Principal component analysis

Article *in* Nature Reviews Methods Primers · December 2022

DOI: 10.1038/s43586-022-00184-w

6 authors, including:

Alfonso Iodice D'Enza
University of Naples Federico II
45 PUBLICATIONS **789** CITATIONS
SEE PROFILE

Angelos Markos
Democritus University of Thrace
108 PUBLICATIONS **1,879** CITATIONS
SEE PROFILE

Elena Tuzhilina
Stanford University
14 PUBLICATIONS **138** CITATIONS
SEE PROFILE

Principal Component Analysis

Michael Greenacre[1], Patrick J. F. Groenen[2], Trevor Hastie[3], Alfonso Iodice d'Enza[4], Angelos Markos[5], and Elena Tuzhilina[3],

[1] *Universitat Pompeu Fabra and Barcelona School of Management, Barcelona, Spain*
[2] *Erasmus School of Economics, Erasmus University, Rotterdam, The Netherlands*
[3] *Stanford University, Palo Alto, California, USA*
[4] *University of Naples Federico II, Naples, Italy*
[5] *Democritus University of Thrace, Alexandroupolis, Greece*

This is a preprint of an earlier version of the review published in *Nature Reviews Methods Primers*.

Principal component analysis is a versatile statistical method for reducing a cases-by-variables data table to its essential features, called principal components. Principal components are a few linear combinations of the original variables that maximally explain the variance of all the variables. In the process, the method provides an approximation of the original data table using only these few major components. In this review we present a comprehensive review of the method's definition and geometry, as well as the interpretation of its numerical and graphical results. The main graphical result is often in the form of a biplot, using the major components to map the cases and adding the original variables to support the distance interpretation of the cases' positions. Variants of the method are also treated, such as the analysis of grouped data as well as the analysis of categorical data, known as correspondence analysis. We also describe and illustrate the latest innovative applications of principal component analysis: its use for estimating missing values in huge data matrices, sparse component estimation, and the analysis of images, shapes and functions. Supplementary material includes video animations and computer scripts in the R environment.

1 Introduction

Principal component analysis[1–9] (abbreviated as PCA) is a multivariate statistical method that combines the information from several variables observed on the same subjects into fewer variables, called principal components (PCs). "Information" is measured by the total variance of the original variables, and the PCs optimally account for the major part of that variance. The PCs have geometric properties that allow for an intuitive and structured interpretation of the main features inherent in a complex multivariate dataset.

An introductory example is from the World Happiness Report[10] conducted in 2021 as part of the Gallup World Poll in 149 countries. This international study contains a measure of happiness on a 0 to 10 scale, called the Cantril ladder[11], as well as several indicators that possibly explain this happiness score. Here we consider five of these indicators: social support (abbreviated as *Social*), healthy life expectancy (*Life*), freedom to make your own life choices (*Choices*), generosity of the general population (*Generosity*) and perceptions of internal and external corruption levels (*Corruption*). PCA capitalizes on the relationships between these five indicators, so if the data were random and there was no correlation between any of the indicators, this approach would be fruitless. PCA looks for a linear combination of the indicators that has maximum variance; in other words, it combines them together in a way that reflects the greatest variation across the 149 countries. The following linear combination achieves this objective, and it defines the first principal component, PC1:

$$\begin{aligned}\text{PC1} \;=\; & 0.538\ \textit{Social} + 0.563\ \textit{Life} + 0.498\ \textit{Choices} \\ & - 0.004\ \textit{Generosity} - 0.381\ \ \textit{Corruption} \quad (1)\end{aligned}$$

Since the original indicators, usually called statistical variables, have different scales and ranges, they have each been standardized to have mean 0 and variance 1,

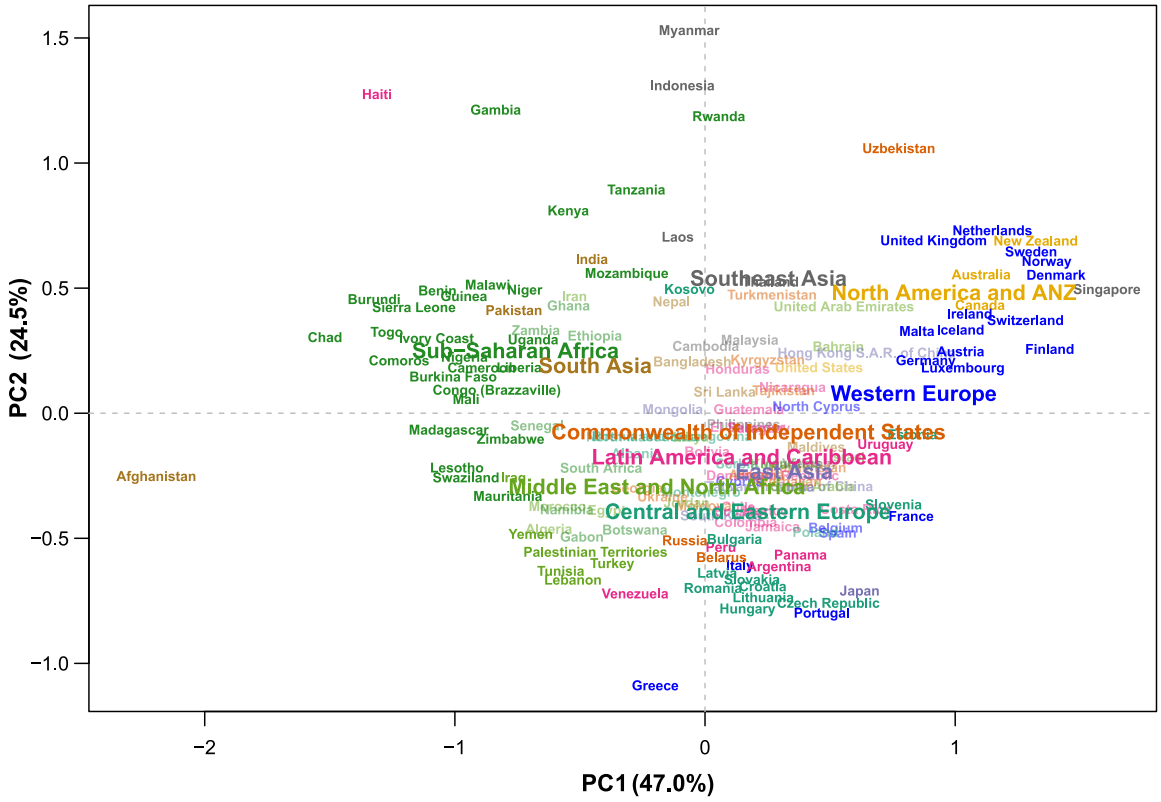

Figure 1: *Plot of multivariate data for 149 countries using the first two principal components as coordinate axes. The 82 countries that contribute more than average to the two-dimensional solution are shown in darker font and are generally further from the centre. The mean positions of the ten regions are added (each mean is at the centre of its label).*

so their total variance is 5. Thanks to this standardization, the coefficients of the variables, sometimes called loadings, indicate the strength of their contributions to the principal component and their signs indicate whether they influence it positively or negatively. PC1 can also be thought of as coming as close as possible in terms of correlation to all five variables, in other words a single variable summary of what they most have in common. If each of these five variables with a variance of 1 is regressed on PC1, their explained variances, usually denoted by R^2 and being identical to the squared correlations with PC1, are 0.680, 0.744, 0.583, 0.000, and 0.341. Hence, the second variable (*Life*) makes the largest contribution to PC1, whereas the fourth variable (*Generosity*) has almost none. The sum of these explained variances divided by the total 5, is 0.470, so that PC1 has "explained" 47.0% of the total variance.

Since 53.0% of the total variance has been left unexplained, a second linear combination of the variables is sought to explain as much of this residual variance as possible. The solution is the second principal component, PC2:

$$\begin{aligned}\text{PC2} = &-0.266\ \textit{Social} - 0.243\ \textit{Life} + 0.258\ \textit{Choices}\\ &+ 0.799\ \textit{Generosity} - 0.407\ \textit{Corruption}\end{aligned} \quad (2)$$

A condition in finding PC2 is that it should be uncorrelated with PC1, so that the principal components measure different features in the data. Again, the five original variables can each be regressed on the two principal components, leading to increased R^2 values of 0.767, 0.816, 0.664, 0.782, and 0.544 respectively, with an overall explained variance of 0.715, that is 71.5%. Hence, PC2 has explained an additional 24.5% of the variance.

The maximum number of PCs is the number of variables, five in this case, so this process can continue three more times to obtain PC3, PC4 and PC5, by which time 100% of the total variance will be explained. The first two PCs identified above can be computed for each of the 149 countries and plotted in a scatterplot (Fig. 1). The countries were classified into ten regions, so the positions of the regional averages can also be shown.

Notice that the signs of the coefficients are indeterminate, and different computer algorithms can produce the negative of PC1 or PC2, with all the signs reversed, and the interpretation of the components similarly reversed. The user is at liberty to multiply any principal component by -1, which simply inverts the corresponding axis in Fig. 1, in order to facilitate the interpretation.

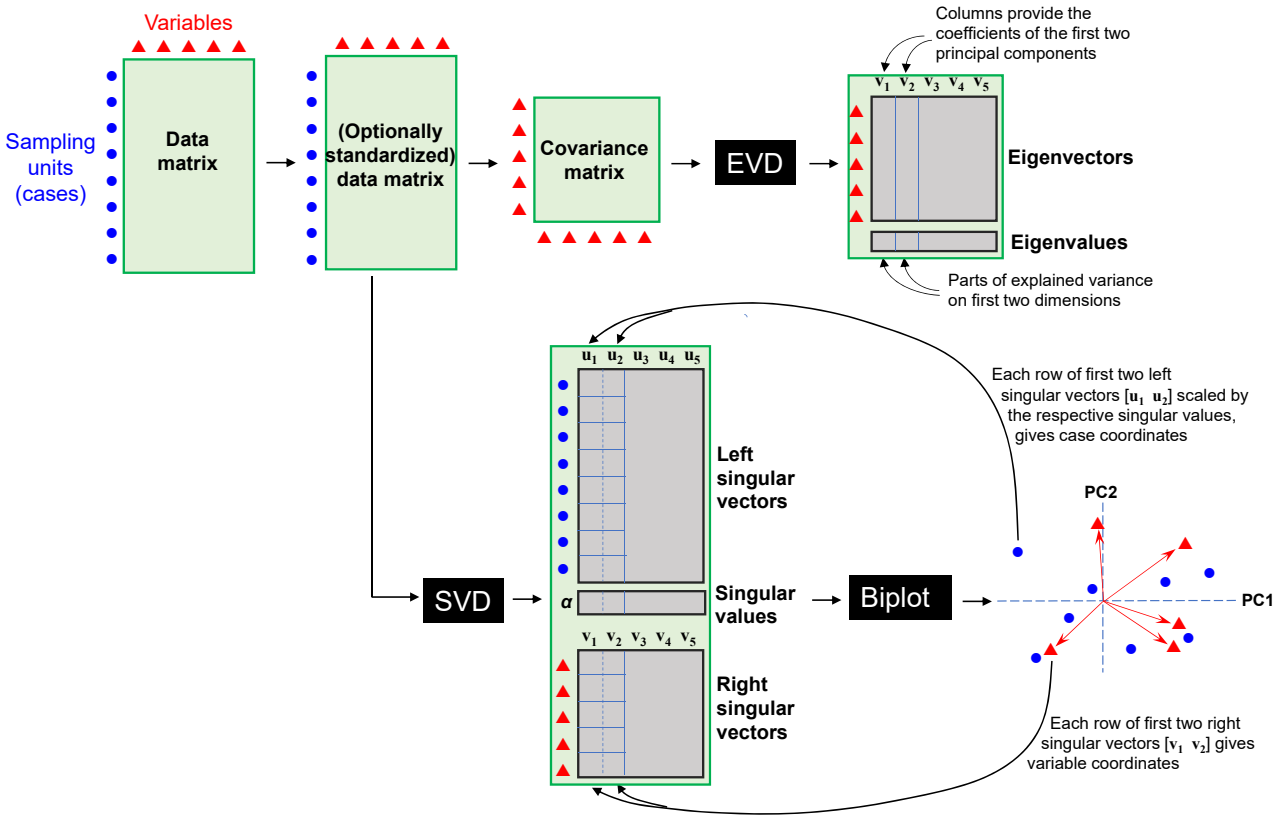

Figure 2: *Schematic view of the PCA workflow. The definition of the principal components (PCs) can be obtained using the eigenvalue decomposition (EVD) of the covariance matrix of the variables, Standardization is optional, but centring is mandatory, and if the variables are divided by their standard deviations, then the covariance matrix is the correlation matrix and the analysis is sometimes referred to as "correlation PCA". The first two PCs,* PC1 *and* PC2*, using the coefficients defined by the first two eigenvectors, given in (1) and (2), were used in Fig. 1 to obtain a spatial map of the countries. The lower pathway is a more efficient one, using the singular value decomposition (SVD) that leads directly to the positions of the countries as well as vectors serving as directions for the variables in the joint representation in Fig. 2. The eigenvectors are identical to the right singular vectors. For the lower pathway to be exactly equivalent to the upper one, the (optionally standardized) data matrix should be divided by* $\sqrt{n}$ *(see Box 1) .*

The visualization in Fig. 1 shows the approximate positions of the countries in terms of all five variables condensed into the two principal components, thus spreading out the countries in the two-dimensional plot as much as possible (i.e., maximizing variance). Interpretation of the country positions will be facilitated after enhancing the plot by showing the variables themselves in the display as well as any other variables observed on the countries (e.g., economic indicators), as explained in the following section.

2 Experimentation

2.1 PCA workflow

Step 1: Standardization of variables. The first and most important step in the PCA workflow is to make a decision about the standardization of the variables. PCA aims to explain the variables' variances, so it is essential that certain variables do not contribute excessively to that variance for extraneous reasons unrelated to the research question. For example, the variable *Life* (expectancy), was measured in years, *Generosity* was measured in positive and negative amounts and the other three variables lay in a 0 to 1 interval. In particular, *Life* has a very large variance due to its high numerical range of years. So, if no adjustment were made to its scale, it would dominate the total variance, with the PCA consequently being biased towards explaining that variable at the expense of the others.

In such a situation, with variables on different scales, a standardization is imposed on the variables. Dividing each variable's values by the respective variable's standard deviation is sufficient for removing the scale effect, but at the same time each variable is usually centred by subtracting its mean. This results in a set of scale-free variables each with mean 0 and variance 1, as was done here for the five variables. The contributions of these variables to the total variance are thus equalized, irrespective of the possible differences in the variables' substantive importance for the research question (see the later comments about weighting in PCA). As a general rule, software for PCA does not include automatic standardization of the variables, so

if standardization is required the user has to perform this manually before applying PCA or choose an option for standardization if the software includes it.

Alternative forms of standardization are possible, and sometimes pre-standardization is not necessary at all[8], for example if all the variables are on the same scale. If positive ratio-scale data are log-transformed, this is already a form of standardization of the variables to have comparable additive scales that reflect the multiplicative differences in the variables, and generally no further transformation is required[12].

Step 2: Dimension reduction. The present dataset, with $n = 149$ rows and $p = 5$ columns, is five-dimensional. The process of extracting the best small set of dimensions (often two), to facilitate interpretation and visualization is called dimension reduction, or (in algebraic parlance) low-rank matrix approximation. The top pathway of Fig. 2 shows how the principal components can be computed using the eigenvalue decomposition (EVD) of the covariance matrix. The EVD computes eigenvalues, denoted usually by $\lambda_1, \lambda_2, \ldots$ which in our five-dimensional example consist of five positive values in descending order, as well as eigenvectors corresponding to each eigenvalue, denoted by $\mathbf{v}_1, \mathbf{v}_2, \ldots$. The coefficients defining the principal components PC1 and PC2 in (1) and (2) are the elements of the two eigenvectors corresponding to the two highest eigenvalues. The eigenvalues themselves are the parts of variance that each PC explains, and the sum of all the eigenvalues is equal to the total variance. Hence, the percentages on the axes of Fig. 1 are λ_1 and λ_2 as percentages of the sum of all five.

The lower pathway shows the more efficient computational workflow. The singular value decomposition (SVD), which is a generalization of the EVD to arbitrary rectangular matrices, is applied directly to the matrix (optionally standardized, but at least centred), resulting in a set of positive singular values and two sets of vectors, the left and right singular vectors, for the rows and columns respectively. The singular values are proportional to the square roots of the eigenvalues of the covariance matrix and the left and right singular vectors lead to the joint display of cases and variables in the form of a biplot[13–15]. Specifically, the first two left singular vectors, $\mathbf{u}_1$ and $\mathbf{u}_2$, scaled by the respective singular values, α_1 and α_2, give the coordinates of the cases in Figs 1 and 3 — these coordinates defined by the principal components are also called principal coordinates. The coordinates of the direction vectors representing the variables in the biplot are given by the respective pairs of values in the two right singular vectors, $\mathbf{v}_1$ and $\mathbf{v}_2$, which are identical to the first two eigenvectors of the covariance matrix – these coordinates are also called standard coordinates. Box 1 shows a technical algebraic definition of the PCA coordinates obtained directly from the SVD. For a musical illustration of the SVD, see [16]. As indicated in Note 1 in Box 1, an alternative way of making a biplot is to leave the left singular vectors unscaled and scale the right singular vectors by the singular values, which focuses attention on the covariance and correlation structure of the variables, and less on the geometry of the cases.

Box 1: The singular value decomposition (SVD) and the PCA biplot coordinates

Given a data matrix $\mathbf{X}$, with n rows and p columns, already column-centred (i.e., column means subtracted from respective columns) and possibly column-standardized as well, the SVD decomposes $\mathbf{X}$ into three matrices of simple structure:

$$\mathbf{X} = \mathbf{U}\,\mathbf{D}\,\mathbf{V}^{\mathsf{T}}$$

where

- $\mathbf{D}$ is the diagonal matrix of the (positive) singular values $\alpha_1, \alpha_2, \ldots$ in descending order;
- $\mathbf{U}$ and $\mathbf{V}$ are the matrices of left and right singular vectors (with columns $\mathbf{u}_1, \mathbf{u}_2, \ldots$ and $\mathbf{v}_1, \mathbf{v}_2, \ldots$) and are orthonormal: $\mathbf{U}^{\mathsf{T}}\mathbf{U} = \mathbf{V}^{\mathsf{T}}\mathbf{V} = \mathbf{I}$, i.e. all $\mathbf{u}_k^{\mathsf{T}}\mathbf{u}_\ell$ and all $\mathbf{v}_k^{\mathsf{T}}\mathbf{v}_\ell$ are equal to 0 for $k \neq \ell$ but equal to 1 for $k \neq \ell$.

Written as a sum of products of the individual vectors, the SVD of $\mathbf{X}$ is $\sum_{k=1}^{m} \alpha_k \mathbf{u}_k \mathbf{v}_k^{\mathsf{T}}$, where m is the rank of $\mathbf{X}$. Since the sum of squares of each rank 1 matrix $\mathbf{u}_k\mathbf{v}_k^{\mathsf{T}}$ is equal to 1 and the singular values are in descending order, this suggests that taking the first terms of the sum will give an approximation to $\mathbf{X}$.

For the biplot the PCA row (principal) coordinates in r dimensions are in the first r columns of $\mathbf{UD}$, and the column (standard) coordinates in the first r columns of $\mathbf{V}$. The squares of the singular values, expressed relative to their sum, give the percentages of explained variance.

Notice the following:

1. An alternative version of the PCA biplot assigns the singular values to the right singular vectors, so the coordinates are in the first columns of $\mathbf{U}$ (row standard) and $\mathbf{VD}$ (column principal). This biplot focuses more on the internal structure of the column variables, and less on the distances between the row cases.
2. To obtain complete equivalence between the two alternative workflows shown in Fig. 2, the data matrix $\mathbf{X}$ (optionally standardized) should be rescaled prior to decomposition as follows: $\mathbf{X}/\sqrt{n}$, in which case the squared singular values are variances.

Step 3: Scaling and interpretation of the biplot. The resultant biplot is shown in Fig. 3. The countries are in the same positions as in Fig. 1, but now using symbols to make the display less cluttered. Their coordinates are obtained either by computing the linear combina-

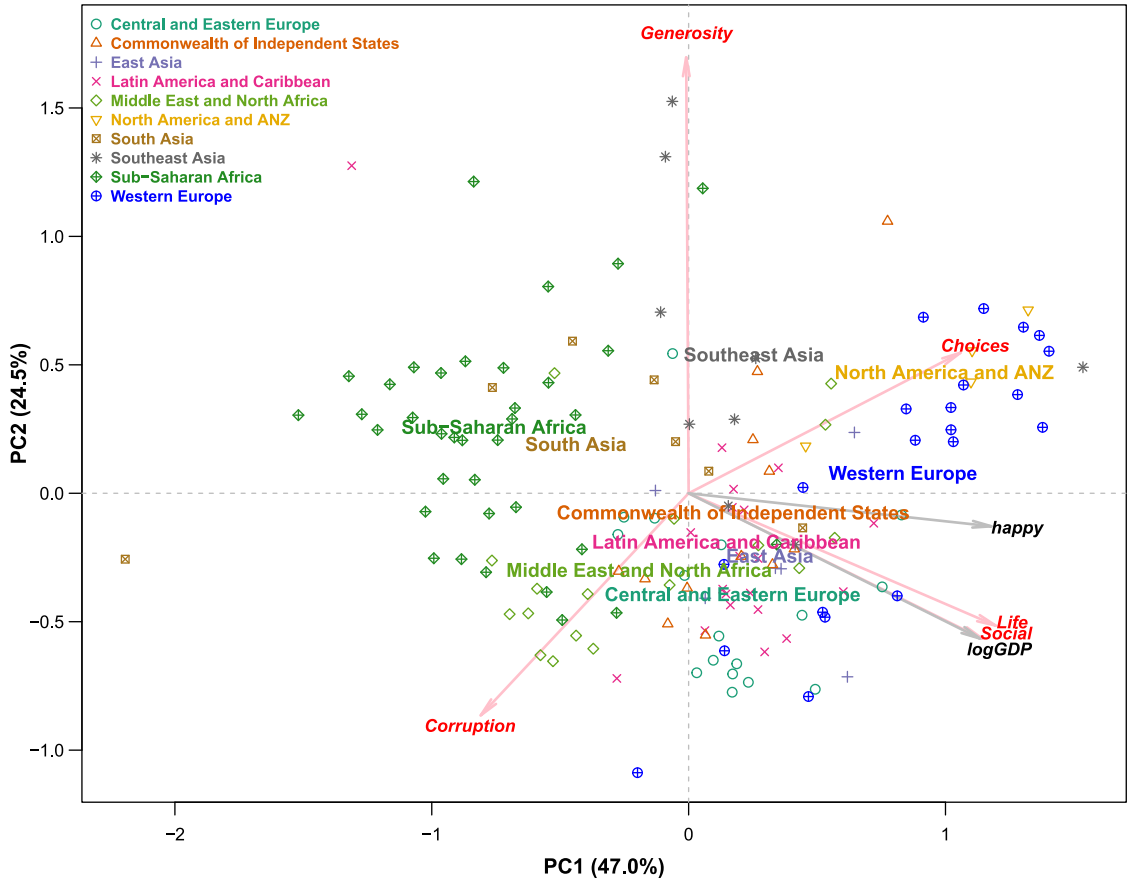

Figure 3: *Same plot as Fig. 1 (explained more fully later in the section on biplots) showing the countries with regional symbols, with regional means indicated by labels Now variables are shown as arrows of increasing values, with the means of all variables at the origin (i.e., point* [0 0]*). The scale of the variables is indicated on the upper and right sides of the plot box. Two supplementary variables,* ***happy*** *(the Cantril ladder happiness score) and* ***logGDP*** *(logarithm of gross domestic product per capita) have been added.*

tions originally defined as the principal components in (1) and (2), for each country, or equivalently using the left singular vectors scaled by the singular values, as just described. The arrows are defined by the pairs of coefficients in the two linear combinations; for example, the vector *Social* has coordinates $[\,0.538 \quad -0.266\,]$ in Fig. 3, according to the scale on the axes for the variables (cf. (1) and (2)).

The five variable directions define biplot axes onto which the countries can be projected perpendicularly. The means of the variables are all at the origin (due to the data centring) and the arrows indicate increasing values (a specific interpretation for this example is given later in the section Results: Interpretation of the biplot). Thus, when two variables point in the same direction, such as *Life* and *Social*, countries will project similarly onto them and suggest that the variables are strongly correlated (their actual correlation is 0.723). Conversely, for two variables that point in opposite directions, such as *Corruption* and *Choices*, this will suggest a negative correlation, since countries' projections onto them will line up in opposite directions (the actual correlation is -0.401). Suggested correlations are closer to actual ones when the dimensions explain a higher percentage of total variance.

Although it is the spatial interpretation with respect to the variable directions that is more important, it is often possible, as in this case, to interpret the principal component directions themselves (i.e., the dimensions, also called principal axes). The first dimension is clearly a negative to positive scale in terms of the four variables apart from *Generosity*, whereas *Generosity* is the main driver of the second dimension, opposing mainly *Corruption*. For example, looking at the positions of the UK, Malta, Germany and France in Fig. 1, they are all at the same position on the first horizontal dimension, but spread out vertically on the second. Thus, they have the same position on their overall "size" on this first dimension, but the composition of their ratings, their "shape", is different in the four countries. UK tends to be higher than average on *Generosity* and lower than average on *Corruption*, also lower on *Life* and *Social*, but nevertheless higher than average. On the other hand, France is higher on

all three variables pointing downwards and less than average on *Generosity*.

Step 4: Optional de-emphasising of cases or variables in the biplot. To show all 149 country names in Fig. 1, we resorted to distinguishing between countries that contributed more than average to the solution dimensions[17]. The left singular vectors corresponding to the countries, without scaling, each have sum of squares equal to 1, and the individual squared values are a direct measure of the proportional contributions to the variance explained on the respective dimension. The average contribution to a dimension is thus 1 divided by the number of points, 1/149 in this case. The countries with contributions greater than this threshold on either of the two dimensions are the ones plotted with higher intensity in Fig. 1, whereas the others, which are less than this threshold on both dimensions, are plotted using lighter labels. The high contributors are thus the points furthest from the origin on the respective dimensions. As an alternative, the countries were represented in Fig. 3 by symbols so that their regional dispersions could be visualized, but without indication of specific countries.

Step 5: Optional adding of supplementary variables to the biplot. If additional variables are available, these can be added to the biplot as supplementary variables, or passive variables. Since the directions of the five variables in the two-dimensional biplot can be equivalently obtained using the regression of these variables on the two principal components, similarly the direction of any other variable observed on the cases can be plotted to enrich the interpretation. The difference is that such a variable has not been optimized in the biplot like the five so-called active variables, which have been used to construct the solution. Two variables, the happiness score itself (abbreviated as *happy*), as well as the logarithm of GDP (*logGDP*), are available for the 149 countries and are represented in Fig. 3 as arrows. The coordinates of their arrowheads are the regression coefficients of each variable, also standardized, when regressed on PC1 and PC2. The principal components have explained variances (R^2) equal to 0.728 and 0.756 in these respective regressions, with *happy* being significantly explained by PC1 ($p < 0.0001$) and *logGDP* significantly explained by PC1 and PC2 (both $p < 0.0001$). The variable *logGDP* follows closely the directions of *Life* and *Social*, whereas the happiness score has a direction close to PC1 between these two indicators and *Choices*. The happiness score has a correlation of 0.850 with the first principal component.

2.2 EVD and SVD matrix decompositions

There are several equivalent ways to explain how the EVD and SVD provide optimal solutions in a PCA. An intuitive way is to accept that the eigenvalues, which are in decreasing order, maximize the explained variances on each dimension, and these dimensions are uncorrelated so that the parts of explained variance can be simply accumulated over the dimensions. Hence, as explained in the Introduction, the first eigenvalue maximizes the explained variance in the first dimension, the second eigenvalue maximizes the explained variance in the second, and the sum of the first two maximizes the explained variance in the plane of the first two dimensions, and so on for higher-dimensional solutions.

Another way is to think of the SVD as the solution of approximating the data matrix in a low-dimensional space, illustrated schematically in Fig. 4. Each row of the standardized data defines a point (shown as a solid dot) in multidimensional space, with as many dimensions as variables. If an approximation of these points in two dimensions is required, any plane through the average point *C* (for centroid) is imagined onto which all the points are projected perpendicularly (their projections are shown as empty dots in Fig. 4B) — this is equivalent to finding the closest point to each multidimensional point on the plane. Fig. 4C shows the right-angled triangle made by each point with its projection and the centroid. The hypotenuse distance d_i of the point to the centroid is fixed, whereas both e_i, the distance of the point to its projection, and $\hat{d}_i$, the distance from the projected point to the centroid, depend on the orientation of the unknown plane. To find the optimal plane in terms of least squares, it should minimize the sum of squared distances $\sum_i e_i^2$, i.e., the closeness of the plane to all the points, which is equivalent to maximizing $\sum_i \hat{d}_i^2$, since the total $\sum_i d_i^2$ is fixed. Averaging by dividing by n turns this into a decomposition of variance.

This is exactly the solution that the SVD finds, a least-squares approximation of the rows of the data matrix in a lower-dimensional subspace. All the approximated rows form a matrix which comes the closest to the data matrix in terms of least squared differences between the original and approximated matrices[18], hence this is often called least-squares matrix approximation. The equivalent approach, using the EVD of the covariance matrix, equivalently identifies the orientation of the two dimensions of the optimum plane (i.e., the principal component directions), leading to the same matrix approximation.

Because of the spatial interpretation of a PCA, it is essential to display the results in a space where the dimensions have the same physical scale – for example, notice in Figs 1 and 3 that unit lengths on the horizontal and vertical axes are physically equal, for each set of scales. In the terminology of image displays, the PCA graphics should have an aspect ratio of 1, like a spatial map or an architectural plan.

2.3 Variations of the PCA theme

There are several multivariate methods that are simple variants of PCA as described here. One possibility is to change the way the distance function is defined,

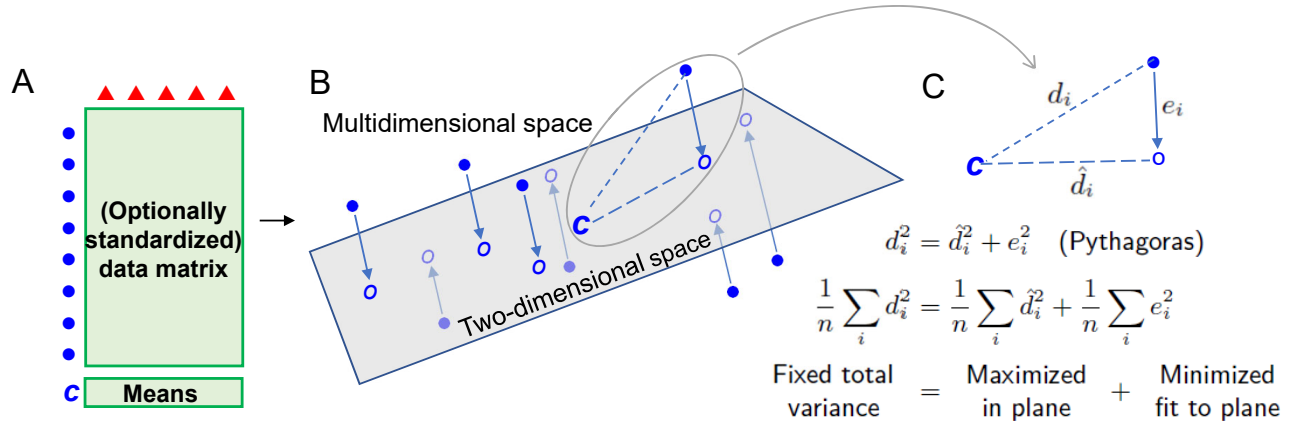

Figure 4: *Schematic view of dimension reduction in PCA. A. The rows of data, optionally standardized, and their mean (or centroid),* ***c****, define points in multidimensional space. B. The first two dimensions of the SVD identify the best-fitting two-dimensional plane in terms of least-squared distances between the plane and the points. This plane contains* ***c****, which becomes the zero point (or origin) of the PCA display and represents the averages of the variables. C. Each multidimensional data point defines a right-angled triangle with its projection onto the plane and the centroid. The average sum of squared distances of the points to the centroid is equal to the total variance, which is fixed. The maximization of average squared distances in the plane (i.e., maximizing variance) is equivalent to minimizing the average squared distances from the points to the plane (i.e., minimizing fit).*

which implies a change to the measure of total variance. Another variation is to assign different weights to the cases so that some cases count more than others in determining the PCA solution.

The distances between the projected points in a PCA are approximating the Euclidean distances between the points in the full space. The Euclidean distance between points i and i' is defined as:

$$d(i,i') = \sqrt{\sum_j (y_{ij} - y_{i'j})^2}, \qquad (3)$$

where the y_{ij} are the standardized data. If the original data are denoted by x_{ij} and standardization is performed by subtracting the mean $\bar{x}_j$ and dividing by the standard deviation s_j, then $y_{ij} = (x_{ij} - \bar{x}_j)/s_j$ and (3) reduces to

$$d(i,i') = \sqrt{\sum_j (x_{ij} - x_{i'j})^2 / s_j^2}, \qquad (4)$$

called the standardized Euclidean distance. The inverses of the variances $w_j = 1/s_j^2$ can be considered as weights on the variables.

A variant of PCA is correspondence analysis (CA), which is applicable to two-way cross-tabulations, general frequency data or data in the form of percentages. In CA it is the relative values of the data that are of interest, for example the rows divided by their row totals, called profiles. The distances between profiles, the chi-square distances, have a form similar to the standardized Euclidean distance. Denoting the (row) profile elements by r_{ij}:

$$d(i,i') = \sqrt{\sum_j (r_{ij} - r_{i'j})^2 / c_j}, \qquad (5)$$

where c_j is the j-th element of the average profile. Thus, for such relative frequency data, the mean profile element c_j substitutes the variance s_j^2 in (4), and the implied weights on the variables are the inverses $1/c_j$. In CA weights are also assigned to the profile points, which is explained in more detail later in the Applications section in an analysis of a dataset of abundance counts in marine ecology.

As an example of weighting of the cases in PCA, suppose that there are groups of cases and the object is to find dimensions that discriminate between the groups, that is, explaining between-group variance rather than the total between-case variance. Then weights proportional to the group sizes can be allocated to the group means, and the group means themselves become the points to be approximated by weighted least squares. The group means with higher weight then play a more important role in determining the low-dimensional solution. The original case points receive zero weight but can still be projected onto the plane that approximates the group points – these are called supplementary, or passive, points, as opposed to the group means, which are now the active points. This could have been done for the previous analysis of the five indicators of happiness if the objective had been to discriminate between the 10 regions. An example of PCA applied to weighted group means is given in the Applications section in an analysis of cancer tumours classified into four groups.

Another variant of PCA is logratio analysis (LRA), which has its origin in geochemistry but is increasingly being applied to biological data, especially microbiome data and data from "omics" research[19,20]. These data are generally compositional since the totals of each sample are irrelevant and it is their relative values

that are of interest. LRA is simply the PCA of the log-transformed data that are initially row-centred; that is, each row of the log-transformed data is centred by its respective row mean. Because PCA performs column-centring, LRA is thus the analysis of the double-centred matrix of log-transformed data, which has row and column means equal to 0. This is theoretically equivalent to the PCA of the much wider matrix of $\frac{1}{2}p(p-1)$ pairwise logratios of the form $\log(x_j/x_k)$ for all unique pairs of the p compositional variables[21,22]. LRA uses the logratio distance, which is the Euclidean distance computed on the logratios, and weights w_j can be optionally allocated to the compositional variables (as opposed to the cases, as described above)[20].

3 Results

3.1 Dimensionality of a PCA solution

Usually, the first question of interest is how much of the data variance is explained by the consecutive dimensions of the solution. PCA sorts the data variance into the major features in the data on the leading dimensions and what is considered random noise on the minor dimensions. The sequence of percentages of variance explained suggests how many non-random major dimensions there are. Fig. 5 shows the bar chart of the five percentages in the PCA of the five variables, where the percentages on the first two dimensions, 47.0% and 24.5%, can be seen to stand out from the last three. This observation can be reinforced by drawing a line (the red dashed line) through the last three, showing that the first two are above that approximate linear descending pattern. This bar plot is referred to as a scree plot[23] with the decision on the dimensionality made by looking for the "elbow" in the sequence of bars — see the similar line through the first two bars changing slope abruptly compared to the one through the last three. Based on this "elbow rule", the conclusion is that the data are two-dimensional, and the two-dimensional solutions presented before are thus validly representing the relevant data structure, with 47.0 + 24.5 = 71.5% of the variance explained and 28.5% of the variance declared random or unexplained. There are several more formal ways of deciding on the number of non-random dimensions in PCA[23–30].

It is not expected that datasets always turn out to have exactly two major dimensions; they could have a single major dimension or more than two. The former case is not problematic — usually the first two dimensions would be visualized anyway, with the caveat that the second dimension is possibly compatible with random variation, and interpretation should be restricted to the dispersion of points and variables along the first dimension. In the latter case, for a three-dimensional solution, three-dimensional graphics can be used, or a selection of planar views of the points made, for example, dimensions 1 and 2, and then separately, dimensions 1 and 3, or for four-dimensional solutions, a plot of dimensions 1 and 2, and a plot of dimensions 3 and 4, an example of which is given in [31].

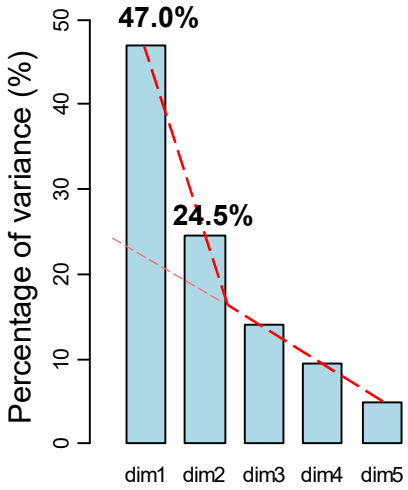

Figure 5: *Scree plot of the percentages of variance explained by the first two PCs as well as the percentages explained by the remaining three, showing the elbow that suggests that the first two dimensions are signal, whereas the last three dimensions are random noise.*

3.2 Interpretation of a PCA biplot

The PCA biplot in Fig. 3, explaining 71.5% of the data variance, consists of points for the cases and vectors for the variables. As shown in Fig. 4, the positions of the points projected onto the reduced-dimensional subspace, usually a plane, are an optimal approximation of their exact positions in the "full" multidimensional space. The distances between the projected points are approximating the distances between the points in the full space. Thus, the case (row) points in the biplot solution have a distance interpretation, and the quality of the distance interpretation is assessed by the percentage of variance explained by the solution dimensions. In fact, the coordinates of the case points are identical, up to a scalar multiplying factor, to the solution coordinates of the distance-based method called classical multidimensional scaling (MDS), which takes the exact interpoint distances as input and produces low-dimensional approximations of the distances[32].

The variables, usually represented by vectors from the origin in different directions, define the directions and sense of the changing values of the variables. Case points can be projected perpendicularly onto these directions in order to understand how the cases line up approximately, but not exactly. To give a concrete example, the country points can be projected perpendicularly onto a biplot axis pointing in the top right direction of Fig. 3, corresponding to *Choices*. Countries such as the Scandinavian ones, Sweden, Norway and Denmark, as well as Singapore (check the country names in Fig. 1) are highest in the positive direction of the arrow, whereas Afghanistan is the lowest on the

negative side of that direction towards bottom left. Remember that the origin of the biplot represents the means of all five variables, so that countries projecting on the upper right of the biplot axis of *Choices* are estimated to be above the mean and those on the lower left are estimated to be below the mean. Taking the projected values for all the countries onto that diagonal sloping axis and correlating them with the original data for this variable, gives a correlation of 0.815. The squared correlation, 0.664, is just the part of variance of *Choices* explained by the first two principal components, which was given earlier in the introduction of this dataset, after defining the PCs in (1) and (2).

The set of countries can be projected in turn on each of the other biplot axes defined by the direction vectors, and the projected positions are as accurate as the proportions of explained variance, the R^2 values of 0.767, 0.816, 0.664, 0.782, and 0.544, as reported in the Introduction. The second variable, *Life*, has the highest R^2, so the way the countries line up on this direction in the biplot will be the most accurate, whereas the projections onto the fifth variable, *Corruption*, with the lowest R^2, will give less accurate estimates. The projected positions of the countries onto the five biplot axes are simply the data values estimated by the two principal components PC1 and PC2 by multiple regression, thus reinforcing the idea that PCA is a method of matrix approximation.

3.3 Numerical results of a PCA

The values of the percentages of variance have already been plotted and interpreted in Fig. 5, and the quality of the approximation of the variables by the principal components has been measured by the respective R^2 values. Additional numerical results are in the form of correlations and contributions. In this particular case where the variance of each of the five standardized variables is 1, the correlations in the columns of Table 1 are the principal component direction vectors (eigenvectors) multiplied by the respective singular values of the standardized data matrix divided by $\sqrt{n}$. For example, the correlation of 0.825 between *Social* and PC1 is equal to 0.538×1.532 (see first coefficient of PC1 in (1)). Since all the eigenvectors have sum of squares equal to 1, and thus equally standardized, this illustrates in a different way why the correlations with the major dimensions are higher, because the singular values are higher.

The sum of squared correlations column-wise in Table 1 are the parts of variance, identical to the squares of the first row, i.e. the squared singular values (eigenvalues) divided by n. The sum of squared correlations of each variable row-wise over the five dimensions in Table 1 is equal to 1, while the sum of squared correlations over the first two dimensions is the corresponding R^2 for the two-dimensional PCA solution; for example, for *Choices*, $0.7642 + 0.2852 = 0.664$. Again, this only holds for this particular case of standardized variables.

	PC1	PC2	PC3	PC4	PC5	Row SS
Singular values/$\sqrt{n}$	1.532	1.107	0.838	0.692	0.495	5
Social	0.825	-0.295	0.303	0.183	0.328	1
Life	0.862	-0.269	0.002	0.252	-0.347	1
Choices	0.764	0.285	0.178	-0.549	-0.050	1
Generosity	-0.007	0.884	0.380	0.268	-0.038	1
Corruption	-0.584	-0.451	0.659	-0.091	-0.114	1
Column variables SS	2.348	1.226	0.703	0.478	0.245	Row sum 5

Table 1: *Correlations of the five variables with the five PCs of the standardized data. The sum of squares (SS) of the correlations for each variable is 1. The sum of squared correlations for each PC is the square of the first row (i.e., squared singular value divided by n) and is that PC's part of variance explained out of a total variance of 5. Expressed as percentages these are the percentages on the PC dimensions, plotted in Fig. 5. For example, on the first dimension,* $100 \times 2.348/5 = 47.0\%$.

Contributions of the variables are the squared correlations in the columns of Table 1 relative to their sum. For example, in column 1, the contributions by the five variables to the first PC are $[\,0.825^2 \;\; 0.862^2 \;\; 0.764^2 \;\; (-0.007)^2 \;\; (-0.584)^2\,]/2.348 = [\,0.290 \;\; 0.317 \;\; 0.248 \;\; 0.000 \;\; 0.145\,]$ — hence, these are just the squares of the PC direction vector elements. Thus, it is mainly the first three variables that contribute highly to the construction of the first principal component. Computing contributions to variance on the major PCs is useful when there are very many variables and the biplot becomes too cluttered – a strategy is then to show only the high contributors, usually defined as those that are above average. This idea can also be applied when there are very many rows, since each row also contributes to the dimensional variance, using the squared elements of the left singular vectors. This tactic was used in Fig. 1, where the above average country contributors were shown in a more intense colour in order to improve the legibility of the biplot. It will also be used later in a genetics application (see the Applications Section 4.1) where there are thousands of variables (genes).

4 Applications

4.1 A high-dimensional dataset with groups of cases

Cases (usually the rows of the data matrix) are frequently grouped and the research question is to identify variables (the columns) that account for this grouping. The Khan child cancer dataset[33,34], consists of a 63×2308 matrix of gene expression data, for 63 children and 2308 genes. The children have small, round

blue-cell tumours, classified into four major types: BL (Burkitt lymphoma, $n = 8$), EW (Ewing's sarcoma, $n = 23$), NB (neuroblastoma, $n = 12$)), and RM (rhabdomyosarcoma, $n = 20$). The data are already given as log-transformed, and no further standardization is required. The number of variables is higher than the number of cases (i.e., tumours), and the dimensionality of the data is thus determined by the number of cases minus 1: $63 - 1 = 62$. To understand this, and remembering that the data are column-centred, notice that 2 cases in a high-dimensional space lie exactly on a line (1-dimensional), 3 cases in a plane (2-dimensional), 4 cases in a 3-dimensional space, and so on.

Fig. 6A shows the PCA of the data, where the four tumour groups are grouped by enclosing them in convex hulls. The genes are displayed as shaded dots, the darkest being the ones that make the highest contributions to the two-dimensional solution. As for the countries in Fig. 1, these high-contributing genes are the most outlying in the biplot, and are similarly contributing to explaining the variance in the individual cases, not necessarily to explaining the variance between the cancer groups. The individual tumors in the different groups can be seen to overlap substantially, especially the groups EW and RM. Also shown in Fig. 6A are confidence ellipses for the group mean points[35]. These are obtained by estimating the bivariate normal distribution for each group of points, and then showing the area containing 95% of the bivariate normal probability for the respective bivariate mean, taking into account the bivariate correlation and margins of error. For the means as well, the confidence ellipses for RM and EW overlap, but their means show significant separation from NB and BL, which themselves appear significantly separated in this PCA solution.

To account for the separation of the groups, a different two-dimensional solution in the 62-dimensional space of the cases can be found, where the group means (i.e., centroids) are optimally separated. This is achieved by computing the means of the groups and using these four points, weighted by their respective group sample sizes, as the data of primary interest. Whereas Fig. 6A can be qualified as an unsupervised PCA, the PCA in Fig. 6B is now supervised to explain group differences. This PCA of the four group means has only three dimensions, and the percentages on the dimensions are thus much higher as they are expressed relative to the between-group variance. The group means are highly separated now, and the convex hulls do not overlap at all, as well as the confidence ellipses, which are now much tighter. In this solution the outlying highly contributing genes will be the ones that account for the group differences. Notice that this weighted PCA of the centroids ignores the covariances within the groups, and is thus a simpler form of Fisher's linear discriminant analysis[36], also called canonical variate analysis[37], which do take these covariances into account. Video 1 of the Supplementary Material shows the exact three-dimensional solution of the group centroids. Video 2 shows an animation of the cases in Fig. 6A transitioning to the group separation in Fig. 6B as weight is taken off smoothly from the individual cases and transferred to the group means. The effect on predicting the tumour group for a hold-out test set is reported in [17].

4.2 Sparsity constraints for wide data

The coefficients that define the principal components are generally all non-zero, or dense. For "wide" data, that is when the number of features is very high, in the hundreds and sometimes thousands, this presents a problem for interpreting so many coefficients. This is the case with the present cancer dataset as well as for microbiome data and "omics" data in general, where there can be thousands of variables compared to a small number of samples. The interpretation would be considerably simplified if some of the coefficients were zero, that is, if they were more sparse. Earlier attempts for partially solving this problem were made by rotating the PCA solution so that variables aligned themselves closer to the dimensions[38,39].

More recently, sparse PCA implementations[40–46] can handle this problem by introducing penalties on the sizes of the coefficients that force some coefficients down to zero, hence eliminating them from the interpretation of the respective principal components. For example, combined with the objective of explaining variance, the lasso penalty[47] restricts the sum of the absolute values of the coefficients, similar to lasso regression. The result is a small sacrifice of the variance-explaining objective in order to shrink the absolute values of some coefficients down to zero. An improvement to achieve coefficient sparsity is also made using the elastic-net penalty[48] which restricts both the sum of the absolute values of the coefficients and their sum of squares. For a recent comprehensive review of sparse PCA methods, see [49]. Sparse PCA is a fairly recent innovation, and is still actively debated in the literature (e.g., see [50,51]).

Fig. 6C shows the effect of the sparse PCA on the results of the Khan gene data shown in Fig. 6A. Most of the 2308 genes have been eliminated from the results, leaving the remaining few with nonzero values either on PC1 or PC2 (103 for PC1 and 84 for PC2), and a few nonzero for both PCs. The configuration of the samples and their averages in Fig. 6C is very similar to Fig. 6A. Within each cancer group there is a vertical separation of samples with positive PC2 and those with negative PC2, which is even more accentuated now. The genes that lie on the vertical axis will be the indicators of this separation. On the horizontal dimension the genes with nonzero values will be related to the separation of the cancer groups, especially RM versus BL. To achieve this simplified interpretation 2.5 percentage points of the explained variance have been sacrificed, compared to Fig. 6A. In the sparse centroid PCA of Fig. 6D the cancer groups are separated and

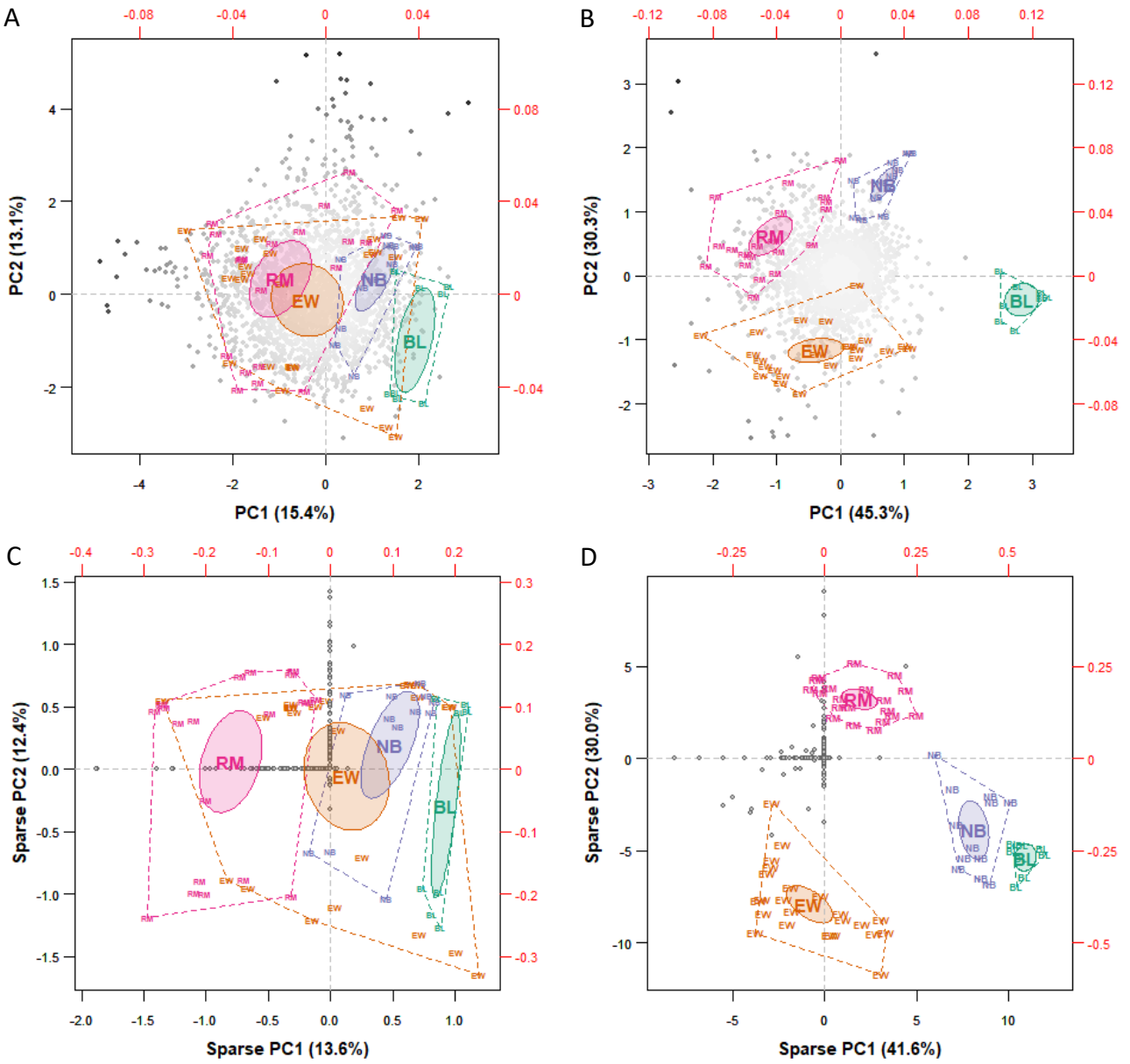

Figure 6: *A. PCA of the Khan child cancer data. The four tumour groups (EW, RM, NB, BL) are enclosed by convex hulls, and 95% confidence ellipses are shown for the group means, which are located at the group label in larger font. The 2308 genes are displayed as dots, where darker dots indicate higher contributions to the separation of individual tumours. B. Supervised PCA of the Khan child cancer data, explaining the between-group variance. The four tumour groups are again enclosed by convex hulls, with confidence ellipses for the group means, which are now all separated. The darker dots now correspond to genes making higher contributions to the group separation. C. Sparse PCA of the Khan data (explained in the next section), comparable to the regular PCA in FIG. 4A. Most of the 2308 genes are eliminated and the remaining genes are now identified with either the first or second principal component, and in a few cases with both PCs. The percentage of explained variance has dropped from 28.5% in the solution of Fig. 6A to 26.0%. D. Sparse PCA of group centroids: 72 and 79 genes have nonzero values on PC1 and PC2 respectively, and the percentage of explained variance has dropped from 75.6% in Fig. 6C to 71.6%.*

the few genes with nonzero values (72 for PC1 and 79 for PC2) will again be indicators of this separation. Notice the clear distinction now between groups RM and EW, and their lower within-group dispersions. In this case the percentage of variance explained by these two sparse PCA dimensions has been reduced by 4 percentage points compared to the regular PCA of the centroids in Fig. 6B.

Video 3 of the Supplementary Material shows an animation of the tumour samples in Fig. 6B transitioning to the sparse solution in Fig. 6D. The outlying genes in Fig. 6B, which contributed the most to the regular PCA solution, can be seen to be the ones not eliminated by the shrinking to zero in the sparse solution.

4.3 Correspondence analysis of categorical data

Correspondence analysis[52,53] (CA) and its constrained version, canonical correspondence analysis[54] (CCA), are among the most popular techniques for visualizing abundance or presence/absence data in ecology, but also extensively used in archaeology, linguistics and sociology. By "constraining" it is meant that the dimensions of the solution are forced to be related (usually, linearly) to external information such as groupings or explanatory variables. Interest is then focused on reducing the dimensionality of the constrained variance rather than the total variance. The analysis of Fig. 6B is, in fact, a constrained PCA, where the constraint is defined by the cancer tumour groups and the between-group variance is of interest – here the constraining variables are the four dummy variables for the tumour groups.

A typical dataset is the Barents Sea fish data[55]: 600 samples over a period of six years, 1999–2004, each obtained by 15-minute trawling in the Barents Sea north of Norway, where the numbers of up to 66 different fish species are counted in each sample. The sampling was performed at a similar time of the year and at similar locations. Such datasets are typically very sparse, since only a few fish species are found in any single sample. In this dataset, 82.6% of the values in the 600×66 data matrix are zeros.

The data to be analysed by CA are the profile vectors of relative frequencies in each row. If the original data matrix has entries n_{ij}, with row sums n_{i+} then the row profiles are the vectors of relative frequencies (proportions) $r_{ij} = n_{ij}/n_{i+}$, $j = 1, \ldots, J$. The interpoint distance function in the multidimensional profile space is the chi-square distance (see (5)), using a weighting of the squared differences between profile elements by the inverse of the average profile with elements $c_j = n_{+j}/n$, the column sums n_{+j} divided by the grand total n. The chi-square distance between two rows thus uses this standardization of the profile data: $(n_{ij}/n_{i+})/\sqrt{c_j}$, followed by the usual Euclidean distance applied to these transformed values (see Section 2.3).

The final property that distinguishes CA from PCA is that the points have weights proportional to their marginal frequencies — that is, the row weights are n_{i+}/n. CA also has the special property that it treats rows and column symmetrically. Hence, it is equivalent to think of the relative frequencies column-wise, the column profiles, as the points to be approximated in multidimensional space, with their corresponding column weights and chi-square distances between column profiles. In other words, the data table can be transposed and identical results will be obtained. This property of symmetric treatment of rows and columns is shared by logratio analysis, which was summarized briefly in Section 2.3.

Similar to the genetic study of the child cancers, there is a specific objective in analysing the Barents Sea fish data, namely to see if there is a temporal evolution of the relative fish abundances across the six years. This is achieved analytically by aggregating the fish abundances into a 6×66 matrix, where the rows are the six years and the counts are now summed for each year. The constraint is thus by the discrete variable year, with six categories. So the CA applied to this aggregated matrix is effectively a CCA, shown in Fig. 7. As before, only the top contributing variables (fish species) are shown, 10 out of the total of 66. In addition, 95% confidence ellipses are shown for the year points, but now based on 1000 bootstrap resamplings of the coordinates of the 600 samples, recomputing the year aggregations for each bootstrap sample, and then computing the ellipse for each year's set of 1000 points using the estimated bivariate normal distribution.

There appears to be a transition from 1999 on the left through to 2004 on the right, with 1999's confidence ellipse separated from the others.The biplot vectors of the species show the reason, with *Pa_bo* (*Pandalus borealis*, shrimp) being highest in 1999 and *Me_ae* (*Melanogrammus aeglefinus*, haddock) and *Tr_es* (*Trisopterus esmarkii*, Norway pout) the highest in 2004. These conclusions can be verified in the table of relative abundances: for example, for the last two species, *Me_ae* and *Tr_es*, their percentages in 2004 were 2.3% and 0.7%, more than twice the next highest relative abundances in the previous years. The difference between 1999 and 2000 appears to be due to *Bo_sa* (*Boreogadus saida*, polar cod): indeed, percentages were the highest (1.2%) in 1999 and the lowest (0.06%) in 2000.

The presence of non-overlapping confidence ellipses suggests that the temporal differences are statistically significant — this can be confirmed by a permutation test[56], which gives a p-value of 0.003. This test computes the between-year variance in the constrained space of the data, which in this case is 5-dimensional, one less than the number of years. Then, the year labels are randomly allocated to the original 600 rows of data, and the between-year variance is again computed, with this random permutation of the year labels being performed a total of 999 times. Assuming the null hypothesis of no difference between years, the obtained p-value of 0.003 means that only two between-year variances based on random allocation were greater than the observed value. These two plus the original observed value give 3 out of the 1000 in the tail of the permutation distribution, and hence the p-value indicating high significance.

4.4 Imposing external constraints

Figs 6B and 7 are both examples of constrained dimension-reduction methods, constrained to explaining between-group variance (between the cancer groups) in the PCA of Fig. 6B, and explaining differences between years in the CA of Fig. 7. Constraints

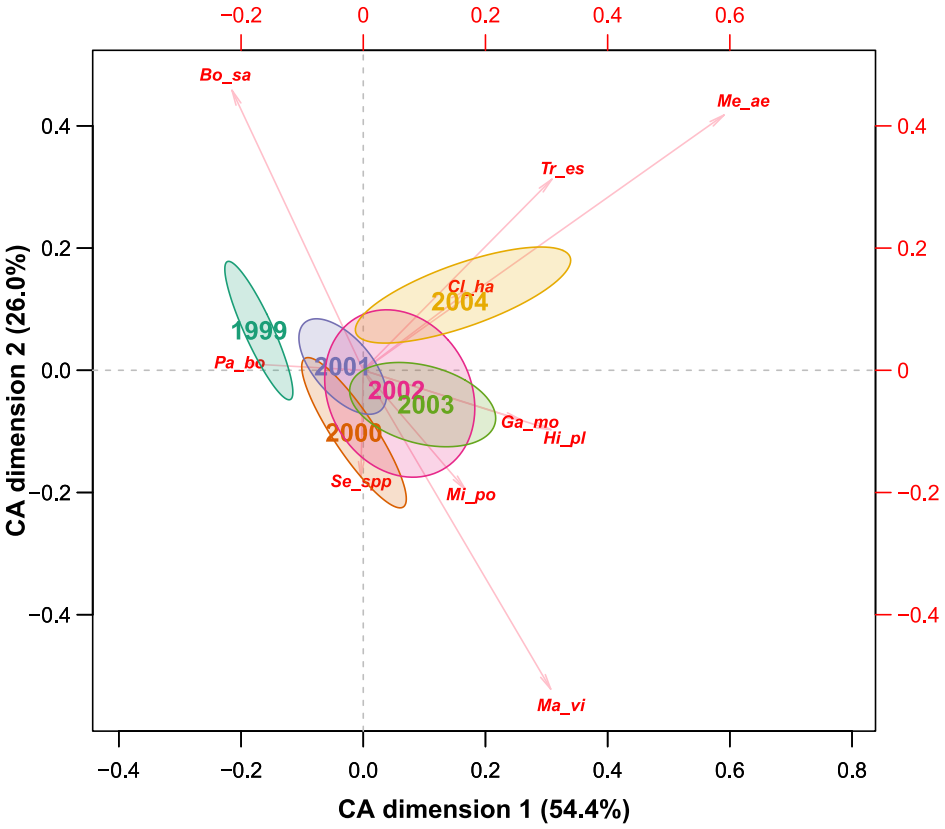

Figure 7: *CA of the Barents Sea fish data, 1999–2004, explaining the between-year variance. The year means are shown as well as their 95% confidence ellipses. The 10 species (out of 66) that contribute more than average to this two-dimensional solution are shown. Only species abbreviations are shown, with the following common names: Pa_bo (shrimp), Bo_sa (polar cod), Tr_es (Norway pout), Cl_ha (herring), Me_ae (haddock), Ga_mo (cod), Hi_pl (long rough dab), Mi_po (blue whiting), Ma_vi (capelin), Se_spp (redfish). The 600 individual sample points, which show great variation due to the sparsity of the data, are not shown.*

can be made with respect to such categorical variables as well as continuous variables, a strategy very common in ecological applications. The data matrix for the PCA is regarded as a set of response variables, for example "biological" variables such as biomasses of different marine species, where the constraining variables are "environmental" variables regarded as explanatory, such as sea temperature and salinity. Categorical variables (e.g., sampling year) are coded as dummy variables as constraining variables. Other examples are morphometric measurements on different fish, or microbial compositions, as multivariate responses, and the constraining variables could be diet variables observed on the same fish.

Rather than explain the total variance of the response dataset, the objective is to focus on that part of variance which is directly related to the explanatory variables. This is achieved through projecting the response dataset on the space defined by the explanatory variables, called the constrained or restricted space, thereby eliminating the biological variance unrelated (in a linear sense) to the environmental variables. The search for principal components is then performed in the constrained space — in the context of PCA this is known as redundancy analysis[57–59]. The result is in the form of a triplot, of cases and response variables as before, with the addition of vectors indicating directions of the continuous constraining explanatory variables or points showing the positions of the categories of constraining categorical variables, as in Figs 6B and 7.

The analogous constrained method for response data such as frequency counts or presence-absence data, which would usually be analysed using correspondence analysis, is called canonical correspondence analysis, one of the most widely used methods in quantitative ecology[54,60,61]. As well as PCA and CA, both redundancy analysis and canonical correspondence analysis are available in the R package vegan[62].

4.5 Multiple correspondence analysis

A popular variant of CA is multiple correspondence analysis (MCA)[53,63], for multivariate categorical data, often found in social surveys where respondents choose response categories in a series of questions[64–66]. The data are coded as zero–one dummy variables, where

each question generates as many dummy variables as categories, and the categories chosen by each respondent are indicated by ones in the corresponding columns. The resultant matrix is called an indicator matrix, with respondents as rows and categories as columns. Then MCA is the application of CA to the indicator matrix, generating biplots of the respondents and the categories. One advantage of this approach is that association patterns of single categories, such as "missing value/not available" or "no opinion" categories, can be investigated[67,68]. In sociological applications it is generally the averages and dispersions of respondents for different demographic categories that are of interest in the MCA results.

4.6 Mixed-scale data

Variants or extensions of PCA have been developed for different data types and structures. The observed variables could be of different types, called mixed-scale data, which often involve both continuous and categorical data. The idea is to come up with a common coding scheme, for example categorizing the continuous variables into crisp categories (dummy variable coding, zero or one) or fuzzy categories (values between zero and one), so that all the variables are of a comparable categorical type[69–72]. A general strategy, called nonlinear multivariate analysis, is to quantify categorical variables so that the resulting principal components explain as much as possible of the variance in the transformed variables[73–75].

Another context related to fuzzy category coding is where the data are intervals of real numbers; for instance, the observation of a variable is its range of values. Interval data are used to represent uncertainty or variability in observed measurements, as would be the case with monthly interval temperatures at meteorological stations, or daily interval stock prices, for example. An interval-valued observation is represented by a hyper-rectangle, rather than a point, in a low-dimensional space. Extensions of PCA for interval-valued data apply classical PCA to the centres or the vertices of the hyper-rectangles[76–83].

4.7 Derivation of scales and indices

PCA has been used to derive composite indices or composite indicators in many disciplines such as socioeconomics, public policy making, environmental and biological sciences[84–87]. A composite indicator is formed when individual indicators are compiled into a single index. For example, if we want to know the opinion on government measures aimed at the reduction of carbon dioxide, we could execute a survey and ask participants to answer a series of questions related to this topic, each answered on an ordinal rating scale. Often, the composite score is taken as the sum of all answers as the approximation of the participants' opinions on the government measures. However, how do we know that taking the direct sum is a good idea? Do all the questions measure the same or possibly different concepts? PCA can be used to do a first exploration. A single large eigenvalue is a strong indication that indeed there is a single dominant scale. Two or more large eigenvalues are indications of the presence of multiple concepts and thus more than one composite indicator. PCA can be helpful in the exploration of such composite indicators, but (confirmatory) factor analysis is recommended for the validation of such composite scales[88]. Note that MCA has also been used to construct indices based on categorical data[89,90], since the method assigns quantitative values to categories to maximize explained variance, and these summed quantifications then constitute new scales[91].

5 Reproducibility and data deposition

5.1 Minimal reporting

Reporting the results of a PCA is generally in the form of a two-dimensional biplot, where the unspoken assumption is often that this is an adequate explanation of the dataset at hand. Percentages of variance should be reported for each dimension, and by "adequate" it is not necessarily meant that the percentages explained by the two dimensions should be high. As in regression analysis, there can be a lot of noise in the data, and low percentages of variance in the leading dimensions might still reflect the only signal contained in the data.

When there are very many cases, it is often not necessary to display them all — when the cases fall into groups, showing the group means and their possible confidence regions is usually sufficient, as in Fig. 7. When there are very many variables, attention can be taken off those that make low contributions to the solution, as in Figs 6A and B, or in Fig. 7 where the low contributors are simply omitted.

As stressed earlier, to avoid distortion, the aspect ratio should be 1 in such a plot, since its interpretation is in terms of distances and perpendicular projections.

5.2 R and Python implementations

PCA is widely implemented in commercial and open-source statistical packages. In the R language, there is a large number of implementations of the PCA algorithm and its several variants. An exhaustive list of the R packages and Python libraries or PCA is beyond the scope of the paper. Box 2 shows the packages and functions that can be used to implement the methods described in this review. The graphics were generally done using base R functions — this sometimes requires more code but gives all the flexibility needed for producing publication quality results.

Box 2: Packages and functions implementing PCA and its variants

R Package	Function	Description
stats[142]	**prcomp**, **princomp**	These base R functions have minimal output and some confusing differences. Plotting is achieved with the **biplot** function, but the result is rather poor.
base	**svd**	Singular value decomposition of a matrix
FactoMineR[143]	**PCA**	These five packages all have options for weighting rows and columns of the data matrix. Options for supplementary rows and supplementary columns are provided in **PCA** (FactoMineR) and **dudi.pca**. **PCA** (easyCODA) has supplementary rows only. These three last-mentioned packages have extensive results in the created objects. The easyCODA package is aimed at compositional data analysis but has functions for PCA, CA, LRA and RDA. Most of these packages have dedicated plotting functions (in the case of the ade4 package there is a separate package adegraphics[147]).
ade4[144]	**dudi.pca**	
amap[145]	**acp**	
easyCODA[22]	**PCA**	
PCAtools[146]	**pca**	
pca3d[148]	**pca3d**	Three-dimensional PCA graphics.
vegan[62]	**rda**	This function computes redundancy analysis (RDA), that is PCA with constraints, but can also perform PCA with no constraints. The same package has function **cca** for CA with or without constraints.
elasticnet[149]	**spca**	Implementations of sparse PCA using a lasso penalized least-squares approach to obtain sparsity. **arrayspc** is specifically designed for the case $p \gg n$, such as microarrays.
	arrayspc	
irlba[150]	**prcomp_irlba**	These fast and memory efficient functions are used when the data are too large to fit in memory, or are arriving in streams.
RSpectra[151]	**svds**	
rsvd[152]	**rpca**	
onlinePCA[153]	**batchpca**	
idm[154]	**i_pca**	
symbolicDA[155]	**PCA.centers.SDA**	PCA for interval-valued data
RSDA[156]	**sym.pca**	
fdapace[157]	**FDA**	PCA of functional data, where data are sparse and longitudinal
softImpute[158]	**softImpute**	Imputation of missing values for PCA or matrix completion; can handle very large and sparse matrices
missMDA[159]	**imputePCA**	Imputation of missing values for PCA.

Python library	Function	Description
scikit-learn[160]	**sklearn.decomposition.PCA**	PCA, also with truncated SVD for large data sets
	sklearn.decomposition.SparsePCA	Sparse PCA using lasso penalty
	sklearn.decomposition.IncrementalPCA	Computes solution by processing data in chunks, when data set is too large to fit in memory
NumPy[161]	**linalg.svd**	SVD of a matrix

6 Limitations and optimizations

6.1 PCA for large datasets

When PCA is used to visualize and explore data, there are practical limitations to the data size and dimensionality that can be handled. In several applications of PCA, such as image classification[92], image compression[93], face recognition[94,95], industrial process modelling[96], quantitative finance[97], neuroscience[98], genetics and genomics[99–102], to name a few, the size and the dimensionality of the datasets can be very large and lead to computational issues. At the core of PCA there is the EVD of the covariance (or correlation) matrix, or the SVD of the centred (possibly standardized) data matrix (see Box 1). Both these matrix decompositions are computationally expensive for very large matrices and require the whole data matrix to fit in memory.

The computations for large-scale EVD and SVD can be enhanced in several ways, where a distinction can be made between batch (or offline) and incremental (or online) approaches. Most batch-enhanced matrix-decomposition methods rely on the fact that interest is usually focused on the first few eigenvalues (or singular values) and corresponding eigenvectors (or singular vectors), that is, a truncated EVD or SVD. To find the largest eigenvalue is the goal of the power method[103], and its adaptation to find the leading eigenvalues and

vectors is the Lanczos algorithm[104]. Some of the most enhanced batch EVD methods are variations of the Lanczos algorithm[105,106]. An alternative probabilistic approach leads to approximate yet accurate matrix decompositions[107].

Batch methods lead to a substantial reduction of the computational cost, but do not solve the case when the matrix cannot be stored in memory, or when new data are constantly produced (i.e., data flows). The general aim of online matrix decomposition methods is to incrementally update an existing EVD or SVD as more data come in. Several approaches have been proposed in the literature[108−112]. An incremental approach to SVD (and PCA) is best suited when the number of variables is much greater than the number of observations ($p \gg n$), and new observations become available. Examples are market basket data[113] and data from recommender systems on e-commerce websites[114] — see the section on PCA for matrix completion below. An example of the continuous arrival of image data is that from surveillance cameras[115−118], where each image is coded as a single vector, with p given by the number of pixels of that image (see an image example in Fig. 11). If nothing happens, the background corresponds to low-variance singular vectors, whereas any disturbance or intruder, however small, creates a big change.

6.2 Missing values using SVD

PCA can be extended to the case when data are partially observed. For example, suppose that 10% of the $149 \times 5 = 745$ entries in the World Happiness Report dataset were corrupted and, as a result, indicated as missing. One of the natural ideas to deal with this situation would be to remove all the rows containing missing observations and perform PCA on the fully-observed samples only. Although convenient, this approach would be very wasteful: in the worst-case scenario, as many as 50% of the 149 samples would be removed. As an alternative, one could replace missing values by the mean of the corresponding column (e.g., missing values for the variable life would be replaced by the average value for all the countries with observed values). Although widely applied in practice, this approach ignores the correlation between the variables.

To explain the goal of PCA with missing values, we first link standard PCA to the low-rank matrix approximation problem. In what follows we assume that $\mathbf{X}$ is a matrix with missing values, which has been pre-centred and pre-scaled using the observed values. As explained earlier, finding the first r principal components is equivalent to searching for the matrix $\mathbf{X}$ of rank r, denoted by $\mathbf{X}_r$, which minimizes the residual sum-of-squares (RSS) in its fit to the original data matrix. For fully-observed data, we measure RSS for all matrix elements, but when some data values are missing, we measure the RSS between the data and $\mathbf{X}_r$ using the observed values only. In this case no explicit solution exists, but the problem can be solved using a simple iterative algorithm detailed in Box 3. An example is given in the online R script of simulating 10% missing data, and finding results quite consistent with those using the complete dataset.

In the next section, more details are given about the imputation of missing data on a massive scale for high-dimensional data, and how the value of r can be inferred.

Box 3: Iterative algorithm for PCA with missing values

Step 1: Initialization for rank $r = 0$.

(a) Set $\mathbf{X}_0 = \mathbf{0}$.

(b) Replace the missing values in $\mathbf{X}$ by the corresponding values in $\mathbf{X}_0$.

(c) Compute RSS between completed $\mathbf{X}$ and $\mathbf{X}_0$ and denote it by RSS_0.

Step 2: Find solutions for ranks $r = 1, 2, \ldots, p$ in a sequential way.

(a) Iterate the following steps until convergence:

(i) Compute the first r principal components of completed $\mathbf{X}$, obtaining the rank r approximation $\mathbf{X}_r$ from the SVD (see Fig. 2 and Box 1) as follows:

$$\mathbf{X}_r = \sum_{k=1}^{r} \mathbf{u}_k \mathbf{v}_k^{\mathsf{T}}$$

(ii) Replace the missing values in $\mathbf{X}$ by the corresponding values in $\mathbf{X}_r$.

(b At convergence, compute RSS between completed $\mathbf{X}$ and $\mathbf{X}_r$ and denote it by RSS_r. The proportion of variance explained by component r can be measured by

$$(\text{RSS}_{r-1} - \text{RSS}_r)/\text{RSS}_0$$

Step 3: The proportions of variances explained by each component define the scree plot. Use it to choose a rank r^* for the final solution. Return the sample principal coordinates$\alpha_k \mathbf{u}_k$ and the variable standard coordinates $\mathbf{v}_k^{\mathsf{T}}$, for $k = 1, 2, \ldots, r^*$, which form the decomposition of $\mathbf{X}_{r^*}$.

Notice the following:

- Because of the pre-centering, Steps 1(a) and (b) amount to imputation with column means of the observed data.
- When proceeding from rank r to $r+1$ in Step 2, the completed data matrix $\mathbf{X}$ carries the filled-in values from $\mathbf{X}_r$.
- Measuring RSS between completed $\mathbf{X}$ and $\mathbf{X}_r$ is equivalent to measuring RSS using the observed values only.

6.3 Matrix completion

In the previous section an algorithm was described for making PCA work on a data matrix with missing data. Attention was not focused there on the values replacing the missing ones, but in other contexts the replaced, or imputed, values are of principal interest. A well-known recent example is that of the Netflix competition[114] where a huge dataset of 480 189 customers and 17 770 movie titles was supplied to contestants (see a tiny part in Fig. 8). On average each customer had rated about 200 movies, which means that only 1% of the matrix was observed. The task is to predict the gaps in the data, the users' ratings of movies they have not seen, based on the ratings they have supplied, and those of other users similar to them. These predictions would then be used to recommend movies to customers. Such recommender systems are widely used in online shopping and other e-commerce systems.

	My Octopus Teacher	Invictus	Tsotsi	Catching Feelings	Mandela	The Kingfisher Caper	Skin	Escape from Pretoria	District 9	Angeliena
Customer 1					4					
Customer 2			3				3			3
Customer 3		2		4					2	
Customer 4	3									
Customer 5	5	5			4					
Customer 6						2	4			
Customer 7			5					3		
Customer 8						2				3
Customer 9	3				5			5		
Customer 10										

Figure 8: *A small portion of the large data matrix* **M** *of movie ratings.*

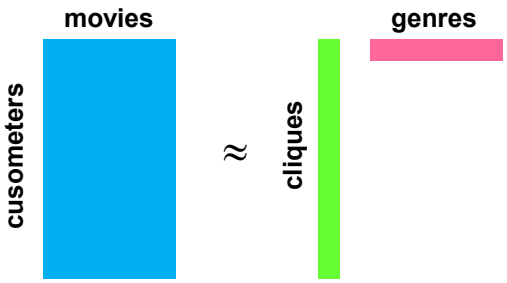

Figure 9: *The matrix factorization of the data matrix* **M** *approximately into a product of low-rank matrices,* **C** *(cliques) and the transpose of* **G** *(genres).*

A low-rank matrix approximation of the PCA type is a natural solution to such a problem (Fig. 9): $\mathbf{M} \approx \mathbf{C}\mathbf{G}^{\mathsf{T}}$. We think of movies belonging to r genres (e.g., thrillers, romance, etc, the rows of the pink matrix), and users belonging to r cliques (e.g., types who like thrillers, types who like romance, etc, the columns of the green matrix). This translates into a matrix approximation $\widehat{\mathbf{M}}$ of rank r, where the general element of the low-rank approximation is $\widehat{m}_{ij} = \sum_{k=1}^{r} c_{ik} g_{jk}$. Notice how the cliques and the genres are combined — hence, the more a customer is in a clique that favours a certain genre, the higher the predicted rating $\widehat{m}_{ij}$ will be. The objective is then to minimize the residual sum of squares (RSS) $\sum_i \sum_j (m_{ij} - \widehat{m}_{ij})^2$, that is, optimize the fit of the $\widehat{m}_{ij}$ to the m_{ij} by least squares, over the observed values in **M** only. Notice that the form of the matrix product $\mathbf{C}\mathbf{G}^{\mathsf{T}}$ is the same as the SVD of low rank (see Box 1), where the singular values have been absorbed into either the left or right singular vectors, or partially into both.

The successive filling-in algorithm for missing data described in Box 3 would be infeasible for this massive imputation task. But the basic algorithm can be significantly enhanced by introducing several computational tricks into what is called the HardImpute algorithm[119]: (i) solving the SVD problem, with filled-in values in **M**, in alternating stages by fixing the genre matrix **G** and then optimizing the fit with respect to **C**, then fixing the clique matrix **C** and optimizing with respect to **G**; (ii) storing only the observed elements of the matrix **M** (which are very few in the Netflix example compared to the elements of the whole matrix) in so-called sparse format and adapting the computations to this format; (iii) a further adaptation, called the SoftImpute algorithm[119], adding a penalty to the singular values that reduces their values, called shrinkage, so that some become zero and effectively estimate the rank of the solution.

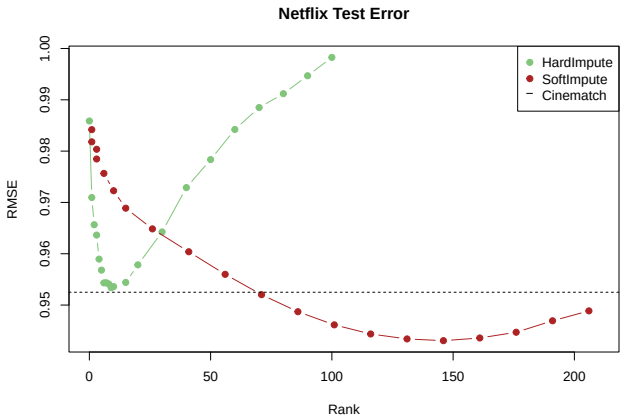

Figure 10: *Performance (root mean square error, RMSE) of HardImpute and SoftImpute on the Netflix test data. "Cinematch" was the in-house algorithm used by Netflix at the time of the competition.*

The SoftImpute algorithm is described more fully in [119–122] and has been demonstrated to give improved performance over HardImpute in many applications — see [123,124]. For the massive Netflix example, Fig. 10 shows how SoftImpute improves over HardImpute. HardImpute starts to overfit at a fairly low-rank solution, while the singular-value shrinkage in SoftImpute delays the overfitting and allows it to find signal in many more dimensions.

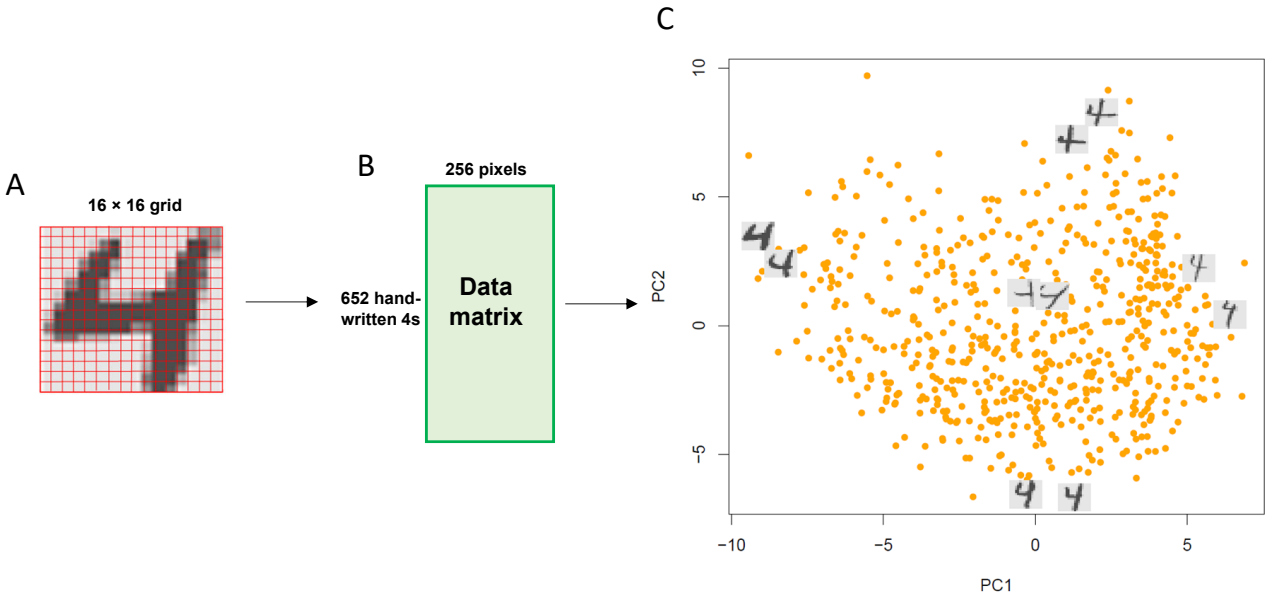

Figure 11: *A. 16×6 grid on one of 652 handwritten "4"s, for coding the gray-scale image for the 256 cells . B. Resultant 652×256 data matrix for all the "4"s. C. PCA plot of the 652 samples, showing some selected images.*

7 Outlook

PCA has been and will remain the workhorse of exploratory data analysis and unsupervised machine learning, as well as being at the heart of many real-life research problems. The future of PCA is its increasing application to a wide range of problems and sometimes quite unexpected areas of research. Here we give some recent innovations where PCA and its core algorithm, the SVD, play an important role, especially in the analysis of very large challenging datasets emanating in genetics, ecology, linguistics, business, finance and signal processing. Some of these have already been described, for example sparse PCA and matrix completion. Images, physical objects, and functions are non-standard data objects, to which PCA can be applied after using clever ways of coding the data in the form of a data matrix.

7.1 PCA of images

Often the observations represented by PCA can be rendered in a recognizable form, such as images. For example, images of birds from closely related species, images of human faces, and retinal images taken during routine eye exams. In this application we have a dataset with 652 handwritten "fours" scanned from the zip codes on letters posted in New York. Each is represented by a 16 × 16 grayscale image (see the grid in Fig. 11A, with pixel values ranging from −1 to +1). Each image of a "4" can then be coded as a single vector of length 256, thus defining a point in a 256-dimensional space, and hence the data matrix in Fig. 11B is 652 × 256. Fig. 11C shows a plot of the first two principal component scores for these data, where we have added some emblematic examples of the points to interpret the configuration: two images each that project to the extremes of PC1 and PC2, and two that project near the middle. We include their images in the plot, and they help understand what components of variation the axes explain. The PC1 axis seems to differentiate 4s with stubby tails (negative side) versus long tails (positive side). The PC2 axis (positive side) has 4s with stubby upturns in the left part of their horizontal arms, and long right arms, contrasted with the opposite pattern on the negative side.

7.2 PCA of shapes

A special case of images is that of shapes. Here we present an example in morphometrics, the study of shape. The plot in Fig. 12A shows one of the mosquito wings, with 100 landmarks indicated along the edge of a wing[125]. Each wing is thus represented by 100 pairs of (x,y) coordinates, 200 numbers in all. The data matrix for the 126 species of mosquitos studied is thus a 126 × 200 matrix of coordinates (Fig. 12B), where the wings were previously rotated while being anchored at the joint part of the wing. This fitting-together of shapes is achieved by Procrustes analysis, yet another multivariate method that relies on the singular value decomposition[126,127]. We use PCA to understand the shape variation of the wings and Fg. 12C shows the positions of the wings in a two-dimensional PCA plot, with some samples labelled at the extremes of the two PC axes. The first principal component PC1 explains 67.2% of the variance, and the plot in Fig. 12D shows all the wings in grey, the mean wing shape in black and then the two extreme wings on PC1 coloured the same

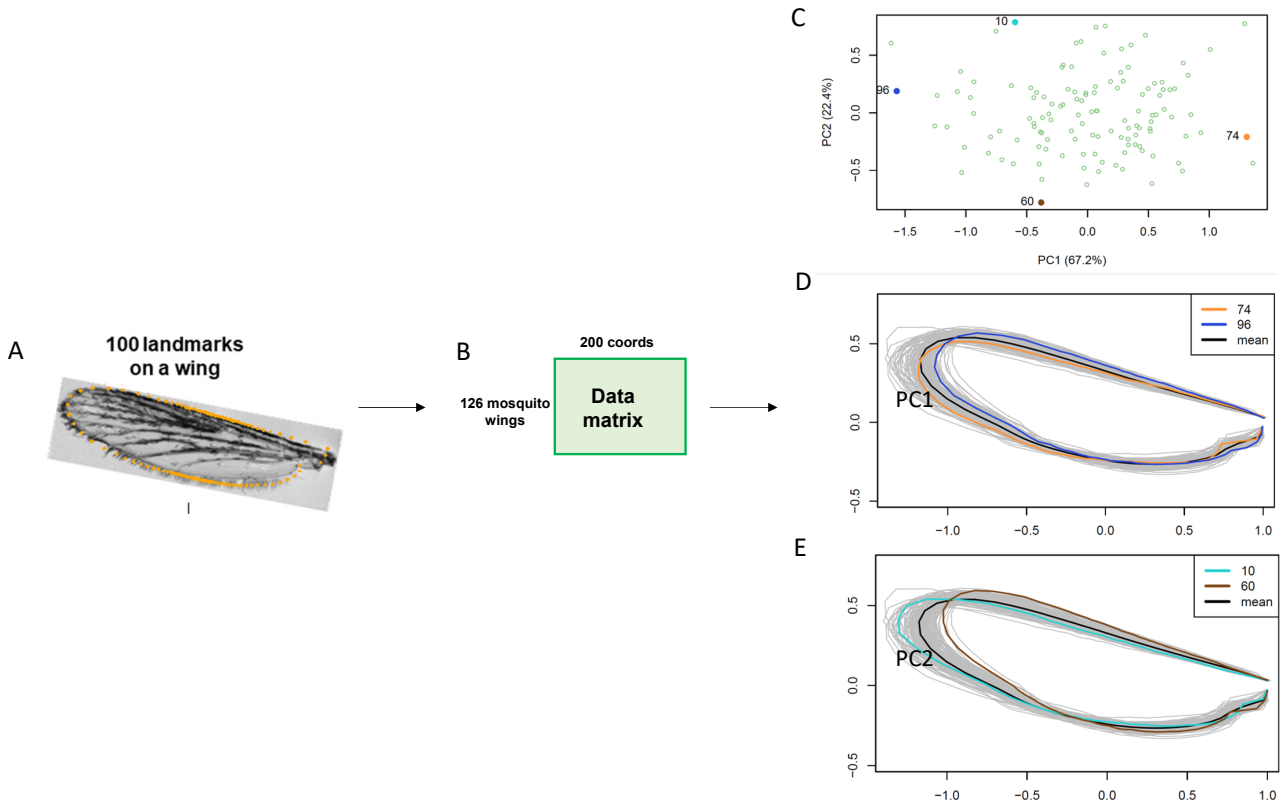

Figure 12: ***A. 100 landmarks, each with an x- and y-coordinate, shown on one of 126 mosquito wings, giving 200 values for each wing. B. The resultant 126 × 200 data matrix of coordinates. C. The PCA of the 126 wings, with some emblematic ones labelled on the extremes of the two dimensions, PC1 and PC2. D. All the wing shapes, showing the mean wing shape and the shapes of the two wings on the extremes of PC1. E. The mean wing shape and the shapes of the two wings on the extremes of PC2.***

as the dots in Fig. 12C. Fig. 12E is a similar plot for PC2. It seems that PC1 has something to do with the shape of the wing, while for PC2 the wings are more or less the same shape but different in length.

7.3 PCA of functions

Functional data are observed as smooth curves or functions. In functional PCA continuous eigenfunctions (rather than eigenvectors) are associated with the major eigenvalues. Since early work in functional PCA[128,129], there have been several developments[130−135]. Suppose that each data feature corresponds to a value of some function evaluated at different points of continuous time, say. The context presented here is the measurement of the angles of knee flexion, shown in Fig. 13A for a set of 1000 patients during a gait cycle, which is the period between successive foot contacts of the same leg. The variables are the successive values of each subject's gait curve evaluated at 100 evenly spaced times along their complete gait cycle. A patient's set of measurements is stored in a row of a 1000 ×100 matrix, and all the functions are represented as a set of curves in Fig. 13B, with the mean curve represented by the thicker black curve. Some emblematic curves are coloured and will be referred to in the next figure, Fig. 13C.

In the usual PCA of a matrix of p variables the axes form a basis in the p-dimensional space and each vector of p observations is approximated in two dimensions, say, by the mean vector (centre of the PCA plot) plus a linear combination of the first two eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$. In the case of functional data, the principal component directions are curves, so now each observed curve is approximated by the mean curve plus linear combinations of the two principal component curves. Fig. 13C shows the PCA plot of the 1000 curves, and by studying the shapes of the curves labelled as extreme points in this plot, an interpretation of what the dimensions are capturing can be suggested. The two principal component curves, shown in Fig. 13D with the same horizontal scale as Fig. 13B, give a more direct interpretation, where it should be remembered that these explain the deviations from the mean curve. (The two points close to the centre in Fig. 13C have curves similar to the mean curve in Fig. 13B). It can be deduced that PC1 is mostly a "size" component in the form of an almost constant vertical knee shift and PC2 a "shape" component in the form of a differential phase shift (PC2). Looking back at the emblematic samples in Figs 13B and 13C confirms this interpretation.

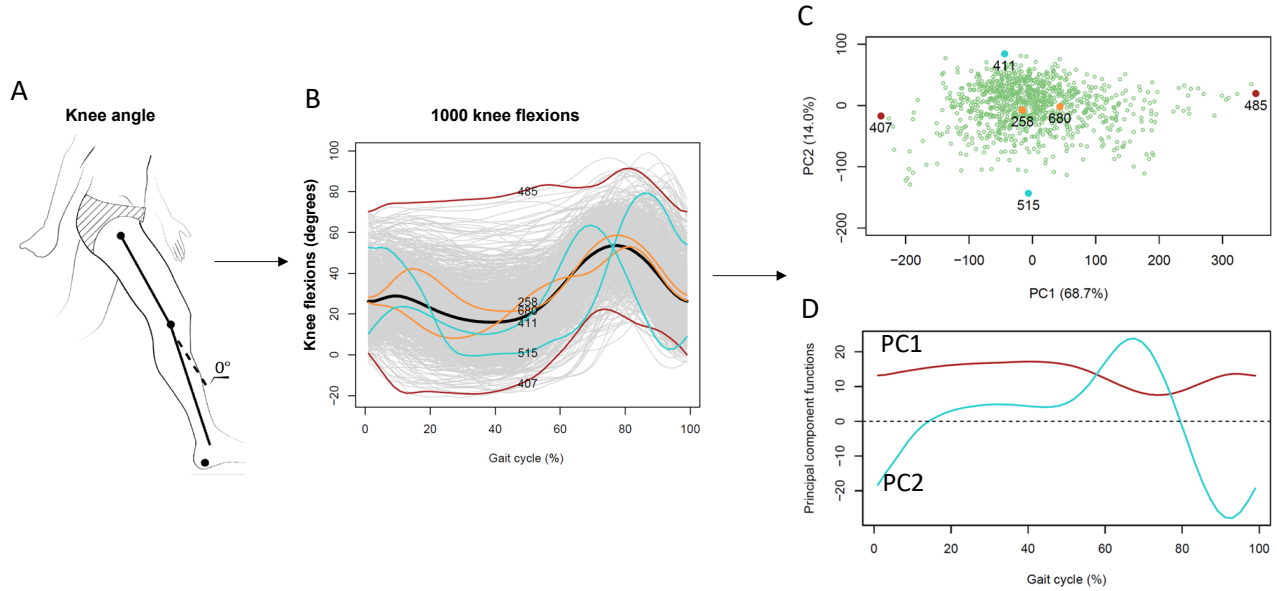

Figure 13: *A. Angle measurement of the knee. B. 1000 knee tension and extension curves, showing the mean curve as the thicker black one in the middle and some other selected ones similar and quite different from the mean (see C). C. The PCA plot of the 1000 curves, showing the positions of the ones highlighted in B. D. The first (red) and second (blue) principal component functions.*

The two principal component curves shown in Fig. 13D are the principal component "direction" vectors, plotted against percentage time in the gait cycle. They are smooth because the original data curves are smooth. Sometimes the function data are noisy, but we would still prefer smooth principal component curves for the solution. In this case we could insist that any solution curve be a linear combination of a small set of smooth functions in the columns of a matrix **S**. These smooth functions can be a basis for polynomials, sines and cosines, or splines, which are simple polynomial functions joined together smoothly to give more flexible curves. In addition, if the columns of **S** are orthonormal (which can be assumed without loss of generality), then the solution for the coefficients that combine the smooth functions can be conveniently obtained from the PCA of the matrix **XS**, where in this application **X** is the original 1000×100 data matrix[129,137].

7.4 PCA unlimited

There are many other innovative uses of PCA in the literature, which take PCA into all sorts of interesting and completely different directions, of which these are a few examples. As before, the art is in coding the appropriate variables, or features, prior to the application of PCA.

Several studies use PCA to understand the structure of songs of Humpback whales. For example, single song sessions by several whales are broken down into themes, then into phrases and finally into units. The units are then coded for various acoustic features based on the sound spectrogram, such as various harmonics and amplitudes[138]. PCA is applied to classify the songs and see their similarities in terms of times of the day and locations. In another study PCA is used to derive a complexity score based on patterns of the song, such as song length, number of units, number of unique units and average phrase length[139].

To understand the patterns of movements of mice[140], continuous three-dimensional imaging data of mice over time were subjected to wavelet decomposition and then analysed by PCA, which transformed the data into continuous trajectories through PC space. The first 10 PCs, explaining 88

Another PCA problem treats the problem of reconstructing images of three-dimensional molecules, using single-particle imaging by X-ray Free Electron Lasers. This work[141] deals with several methodological aspects of PCA that we have discussed and used in the present review: (i) alternative ways to standardization for balancing out the contributions of the image features, using the error standard deviation rather than the usual overall standard deviation; (ii) the weighting of features; and (iii) using shrinkage to determine the number of PCA dimensions.

8 Concluding remarks

PCA was one of the first multivariate analysis techniques proposed in the literature, and has since become an important and universally used tool in the understanding and exploration of data. We have presented several applications in diverse disciplines, showing how this simple and versatile method can extract the essential information from complex multivariate datasets. Recent developments and adaptations of PCA have

expanded its applicability to large datasets of many different types. More innovations of this quintessential statistical method are likely to come in future. We are convinced that PCA, along with its many variants and extensions, will remain one of the cornerstones of data science.

SUPPLEMENTARY MATERIAL

Three video animations of the PCAs of the cancer tumour data in the Applications section:

1. A three-dimensional animation of the centroid analysis of the four tumour groups.
2. A dynamic transition from the regular PCA to the PCA of the four tumour group centroids, as weight is transferred from the individual tumours to the tumour group centroids. This shows how the centroid analysis separates the groups better in the two-dimensional PCA solution, as well as how the highly contributing genes change.
3. A dynamic transition from the PCA of the group centroids to the corresponding sparse PCA solution. This shows how most genes are shrunk to the origin, and are thus eliminated, while the others are generally shrunk to the axes, which means they are contributing to only one PC. A few genes still contribute to both PCs

Several datasets and the R scripts that produce certain results in this Primer can be found on GitHub at:

`https://github.com/michaelgreenacre/PCA`

References

1. Pearson, K. On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dubl. Phil. Mag.* 2, 559–572 (1901).
2. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* 24, 417 (1933).
3. Wold, S., Esbensen, K. & Geladi, P. Principal component analysis. *Chemometr. Intell. Lab. Syst.* 2, 37–52 (1987).
4. Jackson, J. Edward. *A User's Guide to Principal Components.* (Wiley, 1991).
5. Jolliffe, I. T. *Principal Component Analysis (2nd Edition)*. (Springer, 2002).
6. Ringnér, M. What is principal component analysis? *Nat. Biotechnol.* 26, 303-–304 (2008).
7. Abdi, H. & Williams, L. J. Principal component analysis. *WIREs Comp. Stat.* 2, 433—459 (2010).
8. Bro, R. & Smilde, A.K. Principal component analysis. *Anal. Methods* 6, 2812—2831 (2014).
9. Jolliffe, I.T. & Cadima, J. *Principal component analysis: a review and recent developments. Philos. Trans. R. Soc. A* 374, 20150202 (2016).
10. Helliwell, J. F., Huang, H., Wang, S. & Norton, M. World happiness, trust and deaths under COVID-19. *World Happiness Report* 13–56 (2021).
11. Cantril, H. *Pattern of human concerns.* (Rutgers University Press, 1965).
12. Flury, B. D. Developments in principal component analysis. in *Recent Advances in Descriptive Multivariate Analysis* (ed. Krzanowski, W. J.) 14–33 (Clarendon Press, 1995).
13. Gabriel, K.R. The biplot graphic display of matrices with application to principal component analysis. *Biometrika* 58, 453–467 (1971).
14. Gower, J.C. & Hand, D. J. *Biplots.* (Chapman & Hall, 1995).
15. Greenacre, M. *Biplots in Practice.* (BBVA Foundation, 2010).
16. Stephens, G. & Greenacre, M. *It had to be U – the SVD song*. YouTube at https://www.youtube.com/watch?v=JEYLfIVvR9I (2012).
17. Greenacre, M. Contribution Biplots. *J. Comput. Graph. Stat.* 22, 107–122 (2013).
18. Eckart, C. & Young, G. The approximation of one matrix by another of lower rank. *Psychometrika* 1, 211–218 (1936).
19. Greenacre, M., Martínez-Álvaro, M. & Blasco, A. Compositional data analysis of microbiome and any-omics datasets: a validation of the additive logratio transformation. *Front. Microbiol.* 12, 727398 (2021).
20. Greenacre, M. Compositional Data Analysis. *Annu. Rev. Stat. Appl.* 8, 271–299 (2021).
21. Aitchison, J. & Greenacre, M. Biplots of compositional data. *J. R. Stat. Soc. Ser. C* 51, 375–392 (2002).
22. Greenacre, M. *Compositional Data Analysis in Practice.* (Chapman & Hall / CRC Press, 2018).
23. Cattell, R.B. The scree test for the number of factors. *Multivar. Behav. Res.* 1, 245–276 (1966).
24. Jackson, D.A. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology* 74, 2204–2214 (1993).
25. Peres-Neto, P.R., Jackson, D A. & Somers, K.A. How many principal components? stopping rules for determining the number of non-trivial axes revisited. *Comput. Stat. Data Anal.* 49, 974–997 (2005).
26. Auer, P. & Gervini, D. Choosing principal components: a new graphical method based on Bayesian model selection. *Commun. Stat.-Simul. C.* 37, 962–977 (2008).
27. Cangelosi, R. & Goriely, A. Component retention in principal component analysis with application to cDNA microarray data. *Biol. Direct.* 2, 1–21 (2007).
28. Josse, J. & Husson, F. Selecting the number of components in principal component analysis using cross-validation approximations. *Comput. Stat. Data Anal.* 56, 1869–1879 (2012).
29. Choi, Y., Taylor, J. & Tibshirani, R. Selecting the number of principal components: Estimation of the true rank of a noisy matrix. *Ann. Stat.* 259—2617 (2017).
30. Wang, M., Kornblau, S.M. & Coombes, K.R. Decomposing the apoptosis pathway into biologically interpretable principal components. *Cancer Inform.* 17, 1176935118771082 (2018).
31. Greenacre, M. & Degos, L. Correspondence analysis of HLA gene frequency data from 124 population samples. *Am. J. Hum. Genet.* 29, 60–75 (1977).
32. Borg, I. & Groenen, P.J.F. *Modern Multidimensional Scaling: Theory and Applications.* (Springer Science & Business Media, 2005).
33. Khan, J. et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.* 7, 673–679 (2001).
34. Hastie, T., Tibshirani, R. & Friedman, J. H. *The Elements of Statistical Learning Data Mining, Inference, and Prediction.*

(Springer, 2009).
35. Greenacre, M. Data reporting and visualization in ecology. *Polar Biol.* 39, 2189–2205 (2016).
36. Fisher, R.A. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* 7, 179–188 (1936).
37. Campbell, N.A. & Atchley, W.R. The geometry of canonical variate analysis. *Syst. Zool.* 30, 268–280 (1981).
38. Jolliffe, I.T. Rotation of principal components: choice of normalization constraints. *J. Appl. Stat.* 22, 29–35 (1995).
39. Cadima, J.F.C.L. & Jolliffe, I. T. Loadings and correlations in the interpretation of principal components. *J. Appl. Stat.* 22, 203–214 (1995).
40. Jolliffe, I.T., Trendafilov, N.T.T. & Uddin, M. A modified principal component technique based on the LASSO. *J. Comput. Graph. Stat.* 531–547 (2003).
41. Zou, H., Hastie, T. & Tibshirani, R. Sparse Principal Component Analysis. *J. Comput. Graph. Stat.* 15, 265–286 (2006).
42. Shen, H. & Huang, J. Z. Sparse principal component analysis via regularized low rank matrix approximation. *J. Multivar. Anal.* 99, 1015–1034 (2008).
43. Witten, D. M., Tibshirani, R. & Hastie, T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* 10, 515–534 (2009).
44. Journée, M., Nesterov, Y., Richtárik, P. & Sepulchre, R. Generalized Power Method for Sparse Principal Component Analysis. *J. Mach. Learn. Res.* 11, 517–553 (2010).
45. Papailiopoulos, D., Dimakis, A. & Korokythakis, S. Sparse PCA through low-rank approximations. in Proc. Mach. Learn. Res. 747–755 (2013). 46. Erichson, N. B. et al. Sparse principal component analysis via variable projection. SIAM J. Appl. Math. 80, 977–1002 (2020). 47. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* 58, 267–288 (1996).
48. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* 67, 301–320 (2005).
49. Guerra-Urzola, R., van Deun, K., Vera, J. C. & Sijtsma, K. A guide for sparse PCA: Model comparison and applications. *Psychometrika* 86, 893–919 (2021).
50. Camacho, J., Smilde, A. K., Saccenti, E. & Westerhuis, J. A. All sparse PCA models are wrong, but some are useful. Part I: computation of scores, residuals and explained variance. *Chemometr. Intell. Lab. Syst.* 196, 103907 (2020).
51. Camacho, J., Smilde, A. K., Saccenti, E., Westerhuis, J. A. & Bro, R. All sparse PCA models are wrong, but some are useful. Part II: Limitations and problems of deflation. *Chemometr. Intell. Lab. Syst.* 208, 104212 (2021).
52. Benzécri, J.-P. *Analyse des Données, Tôme 2: Analyse des Correspondances.* (Dunod, 1973).
53. Greenacre, M. Correspondence Analysis in Practice (3rd edition). (Chapman & Hall / CRC Press, 2016). 54. ter Braak, C. J. F. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology* 67, 1167–1179 (1986).
55. Greenacre, M. & Primicerio, R. *Multivariate Analysis of Ecological Data.* (Fundacion BBVA Publication Bilbao, Spain, 2013).
56. Good, P. *Permutation Tests: a Practical Guide to Resampling Methods for Testing Hypotheses.* (Springer Science & Business Media, 1994).
57. Legendre, P. & Anderson, M. J. Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecol. Monogr.* 69, 1–24 (1999).
58. van den Wollenberg, A. L. Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika* 42, 207–219 (1977).
59. Capblancq, T. & Forester, B. R. Redundancy analysis: A Swiss Army Knife for landscape genomics. *Methods Ecol. Evol.* 12, 2298–2309 (2021).
60. Palmer, M. W. Putting things in even better order: the advantages of canonical correspondence analysis. *Ecology* 74, 2215–2230 (1993).
61. ter Braak, C. J. F. & Verdonschot, P. F. M. Canonical correspondence analysis and related multivariate methods in aquatic ecology. *Aquat. Sci.* 57, 255–289 (1995).
62. Oksanen, J. et al. vegan: Community Ecology Package. URL: https://CRAN.R-project.org/package=vegan (2022).
63. Abdi, H. & Valentin, D. Multiple Correspondence Analysis. *Encycl. Meas. Stat.* 2, 651–657 (2007).
64. Richards, G. & van der Ark, L. A. Dimensions of cultural consumption among tourists: Multiple correspondence analysis. *Tour. Manag.* 37, 71–76 (2013).
65. Glevarec, H. & Cibois, P. Structure and historicity of cultural tastes. Uses of multiple correspondence analysis and sociological theory on age: The case of music and movies. *Cult. Sociol.* 15, 271–291 (2021).
66. Jones, I. R., Papacosta, O., Whincup, P. H., Goya Wannamethee, S. & Morris, R. W. Class and lifestyle 'lock-in' among middle-aged and older men: a Multiple Correspondence Analysis of the British Regional Heart Study. Sociol. *Health Illn.* 33, 399–419 (2011).
67. Greenacre, M. & Pardo, R. Subset correspondence analysis: visualizing relationships among a selected set of response categories from a questionnaire survey. *Sociol. Methods Res.* 35, 193–218 (2006).
68. Greenacre, M. & Pardo, R. Multiple correspondence analysis of subsets of response categories. in Multiple Correspondence Analysis and Related Methods (eds. Greenacre, M. & Blasius, J.) 197–217 (Chapman & Hall/CRC Press, 2008).
69. Aşan, Z. & Greenacre, M. Biplots of fuzzy coded data. *Fuzzy Sets Syst.* 183, 57–71 (2011).
70. Vichi, M., Vicari, D. & Kiers, H. A. L. Clustering and dimension reduction for mixed variables. *Behaviormetrika* 46, 243–269 (2019).
71. van de Velden, M., Iodice D'Enza, A. & Markos, A. Distance-based clustering of mixed data. *Wiley Interdiscip. Rev. Comput. Stat.* 11, e1456 (2019).
72. Greenacre, M. Use of Correspondence Analysis in Clustering a Mixed-Scale Data Set with Missing Data. *Arch. Data Sci. Ser.* B 1, 1–12 (2019).
73. Gifi, A. *Nonlinear multivariate analysis.* (Wiley-Blackwell, 1990).
74. Michailidis, G. & de Leeuw, J. The Gifi system of descriptive multivariate analysis. *Stat. Sci.* 307–336 (1998).
75. Linting, M., Meulman, J. J., Groenen, P. J. F. & van der Koojj, A. J. Nonlinear principal components analysis: Introduction and application. *Psychol. Methods* 12, 336 (2007).
76. Cazes, P., Chouakria, A., Diday, E. & Schektman, Y. Extension de l'analyse en composantes principales à des données de type intervalle. *Rev. Stat. Appl.* 45, 5–24 (1997).
77. Bock H.-H. and Chouakria, A. and C. P. & D. E. Symbolic principal component analysis. in *Analysis of Symbolic Data* (eds Bock H.-H. and Diday, E.) 200–212 (Springer Berlin Heidelberg, 2000).
78. Lauro, C. N. & Palumbo, F. Principal component analysis of interval data: a symbolic data analysis approach. *Comput. Stat.* 15, 73–87 (2000).

79. Gioia, F. & Lauro, C. N. Principal component analysis on interval data. *Comput. Stat.* 21, 343–363 (2006).
80. Giordani, P. & Kiers, H. A comparison of three methods for principal component analysis of fuzzy interval data. *Comput. Stat. Data Anal.* 51, 379–397 (2006).
81. Makosso-Kallyth, S. & Diday, E. Adaptation of interval PCA to symbolic histogram variables. *Adv. Data Anal. Classif.* 6, 147–159 (2012).
82. Brito, P. Symbolic data analysis: another look at the interaction of data mining and statistics. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 4, 281–295 (2014).
83. Le-Rademacher, J. & Billard, L. Principal component analysis for histogram-valued data. *Adv. Data Anal. Classif.* 11, 327–351 (2017).
84. Booysen, F. An overview and evaluation of composite indices of development. *Social Indic. Res.* 59, 115–151 (2002).
85. Lai, D. Principal component analysis on human development indicators of China. *Social Indic. Res.* 61, 319–330 (2003).
86. Krishnakumar, J. & Nagar, A. L. On exact statistical properties of multidimensional indices based on principal components, factor analysis, MIMIC and structural equation models. *Social Indic. Res.* 86, 481–496 (2008).
87. Mazziotta, M. & Pareto, A. Use and misuse of PCA for measuring well-being. *Social Indic. Res.* 142, 451–476 (2019).
88. Fabrigar, L. R., Wegener, D. T., MacCallum, R. C. & Strahan, E. J. Evaluating the use of exploratory factor analysis in psychological research. *Psychol. Methods* 4, 272 (1999).
89. Booysen, F., van der Berg, S., Burger, R., von Maltitz, M. & du Rand, G. Using an asset index to assess trends in poverty in seven Sub-Saharan African countries. *World Dev.* 36, 1113–1130 (2008).
90. Wabiri, N. & Taffa, N. Socio-economic inequality and HIV in South Africa. *BMC Public Health* 13, 1–10 (2013).
91. Lazarus, J. v et al. The global NAFLD policy review and preparedness index: Are countries ready to address this silent public health challenge? *J. Hepatol.* 76, 771–780 (2022).
92. Rodarmel, C. & Shan, J. Principal component analysis for hyperspectral image classification. *Surv. Land Inf. Sci.* 62, 115–122 (2002).
93. Du, Q. & Fowler, J. E. Hyperspectral image compression using JPEG2000 and principal component analysis. IEEE Geosci. *Remote Sens. Lett.* 4, 201–205 (2007).
94. Turk, M. & Pentland, A. Eigenfaces for recognition. J. *Cognit. Neurosci.* 3, 71–86 (1991).
95. Paul, L. & Suman, A. Face recognition using principal component analysis method. *Int. J. Adv. Res. Comput. Eng. Technol.* 1, 135–139 (2012).
96. Zhu, J., Ge, Z., Song, Z. & Gao, F. Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data. *Annu. Rev. Control* 46, 107–133 (2018).
97. Ghorbani, M. & Chong, E. K. P. Stock price prediction using principal components. *PLoS One* 15, e0230124 (2020).
98. Pang, R., Lansdell, B. J. & Fairhall, A. L. Dimensionality reduction in neuroscience. *Curr. Biol.* 26, R656–R660 (2016).
99. Abraham, G. & Inouye, M. Fast principal component analysis of large-scale genome-wide data. *PLoS One* 9, e93766 (2014).
100. Alter, O., Brown, P. O. & Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci.* 97, 10101–10106 (2000).
101. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* 2, e190 (2006).
102. Tsuyuzaki, K., Sato, H., Sato, K. & Nikaido, I. Benchmarking principal component analysis for large-scale single-cell RNA-sequencing. *Genome Biol.* 21, 1–17 (2020).
103. Golub, G. H. & van Loan, C. F. *Matrix Computations.* (JHU press, 2013).
104. Lanczos, C. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J.Res. Nat. Bur. Stand.* 45, 255–282 (1950).
105. Baglama, J. & Reichel, L. Augmented GMRES-type methods. Numer.*Linear Algebra Appl.* 14, 337–350 (2007).
106. Wu, K. & Simon, H. Thick-restart Lanczos method for large symmetric eigenvalue problems. *SIAM J. Matrix Anal. Appl.* 22, 602–616 (2000).
107. Halko, N., Martinsson, P.-G. & Tropp, J. A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.* 53, 217–288 (2011).
108. Weng, J., Zhang, Y. & Hwang, W.-S. Candid covariance-free incremental principal component analysis. IEEE Trans. Pattern Anal. *Mach. Intell.* 25, 1034–1040 (2003).
109. Ross, D. A., Lim, J., Lin, R.-S. & Yang, M.-H. Incremental learning for robust visual tracking. *Int. J. Comput. Vis.* 77, 125–141 (2008).
110. Cardot, H. & Degras, D. Online principal component analysis in high dimension: Which algorithm to choose? *Int. Stat. Rev.* 86, 29–50 (2018).
111. Iodice D'Enza A. & Greenacre, M. Multiple Correspondence Analysis for the Quantification and Visualization of Large Categorical Data Sets. In *Advanced Statistical Methods for the Analysis of Large Data Sets* (eds di Ciaccio Agostino and Coli, M. & A. I. J.-M.) 453–463 (Springer Berlin Heidelberg, 2012).
112. Iodice D'Enza, A., Markos, A. & Palumbo, F. Chunkwise regularised PCA-based imputation of missing data. *Stat. Methods Appl.* 1–22 (2021).
113. Shiokawa, Y., Kikuchi, J. & others. Application of kernel principal component analysis and computational machine learning to exploration of metabolites strongly associated with diet. *Sci. Rep.* 8, 1–8 (2018).
114. Koren, Y., Bell, R. & Volinsky, C. Matrix factorization techniques for recommender systems. *Computer* (Long Beach Calif) 42, 30–37 (2009).
115. Li, Y. On incremental and robust subspace learning. *Pattern Recognit.* 37, 1509–1518 (2004).
116. Bouwmans, T. Subspace learning for background modeling: A survey. *Recent Pat. Comput. Sci.* 2, 223–234 (2009).
117. Guyon, C., Bouwmans, T. & Zahzah, E.-H. Foreground detection via robust low rank matrix decomposition including spatio-temporal constraint. In *Asian Conf. Computer Vision* (eds Park Jong-Il and Kim, J.) 315–320 (Springer Berlin Heidelberg, 2012).
118. Bouwmans, T. & Zahzah, E. H. Robust PCA via principal component pursuit: A review for a comparative evaluation in video surveillance.*Comput. Vis. Image Underst.* 122, 22–34 (2014).
119. Mazumder, R., Hastie, T. & Tibshirani, R. Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.* 11, 2287–2322 (2010).
120. Josse, J. & Husson, F. Handling missing values in exploratory multivariate data analysis methods. *J. Soc. Fr. Stat.* 153, 79–99 (2012).

121. Hastie, T., Tibshirani, R. & Wainwright, M. *Statistical learning with sparsity: the LASSO and generalizations.* (Boca Raton, FL: CRC Press, 2015).
122. Hastie, T., Mazumder, R., Lee, J. D. & Zadeh, R. Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares. *J. Mach. Learn. Res.* 16, 3367–3402 (2015).
123. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J.-P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* 9, 1–17 (2018).
124. Ioannidis, A. G. et al. Paths and timings of the peopling of Polynesia inferred from genomic networks. *Nature* 597, 522–526 (2021).
125. Rohlf, F. J. & Archie, J. W. A comparison of Fourier methods for the description of wing shape in mosquitoes (Diptera: Culicidae). *Syst. Zool.* 33, 302–317 (1984).
126. Gower, J. C. Generalized Procrustes analysis. *Psychometrika* 40, 33–51 (1975).
127. Dryden, I. L. & Mardia, K. V. *Statistical Shape Analysis: With Applications in R (2nd edition)*. (John Wiley & Sons, 2016).
128. Ocaña, F. A., Aguilera, A. M. & Valderrama, M. J. Functional principal components analysis by choice of norm. *J. Multivar. Anal.* 71, 262–276 (1999).
129. Ramsay, J. O. & Silverman, B. W. Principal components analysis for functional data. In *Functional Data Analysis* 147–172 (Springer, 2005).
130. James, G. M., Hastie, T. J. & Sugar, C. A. Principal component models for sparse functional data. *Biometrika* 87, 587–602 (2000).
131. Yao, F., Müller, H.-G. & Wang, J.-L. Functional data analysis for sparse longitudinal data. *J. Am. Stat. Assoc.* 100, 577–590 (2005).
132. Hörmann, S., Kidziński, Ł. & Hallin, M. Dynamic functional principal components. *J. R. Stat. Soc. Ser. B* 77, 319–348 (2015).
133. Bongiorno, E. G. & Goia, A. Describing the concentration of income populations by functional principal component analysis on Lorenz curves. *J. Multivar. Anal.* 170, 10–24 (2019).
134. Li, Y., Huang, C. & Härdle, W. K. Spatial functional principal component analysis with applications to brain image data. *J. Multivar. Anal.* 170, 263–274 (2019).
135. Song, J. & Li, B. Nonlinear and additive principal component analysis for functional data. *J. Multivar. Anal.* 181, 104675 (2021).
136. Kidziński, Ł. et al. Deep neural networks enable quantitative movement analysis using single-camera videos. *Nat. Commun.* 11, 1–10 (2020).
137. Tuzhilina, E., Hastie, T. J. & Segal, M. R. Principal curve approaches for inferring 3D chromatin architecture. *Biostatistics* 23, 626–642 (2022).
138. Maeda, H., Koido, T. & Takemura, A. Principal component analysis of song units produced by humpback whales (Megaptera novaeangliae) in the Ryukyu region of Japan. *Aquat. Mamm.* 26, 202–211 (2000).
139. Allen, J. A., Garland, E. C., Garrigue, C. & et al. Song complexity is maintained during inter population cultural transmission of humpback whale songs. *Sci. Rep.* 12, 8999 (2022).
140. Wiltschko, A. B., Johnson, M. J., Iurilli, G. & et al. Mapping sub-second structure in mouse behavior. *Neuron* 88, 1121–1135 (2015).
141. Liu, L. T., Dobriban, E. & Singer, A. ePCA: high dimensional exponential family PCA. *Ann. Appl. Stat.* 12, 2121–2150 (2018).
142. R Core Team. R: A Language and Environment for Statistical Computing. URL: https://www.R-project.org/ (2021).
143. Lê, S., Josse, J. & Husson, F. FactoMineR: an R package for multivariate analysis. *J. Stat. Softw.* 25, 1–18 (2008).
144. Thioulouse, J. et al.*Multivariate Analysis of Ecological Data with ade4.* (Springer, 2018).
145. Lucas, A. amap: Another Multidimensional Analysis Package. URL: https://CRAN.R-project.org/package=amap (2019).
146. Blighe, K. & Lun, A. PCAtools: PCAtools: Everything Principal Components Analysis. URL: https://github.com/kevinblighe/PCAtools (2021).
147. Siberchicot, A., Julien-Laferrière, A., Dufour, A.-B., Thioulouse, J. & Dray, S. adegraphics: an S4 lattice-based package for the representation of multivariate data. *R J.* 9, 198–212 (2017).
148. Weiner, J. pca3d: Three Dimensional PCA Plots. URL: https://CRAN.R-project.org/package=pca3d (2020).
149. Zou, H. & Hastie, T. elasticnet: Elastic-Net for Sparse Estimation and Sparse PCA. URL: https://CRAN.R-project.org/package=elasticnet (2020).
150. Baglama, J., Reichel, L. & Lewis, B. W. irlba: Fast Truncated Singular Value Decomposition and Principal Components Analysis for Large Dense and Sparse Matrices. R package version 2.3.5 URL: https://CRAN.R-project.org/package=irlba (2021).
151. Qiu, Y. & Mei, J. RSpectra: Solvers for Large-Scale Eigenvalue and SVD Problems. URL: https://CRAN.R-project.org/package=RSpectra (2019).
152. Erichson, N. B., Voronin, S., Brunton, S. L. & Kutz, J. N. Randomized Matrix Decompositions Using R. *J. Stat. Softw.* 89, 1–48 (2019).
153. Degras, D. & Cardot, H. onlinePCA: Online Principal Component Analysis. URL: https://CRAN.R-project.org/package=onlinePCA (2016).
154. Iodice D'Enza, A., Markos, A. & Buttarazzi, D. The idm package: incremental decomposition methods in R. *J. Stat. Softw.* 86, 1–24 (2018).
155. Dudek, A., Pelka, M., Wilk, J. & Walesiak, M. symbolicDA: Analysis of Symbolic Data. URL: https://CRAN.R-project.org/package=symbolicDA (2019).
156. Rodriguez, O. RSDA: R to Symbolic Data Analysis. URL: https://CRAN.R-project.org/package=RSDA (2021).
157. Gajardo, A. et al. fdapace: Functional Data Analysis and Empirical Dynamics. URL: https://CRAN.R-project.org/package=fdapace (2021).
158. Hastie, T. & Mazumder, R. softImpute: Matrix Completion via Iterative Soft-Thresholded SVD. URL: https://CRAN.R-project.org/package=softImpute (2021).
159. Josse, J. & Husson, F. missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. *J. Stat. Softw.* 70, 1–31 (2016).
160. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830 (2011).
161. Harris, C. R. et al. Array programming with NumPy. *Nature* 585, 357–362 (2020).