives a pattern which is commonly

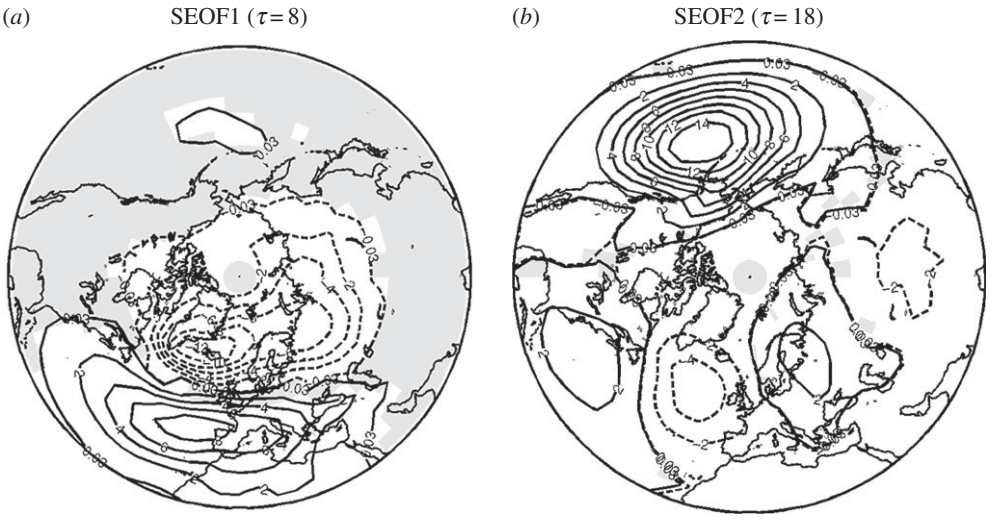

Figure 4. (*a*,*b*) LASSO-based simplified EOFs for the SLP data. Grey areas are grid-points with exactly zero loadings. (Adapted from [36].)

known as the Arctic Oscillation (AO). It is a measure of the north–south pressure gradient in the Atlantic Ocean and, to a lesser extent, in the Pacific Ocean and is a major source of variation in weather patterns. The second EOF is dominated by variation in the Pacific Ocean. The PCs for examples of this type are time series so the first PC, for example, will display which years have high values of the AO and which have low values.

Figure 4 shows simplified EOFs based on SCoTLASS [36]. The main difference from the EOFs in figure 3 is for the first EOF, which is now completely dominated by the north–south pressure gradient in the Atlantic (the North Atlantic Oscillation) with exactly zero loadings for many grid-points. The simplification is paid for by a reduction in percentage of variation explained for the corresponding simplified PC (17% compared with 21%). The second simplified PC is very similar to the original second EOF, also explaining 13% of variation.

(c) Robust principal component analysis

By its very nature, PCA is sensitive to the presence of outliers and therefore also to the presence of gross errors in the datasets. This has led to attempts to define robust variants of PCA and the expression RPCA has been used for different approaches to this problem. Early work by Huber [37,38] discussed robust alternatives to covariance or correlation matrices and ways in which they can be used to define robust PCs. This work was extended in [39,40]; see also [41].

The need for methods to deal with very large datasets in areas such as image processing, machine learning, bioinformatics or Web data analysis has generated a recent renewed interest in robust variants of PCA and has led to one of the most vigorous lines of research in PCA-related methods. A discussion of this issue can be found in [42]. Wright *et al.* [43] defined RPCA as a decomposition of an $n \times p$ data matrix $\mathbf{X}$ into a sum of two $n \times p$ components: a low-rank component $\mathbf{L}$ and a sparse component $\mathbf{S}$. More precisely, a convex optimization problem was defined as identifying the matrix components of $\mathbf{X} = \mathbf{L} + \mathbf{S}$ that minimize a linear combination of two different norms of the components:

$$\min_{\mathbf{L},\mathbf{S}} \|\mathbf{L}\|_* + \lambda \|\mathbf{S}\|_1, \tag{3.5}$$

where $\|\mathbf{L}\|_* = \sum_r \sigma_r(\mathbf{L})$, the sum of the singular values of $\mathbf{L}$, is the *nuclear norm* of $\mathbf{L}$, and $\lambda\|\mathbf{S}\|_1 = \sum_i \sum_j |s_{ij}|$ is the ℓ_1-norm of matrix $\mathbf{S}$. The motivation for such a decomposition is the fact that, in many applications, low-rank matrices are associated with a general pattern (e.g.

the 'correct' data in a corrupted dataset, a face in facial recognition, or a background image in video surveillance data), whereas a sparse matrix is associated with disturbances (e.g. corrupted data values, effects of light or shading in facial recognition, a moving object or person in the foreground of data surveillance images). Sparse components are also called 'noise', in what can be confusing terminology since in some applications it is precisely the 'noise' component that is of interest. Problem 3.5 has obvious points of contact with some of the discussion in §3b. Candès *et al.* [44] return to this problem, also called principal component pursuit, and give theoretical results proving that, under not very stringent conditions, it is possible to exactly recover the low-rank and sparse components with high probability and that the choice of $\lambda = 1/\sqrt{\max(n, m)}$ works well in a general setting, avoiding the need to choose a tuning parameter. Results are extended to the case of data matrices with missing values. Algorithms for the identification of the components are also discussed in [44], an important issue given the computational complexity involved. Further variations consider more complex structures for the 'noise' component. Some such proposals are reviewed in [45], where the results of alternative algorithms in the presence of different types of 'noise' are compared in the context of image-processing and facial recognition problems. Their results show that classical PCA performs fairly well, when compared with these new methods, in terms of both time and the quality of low-rank solutions that are produced. A fairly recent review of work in this area can be found in [46].

(d) Symbolic data principal component analysis

There is a recent body of work with so-called symbolic data, which is a general designation for more complex data structures, such as intervals or histograms [47,48].

Interval data arise when one wishes to retain a measure of underlying variability in the observations. This may occur if we wish to reflect the lack of precision of a measuring instrument or, more fundamentally, because the data are summary observations for which associated variability is considered inherent to the measurement. This is often the case when each observation corresponds to a group, rather than an individual, as would be the case with measurements on species, for which a range of values is considered part of the group value. If all p observed variables are of this type, each observation is represented by a hyper-rectangle, rather than a point, in p-dimensional space. Extensions of PCA for such data [47,49] seek PCs that are also of interval type, and which therefore also reflect ranges of values.

Another common type of symbolic data is given by histograms, which can be considered a generalization of interval-valued data where for each observation there are several intervals (the histogram bins) and associated frequencies. A recent review [50] covers several proposed definitions of PCA-type analyses for histogram data. Most of them require the definition of concepts such as distances between histograms (the Wasserstein distance being a common choice) or the sum and mean of histograms.

4. Conclusion

Although PCA in its standard form is a widely used and adaptive descriptive data analysis tool, it also has many adaptations of its own that make it useful to a wide variety of situations and data types in numerous disciplines. Adaptations of PCA have been proposed, among others, for binary data, ordinal data, compositional data, discrete data, symbolic data or data with special structure, such as time series [4] or datasets with common covariance matrices [6,40]. PCA or PCA-related approaches have also played an important direct role in other statistical methods, such as linear regression (with principal component regression [4]) and even simultaneous clustering of both individuals and variables [51]. Methods such as correspondance analysis, canonical correlation analysis or linear discriminant analysis may be only loosely connected to PCA, but, insofar as they are based on factorial decompositions of certain matrices, they share a common approach with PCA. The literature on PCA is vast and spans many disciplines. Space constraints mean that

it has been explored very superficially here. New adaptations and methodological results, as well as applications, are still appearing.

Data accessibility. The fossil teeth data are available from I.T.J. The atmospheric science data were taken from the publicly accessible NCEP/NCAR reanalysis database (see [36] for details).

Authors' contributions. Both authors were equally involved in drafting the manuscript.

Competing interests. We have no competing interests.

Funding. Research by J.C. is partially supported by the Portuguese Science Foundation FCT - PEst-OE/MAT/UI0006/2014.

Acknowledgements. We thank Pamela Gill and Abdel Hannachi for helpful discussions regarding their data and results.

References

1. Pearson K. 1901 On lines and planes of closest fit to systems of points in space. *Phil. Mag.* **2**, 559–572. (doi:10.1080/14786440109462720)
2. Hotelling H. 1933 Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24**, 417–441, 498–520. (doi:10.1037/h0071325)
3. Jackson JE. 1991 *A user's guide to principal components*. New York, NY: Wiley.
4. Jolliffe IT. 2002 *Principal component analysis*, 2nd edn. New York, NY: Springer-Verlag.
5. Diamantaras KI, Kung SY. 1996 *Principal component neural networks: theory and applications*. New York, NY: Wiley.
6. Flury B. 1988 *Common principal components and related models*. New York, NY: Wiley.
7. Horn R, Johnson C. 1985 *Matrix analysis*. Cambridge, UK: Cambridge University Press.
8. Hudlet R, Johnson RA. 1982 An extension of some optimal properties of principal components. *Ann. Inst. Statist. Math.* **34**, 105–110. (doi:10.1007/BF02481011)
9. Okamoto M. 1969 Optimality of principal components. In *Multivariate analysis II* (ed. PR Krishnaiah), pp. 673–685. New York, NY: Academic Press.
10. McCabe GP. 1984 Principal variables. *Technometrics* **26**, 137–144. (doi:10.1080/00401706.1984.10487939)
11. Cadima J, Cerdeira JO, Minhoto M. 2004 Computational aspects of algorithms for variable selection in the context of principal components. *Comp. Stat. Data Anal.* **47**, 225–236. (doi:10.1016/j.csda.2003.11.001)
12. Gill PG, Purnell MA, Crumpton N, Brown KR, Gostling NJ, Stampanoni M, Rayfield EJ. 2014 Dietary specializations and diversity in feeding ecology of the earliest stem mammals. *Nature* **512**, 303–305. (doi:10.1038/nature13622)
13. R Development Core Team. 2015 *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. See http://www.R-project.org.
14. Gabriel KR. 1971 The biplot graphical display of matrices with application to principal component analysis. *Biometrika* **58**, 453–467. (doi:10.1093/biomet/58.3.453)
15. Cadima J, Jolliffe IT. 2009 On relationships between uncentred and column-centred principal component analysis. *Pak. J. Stat.* **25**, 473–503.
16. Ringner M. 2008 What is principal component analysis? *Nat. Biotechnol.* **26**, 303–304. (doi:10.1038/nbt0308-303)
17. Lee D, Lee W, Lee Y, Pawitan Y. 2010 Super-sparse principal component analyses for high-throughput genomic data. *BMC Bioinform.* **11**, 296. (doi:10.1186/1471-2105-11-296)
18. Birnbaum A, Johnstone IM, Nadler B, Paul D. 2013 Minimax bounds for sparse PCA with noisy high-dimensional data. *Ann. Stat.* **41**, 1055–1084. (doi:10.1214/12-AOS1014)
19. Johnstone IM, Lu AY. 2009 On consistency and sparsity for principal components analysis in high dimensions. *J. Am. Stat. Assoc.* **104**, 682–693. (doi:10.1198/jasa.2009.0121)
20. Ramsay JO, Silverman BW. 2006 *Functional data analysis*, 2nd edn. Springer Series in Statistics. New York, NY: Springer.
21. Rao CR. 1958 Some statistical methods for comparison of growth curves. *Biometrics* **14**, 1–17. (doi:10.2307/2527726)
22. Li Y, Wang N, Carroll RJ. 2013 Selecting the number of principal components in functional data. *J. Am. Stat. Assoc.* **108**, 1284–1294. (doi:10.1080/01621459.2013.788980)
23. Boente G, Silibian-Barrera M. 2015 S-estimators for functional principal components. *J. Am. Stat. Assoc.* **110**, 1100–1111. doi:10.1080/01621459.2014.946991)

24. Hörmann S, Kidziński L, Hallin M. 2015 Dynamic functional principal components. *J. R. Stat. Soc. B* **77**, 319–348. (doi:10.1111/rssb.12076)
25. Brillinger DR. 1981 *Time series: data analysis and theory*, Expanded edn. San Francisco, CA: Holden-Day.
26. Cattell RB. 1978 *The scientific use of factor analysis in behavioral and life sciences*. New York, NY: Plenum Press.
27. Jolliffe IT, Trendafilov N, Uddin M. 2003 A modified principal component technique based on the LASSO. *J. Comput. Graph. Stat.* **12**, 531–547. (doi:10.1198/1061860032148)
28. Hastie T, Tibshirani R, Wainwright M. 2015 *Statistical learning with sparsity: the LASSO and generalizations*. Boca Raton, FL: CRC Press.
29. Zou H, Hastie T, Tibshirani R. 2006 Sparse principal components. *J. Comput. Graph. Stat.* **15**, 262–264. (doi:10.1198/jcgs.2006.s7)
30. Witten D, Tibshirani R, Hastie T. 2009 A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**, 515–534. (doi:10.1093/biostatistics/kxp008)
31. d'Aspremont A, El Ghaoui L, Jordan MI, Lanckriet GRG. 2007 A direct formulation for sparse PCA using semidefinite programming. *SIAM Rev.* **49**, 434–448. (doi:10.1137/050645506)
32. Lei J, Vu VQ. 2015 Sparsistency and agnostic inference in sparse PCA. *Ann. Stat.* **43**, 299–322. (doi:10.1214/14-AOS1273)
33. Obukhov AM. 1947 Statistically homogeneous fields on a sphere. *Usp. Mat. Navk.* **2**, 196–198.
34. Lorenz EN. 1956 Empirical orthogonal functions and statistical weather prediction. Technical report, Statistical Forecast Project Report 1, Dept. of Meteor. MIT: 49.
35. Preisendorfer RW, Mobley CD. 1988 *Principal component analysis in meteorology and oceanography*. Amsterdam, The Netherlands: Elsevier.
36. Hannachi A, Jolliffe IT, Stephenson DB, Trendafilov N. 2006 In search of simple structures in climate: simplifying EOFs. *Int. J. Climatol.* **26**, 7–28. (doi:10.1002/joc.1243)
37. Huber PJ. 1977 *Robust statistical procedures*. Philadelphia, PA: Society for Industrial and Applied Mathematics.
38. Huber PJ. 1981 *Robust statistics*. New York, NY: Wiley.
39. Ruymagaart FH. 1981 A robust principal component analysis. *J. Multivariate Anal.* **11**, 485–497. (doi:10.1016/0047-259X(81)90091-9)
40. Hallin M, Paindaveine D, Verdebout T. 2014 Efficient R-estimation of principal and common principal components. *J. Am. Stat. Assoc.* **109**, 1071–1083. (doi:10.1080/01621459.2014.880057)
41. Huber PJ, Ronchetti EM. 2009 *Robust statistics*, 2nd edn. Wiley Series in Probability and Statistics. New York, NY: Wiley.
42. De la Torre F, Black MJ. 2003 A framework for robust subspace learning. *Int. J. Comput. Vis.* **54**, 117–142. (doi:10.1023/A:1023709501986)
43. Wright J, Peng Y, Ma Y, Ganesh A, Rao S. 2009 Robust principal component analysis: exact recovery of corrupted low-rank matrices by convex optimization. In *Proc. of Neural Information Processing Systems 2009* (*NIPS 2009*), *Vancouver, BC, Canada, 7–10 December 2009*. See http://papers.nips.cc/paper/3704-robust-principal-component-analysis-exact-recovery-of-corrupted-low-rank-matrices-via-convex-optimization.pdf.
44. Candès EJ, Li X, Ma Y, Wright J. 2011 Robust principal component analysis? *J. ACM* **58**, 11:1–11:37.
45. Zhao Q, Meng D, Xu Z, Zuo W, Zhang L. 2014 Robust principal component analysis with complex noise. In *Proc. of the 31st Int. Conf. on Machine Learning, Beijing, China, 21–26 June 2014*. See http://jmlr.org/proceedings/papers/v32/zhao14.pdf.
46. Bouwmans T, Zahzah E. 2014 Robust PCA via principal component pursuit: a review for a comparative evaluation in video surveillance. *Comput. Vis. Image Underst.* **122**, 22–34. (doi:10.1016/j.cviu.2013.11.009)
47. Bock H-H, Diday E. 2000 *Analysis of symbolic data*. Berlin, Germany: Springer.
48. Brito P. 2014 Symbolic data analysis: another look at the interaction of data mining and statistics. *WIREs Data Mining Knowl. Discov.* **4**, 281–295. (doi:10.1002/widm.1133)
49. Ichino M, Yaguchi H. 1994 Generalized Minkowski metrics for mixed feature type data analysis. *IEEE Trans. Syst. Man Cybern.* **24**, 698–708. (doi:10.1109/21.286391)
50. Makosso-Kallyth S. In press. Principal axes analysis of symbolic histogram variables. *Stat. Anal. Data Mining*. (doi:10.1002/sam.11270)
51. Vichi M, Saporta G. 2009 Clustering and disjoint principal component analysis. *Comp. Stat. Data Anal.* **53**, 3194–3208. (doi:10.1016/j.csda.2008.05.028)