nature reviews methods primers

https://doi.org/10.1038/s43586-022-00184-w

Primer

Check for updates

Principal component analysis

Michael Greenacre[1], Patrick J. F. Groenen[2], Trevor Hastie[3], Alfonso Iodice D'Enza[4], Angelos Markos[5] & Elena Tuzhilina[6]

Abstract

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. This Primer presents 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. Also described and illustrated are the latest innovative applications of principal component analysis: 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.

Sections

[1]Department of Economics and Business, Universitat Pompeu Fabra and Barcelona School of Management, Barcelona, Spain. [2]Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam, Netherlands. [3]Departments of Statistics and Biomedical Science, Stanford University, Stanford, CA, USA. [4]Department of Political Sciences, University of Naples Federico II, Naples, Italy. [5]Department of Primary Education, Democritus University of Thrace, Alexandroupolis, Greece. [6]Department of Statistics, Stanford University, Stanford, CA, USA. e-mail: michael.greenacre@upf.edu

Introduction

Principal component analysis[1–9] (PCA) is a multivariate statistical method that combines 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/territories. 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, five indicators are considered: 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. If the data were random, with 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/territories. The following linear combination achieves this objective, and it defines the first principal component, PC1:

$$\begin{aligned}\mathrm{PC1} = 0.538\ \textit{Social} + 0.563\ \textit{Life} + 0.498\ \textit{Choices}\\ -0.004\ \textit{Generosity} - 0.381\ \textit{Corruption}\end{aligned} \tag{1}$$

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. As a result, their total variance is 5. Thanks to this standardization, the coefficients of the variables, sometimes called loadings, indicate the strength of contribution to the principal component, while their signs indicate whether the influence is positive or negative. PC1 can also be considered the closest correlate to all five variables. In other words, PC1 is a single-variable summary of what the original indicators 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 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, meaning that PC1 has explained 47.0% of the total variance.

As 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:

$$\begin{aligned}\mathrm{PC2} = -0.266\ \textit{Social} - 0.243\ \textit{Life} + 0.258\ \textit{Choices}\\ +0.799\ \textit{Generosity} - 0.407\ \textit{Corruption}\end{aligned} \tag{2}$$

A condition in finding PC2 is that it should be uncorrelated with PC1, so 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. The overall explained variance is 0.715, that is, 71.5%. PC2 has, therefore, explained an additional 24.5% of the variance.

The maximum number of PCs is equal to 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 in Eqs. 1 and 2 can be computed for each of the 149 countries/territories and plotted in a scatterplot (Fig. 1). The countries/territories are classified into ten regions, so the positions of the regional averages can also be shown.

The signs of the coefficients are indeterminate. Different computer algorithms can produce the negative of PC1 or PC2, with all the signs reversed, resulting in an equivalent interpretation, but in the opposite direction. The user is at liberty to multiply any principal component by −1, which inverts the corresponding axis in Fig. 1, to facilitate the interpretation.

This visualization in Fig. 1 shows the approximate positions of the countries/territories in terms of all five variables condensed into the two principal components. The countries/territories are spread out in the two-dimensional plot as much as possible, maximizing the variance. In the following sections, interpretation of the country/territory positions will be facilitated by showing the variables themselves in the display, alongside any other variables observed on the countries/territories, such as economic indicators.

Experimentation

PCA workflow

Step 1: standardization of variables. The first and most important step in the PCA workflow is to make a decision about 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* was measured in years, *Generosity* 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 owing to its high numerical range of years. If no adjustment is 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. Dividing each variable's values by the respective standard deviation is sufficient for removing the scale effect. 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 done here for the five variables. The contributions of these variables to the total variance are thus equalized, irrespective of the possible differences in their substantive importance for the research question. As a general rule, software for PCA does not include automatic standardization of the variables. 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. Sometimes, prestandardization is not necessary[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 variable standardization, which gives comparable additive scales to reflect multiplicative differences in the variables, meaning that 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 the 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 Eqs. 1 and 2 are the elements of the two eigenvectors corresponding to the two highest eigenvalues. The eigenvalues themselves are the parts of the 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. 1a and b are λ_1 and λ_2 as percentages of the sum of all five eigenvalues.

The lower pathway in Fig. 2 shows the more efficient computational workflow. The singular value decomposition (SVD), a generalization of the EVD to arbitrary rectangular matrices, is applied directly to the matrix, which is optionally standardized, but at least centred. This results 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 Fig. 1a and b. 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, also summarized in this musical illustration of the SVD. As indicated in the notes in Box 1, an alternative way of making a biplot is to leave the left singular vectors unscaled and to 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.

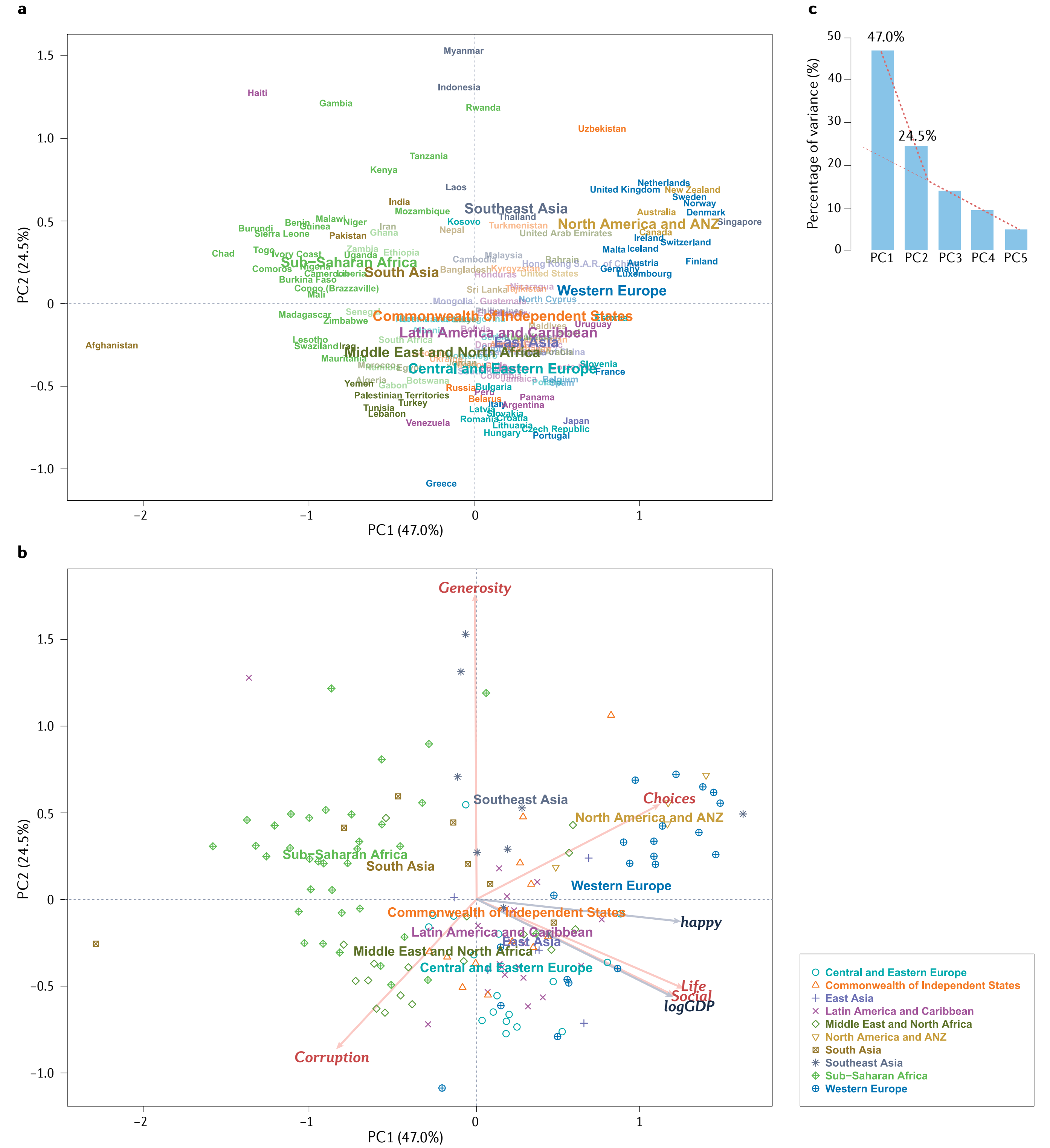
Step 3: scaling and interpretation of the biplot. The biplot of the data from the 149 countries/territories is shown in Fig. 1b. The countries/territories are in the same positions as the scatterplot in Fig. 1a, but now use symbols to make the display less cluttered. Their coordinates are obtained either by computing the linear combinations originally defined as the principal components in Eqs. 1 and 2 for each country/territory, or equivalently using the left singular vectors scaled by the singular values. The arrows are defined by the pairs of coefficients in the two linear combinations. For example, the vector *Social* has coordinates [0.538, −0.266] in Fig. 1b, according to the scale on the axes for the variables; see Eqs. 1 and 2.

The five variable directions define biplot axes onto which the countries/territories can be projected perpendicularly. The means of the variables are all at the origin owing to the data centring and the arrows indicate increasing values. Therefore, when two variables point in the same direction, such as *Life* and *Social*, countries/territories will project similarly onto them, suggesting 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 the projections of the countries/territories 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 high percentage of total variance.

Although the spatial interpretation with respect to the variable direction is most important, it is often possible to interpret the principal component directions themselves, namely the dimensions, called principal axes. The first dimension is clearly a negative-to-positive scale in terms of the four variables apart from *Generosity*. By contrast, *Generosity* is the main driver of the second dimension, mainly opposing *Corruption*. For example, looking at the positions of the UK, Malta, Germany and France in Fig. 1a, 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 for their overall size on this first dimension, but the composition of their ratings (their shape) is different for the four locations. The UK tends to be higher than average on *Generosity* and lower than average on *Corruption*, and 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 lower than average on *Generosity*.

Step 4: optional de-emphasizing of cases or variables in the biplot. To show all 149 country/territory names in Fig. 1a, it was necessary to distinguish between countries/territories that contributed more than average to the solution dimensions[16]. The left singular vectors corresponding to the countries/territories, without scaling, each have a sum of squares equal to 1. 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 1 divided by the number of points; 1/149 in this case. The countries/territories with contributions greater than this threshold on either of the two dimensions are the ones plotted in higher intensity in Fig. 1a. The others, which are less than this threshold on both dimensions, are plotted using lighter labels. Consequently, the high contributors are the points furthest from the origin on the respective dimensions. As an alternative, the countries/territories were represented in Fig. 1b by symbols so that their regional dispersions could be visualized without indication of specific countries/territories.

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. The directions of the five variables in the two-dimensional biplot can be equivalently obtained by regressing the variables on the two principal components. Similarly, the direction of any other variable observed in the cases can be plotted to enrich the interpretation. The difference is that a supplementary variable has not been optimized in the biplot, like the five active variables that were used to construct the solution. Two variables, the happiness score itself (*happy*), and the logarithm of gross domestic product (*logGDP*), are available for the 149 countries/territories, represented in Fig. 1b 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.



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 parts of the explained variance can be accumulated over the dimensions. As a result, the first eigenvalue maximizes the explained variance in the first dimension, the second

Fig. 1 | PCA of the indicators in the World Happiness Report. a, Plot of multivariate data for 149 countries/territories using the first two principal components (PCs) as coordinate axes. The 82 countries/territories 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, with each mean at the centre of its label. **b**, Same plot as panel **a**, but showing the countries/territories with regional symbols, with regional means indicated by labels. Variables are now shown as arrows of increasing values, with the means of all variables at the origin (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. *Social*, social support; *Life*, healthy life expectancy; *Choices*, freedom to make life choices; *Generosity*, generosity of the general population; *Corruption*, perceptions of internal and external corruption. **c**, 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. PCA, principal component analysis.

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. 3. 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. 3b. This is equivalent to finding the closest point to each multidimensional point on the plane. Figure 3c 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, the aim is to minimize the sum of squared distances $\sum_i e_i^2$, in other words, the closeness of the plane to all the points. This 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 closest to the data matrix in terms of least squared differences between the original and approximated matrices[17], often called a 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, 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, in Fig. 1a and b, the 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.

Variations of the PCA theme

There are several multivariate methods that are simple variants of PCA. One possibility is to change the way the distance function is defined, 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 approximates the Euclidean distances between the points in the full space. The Euclidean distance between points i and i' is defined as:

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

where y_{ij} and $y_{i'j}$ refer to 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

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

called the standardized Euclidean distance, where 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, which is generally applicable to two-way cross-tabulations, general frequency data or data in the form of percentages. In correspondence analysis 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) profiles by r_{ij}:

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

where c_j is the jth element of the average profile. Thus, for such relative frequency data, the mean profile element c_j substitutes the variance s_j^2 in Eq. 4, and the implied weights on the variables are the inverses $1/c_j$. In correspondence analysis, weights are also assigned to the profile points.

As an example of case weighting in PCA, suppose that there are groups of cases and that the object is to find dimensions that discriminate between the groups, to explain between-group variance rather than the total between-case variance. 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 have 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 ten regions.

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 omics research[18,19]. These data are generally compositional, since the totals of each sample are irrelevant and it is the relative values that are of interest. LRA is the PCA

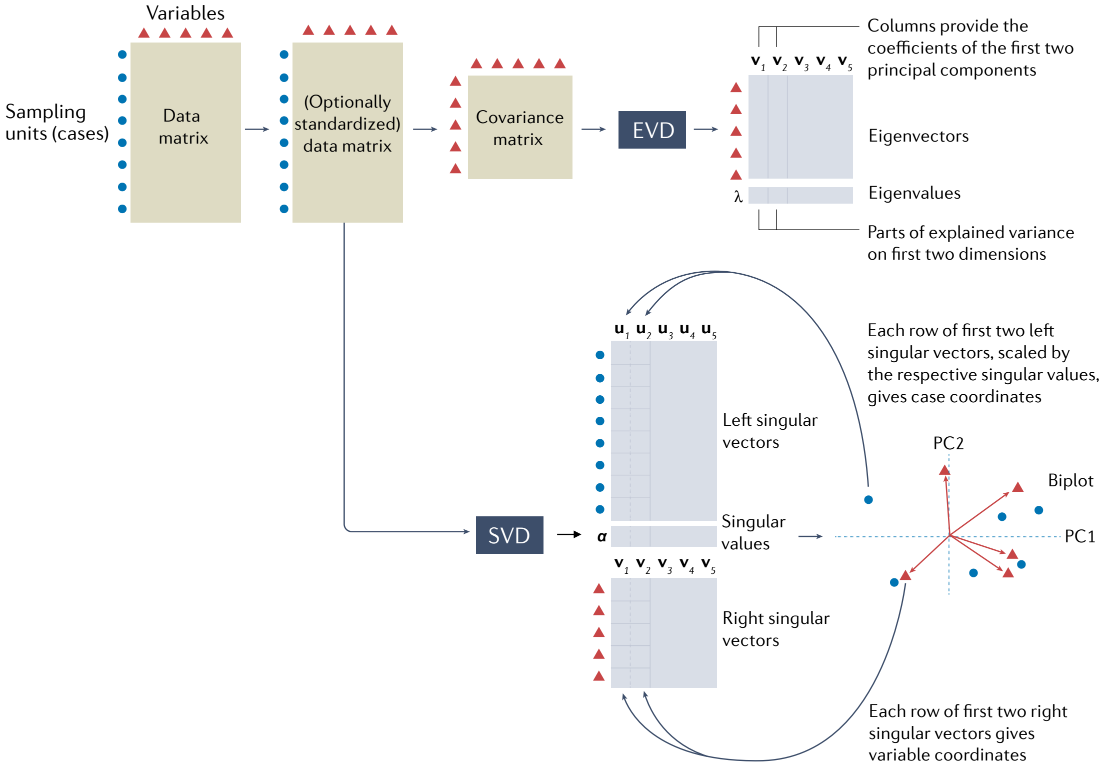

Variables
Sampling units (cases)
Data matrix
(Optionally standardized) data matrix
Covariance matrix
EVD
$\mathbf{v}_1$ $\mathbf{v}_2$ $\mathbf{v}_3$ $\mathbf{v}_4$ $\mathbf{v}_5$
Columns provide the coefficients of the first two principal components
Eigenvectors
λ
Eigenvalues
Parts of explained variance on first two dimensions
$\mathbf{u}_1$ $\mathbf{u}_2$ $\mathbf{u}_3$ $\mathbf{u}_4$ $\mathbf{u}_5$
Each row of first two left singular vectors, scaled by the respective singular values, gives case coordinates
Left singular vectors
SVD
α
Singular values
$\mathbf{v}_1$ $\mathbf{v}_2$ $\mathbf{v}_3$ $\mathbf{v}_4$ $\mathbf{v}_5$
Right singular vectors
PC2
Biplot
PC1
Each row of first two right singular vectors gives variable coordinates

Fig. 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 principal component analysis (PCA). The lower pathway is a more efficient one, using the singular value decomposition (SVD) to directly give the positions and vectors of the variables in a joint representation. 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 $\sqrt{n}$, where n is the number of cases (rows).

of log-transformed data that are initially row-centred, meaning that each row of the log-transformed data is centred by its respective row mean. Because PCA performs column-centring, LRA is 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[20,21]. 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[19].

Results

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 on the leading dimensions and what can be considered random noise on the minor dimensions. The sequence of explained variance percentages suggests how many non-random major dimensions there are. Figure 1c 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 (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[22] with the decision on the dimensionality made by looking for the elbow in the sequence of bars. There is a similar line passing through the first two bars, which changes slope abruptly compared with the line through the last three. Based on this elbow rule, the conclusion is that the data are two-dimensional. Therefore, the two-dimensional solutions presented previously are a valid representation of 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[22–29].

It is not expected that datasets always 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 (see an example in the Supplementary Information), or a selection of planar views of the points made. For example, dimensions 1 and 2 could be plotted, and then separately, dimensions 1 and 3, or for four-dimensional solutions, a plot of dimensions 1 and 2, and also a plot of dimensions 3 and 4, could be produced, an example of which is given in ref. [30].

Interpretation of a PCA biplot

The PCA biplot in Fig. 1b, explaining 71.5% of the data variance, consists of points for the cases and vectors for the variables. As shown in Fig. 3, 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 approximate 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, which takes the exact interpoint distances as input and produces low-dimensional approximations of the distances[31].

Box 1

The singular value decomposition and the PCA biplot coordinates

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

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

where

- $\mathbf{D}$ is the diagonal matrix of the (positive) singular values α_1, α_2, … in descending order.
- $\mathbf{U}$ and $\mathbf{V}$ are the matrices of left and right singular vectors (columns $\mathbf{u}_1$, $\mathbf{u}_2$, … and $\mathbf{v}_1$, $\mathbf{v}_2$, …) and are orthonormal: $\mathbf{U}^{\mathrm{T}}\mathbf{U}=\mathbf{V}^{\mathrm{T}}\mathbf{V}=\mathbf{I}$, namely all $\mathbf{u}_k^{\mathrm{T}}\mathbf{u}_\ell$ and all $\mathbf{v}_k^{\mathrm{T}}\mathbf{v}_\ell$ are equal to 0 for $k \neq \ell$ but equal to 1 for $k=\ell$.

Written as a sum of products of the individual vectors, the SVD of $\mathbf{X}$ is $\sum_{k=1}^{m} \alpha_k \mathbf{u}_k \mathbf{v}_k^{\mathrm{T}}$, where m is the rank of $\mathbf{X}$. Since the sum of squares of each rank 1 matrix $\mathbf{u}_k \mathbf{v}_k^{\mathrm{T}}$ 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 principal component analysis (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.

Notes

- 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 samples.
- To obtain complete equivalence between the two alternative workflows of the SVD and the eigenvalue decomposition (EVD), 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.

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 to understand how the cases line up approximately, but not exactly. To give a concrete example, the country/territory points can be projected perpendicularly onto a biplot axis pointing in the top right direction of Fig. 1b, corresponding to *Choices*. Countries/territories, such as those in Scandinavia – Sweden, Norway and Denmark – as well as Singapore are highest in the positive direction of the arrow, whereas Afghanistan is the lowest on the negative side, towards bottom left. As the origin of the biplot represents the means of all five variables, countries/territories projecting on the upper right of the biplot axis of *Choice*s are estimated to be above the mean, while those on the lower left are estimated to be below the mean. Taking the projected values for all countries/territories onto the diagonal sloping axis and correlating them with the original data for the variable *Choices* gives a correlation of 0.815. The square of this correlation, 0.664, is the part of the variance of the variable *Choices* that is explained by the first two principal components.

The set of countries/territories can be projected on each of the other biplot axes defined by the direction vectors. 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. The second variable, *Life*, has the highest R^2, so the way the countries/territories line up on this direction in the biplot will be the most accurate. By contrast, the projections onto the fifth variable, *Corruption*, with the lowest R^2, will give less accurate estimates. The projected positions of the countries/territories onto the five biplot axes are simply the data values estimated by the two principal components PC1 and PC2 by multiple regression, reinforcing the idea that PCA is a method of matrix approximation.

Numerical results of a PCA

The percent of variance values were plotted and interpreted in Fig. 1c. 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 the first coefficient of PC1 in Eq. 1. Since all the eigenvectors have a sum of squares equal to 1, and thus are 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, which are identical to the squares of the first row, that is, the squared singular values (eigenvalues) divided by n. The sum of squared correlations of each variable row-wise over the five dimensions

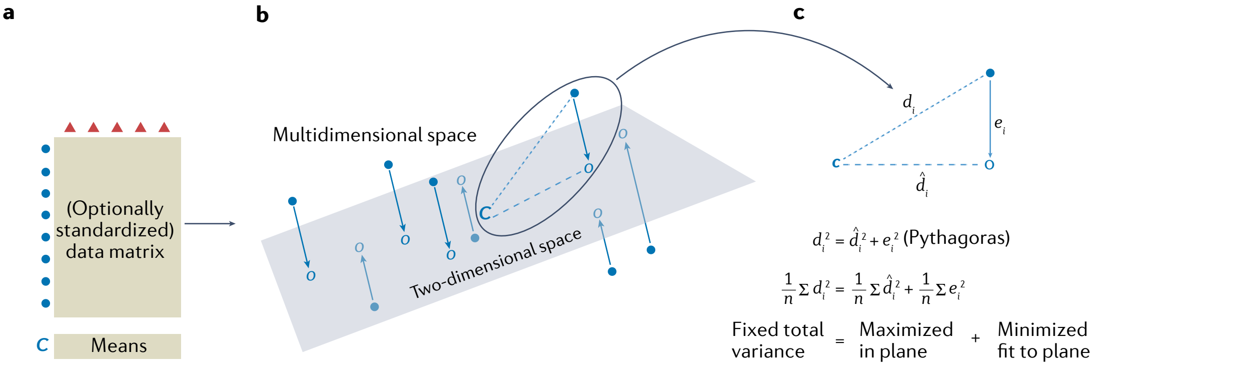

a
b
c
(Optionally standardized) data matrix
C
Means
Multidimensional space
Two-dimensional space
C
d_i
e_i
$\hat{d}_i$
$d_i^2 = \hat{d}_i^2 + e_i^2$ (Pythagoras)
$\frac{1}{n}\Sigma d_i^2 = \frac{1}{n}\Sigma \hat{d}_i^2 + \frac{1}{n}\Sigma e_i^2$
Fixed total variance
=
Maximized in plane
+
Minimized fit to plane

Fig. 3 | 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 singular value decomposition 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 (maximizing variance) is equivalent to minimizing the average squared distances from the points to the plane (minimizing fit).

in Table 1 is equal to 1. 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*, $0.764^2 + 0.285^2 = 0.664$. Again, this only holds for this particular case of standardized variables.

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. As a result, 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 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 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. 1a, where the above average country/territory contributors were shown in a more intense colour to improve the legibility of the biplot.

Applications

A high-dimensional grouped dataset

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[32–34] consists of a 63 × 2,308 matrix of gene expression data, for 63 children and 2,308 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 given as log-transformed, and no further standardization is required. The number of variables is higher than the number of cases, which is the number of tumours, and the dimensionality of the data is determined by the number of cases minus 1, which is $63 - 1 = 62$ here. To understand this, and given that the data are column-centred, two cases in a high-dimensional space lie exactly on a line (1-dimensional), three cases lie in a plane (two-dimensional), four cases lie in a three-dimensional space, and so on.

Figure 4a shows the PCA of the data, where the four tumour types 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. Similarly to the countries/territories in Fig. 1a, these high-contributing genes are the most outlying in the biplot, and likewise aim to explain the variance in the individual cases, rather than the variance between the cancer groups. The individual tumours in the different groups can be seen to overlap substantially, especially the groups EW and RM. Also shown in Fig. 4a 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, 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, their 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. 4a can be qualified as an unsupervised PCA, the PCA in Fig. 4b is now supervised to explain group differences. This PCA of the four group means has only three dimensions. The percentages on the dimensions are much higher, because they are expressed relative to the between-group variance. The group means are now highly separated, the convex hulls do not overlap and the confidence ellipses are 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 does take these covariances into account. Supplementary Video 1 shows the exact three-dimensional solution of the group centroids. Supplementary Video 2 shows an animation of the cases in Fig. 4a transitioning to the group separation in Fig. 4b 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 ref. [16].

Sparsity constraints for wide data

The coefficients that define the principal components are generally all non-zero, or dense. For wide data, when the number of features is very high, in the hundreds or thousands, this presents a problem for interpreting so many coefficients. This is the case with the present cancer dataset as well as microbiome and omics data in general, where there can be thousands of variables compared with 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 to partially solve this problem rotated the PCA solution so that variables aligned themselves closer to the dimensions[38,39].

More recently, sparse PCA implementations[40–46] handle this problem by introducing penalties on the coefficient sizes that force some coefficients down to zero, 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 to shrink the absolute values of some coefficients to zero. An improvement that achieves coefficient sparsity can also be 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 ref. [49]. Sparse PCA is a fairly recent innovation, and is still actively debated in the literature[50,51].

Figure 4c shows the effect of sparse PCA on the results of the Khan gene data shown in Fig. 4a. Most of the 2,308 genes have been eliminated, 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. 4c is very similar to that in Fig. 4a. Within each cancer group there is a vertical separation of samples with positive PC2 and those with negative PC2, which is now accentuated. 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 with Fig. 4a. In the sparse centroid PCA of Fig. 4d, the cancer groups are separated and the few genes with nonzero values (72 for PC1 and 79 for PC2) will be indicators of this separation. Notice that there is now a clear distinction between groups RM and EW, with 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 with the regular PCA of the centroids in Fig. 4b.

Supplementary Video 3 shows an animation of the tumour samples in Fig. 4b transitioning to the sparse solution in Fig. 4d. The outlying genes in Fig. 4b, which contributed the most to the regular PCA solution, can be seen to be the ones that are not eliminated by shrinking to zero in the sparse solution.

Table 1 | Correlations of the five PCs with the five variables

	PC1	PC2	PC3	PC4	PC5	Row sum of squares
Singular values						
Singular values /√n	1.532	1.107	0.838	0.692	0.495	5
Correlations with variables						
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
Summary values						
Column sum of squares	2.348	1.226	0.703	0.478	0.245	Row sum 5

Social, social support; *Life*, healthy life expectancy; *Choices*, freedom to make life choices; *Generosity*, generosity of the general population; *Corruption*, perceptions of internal and external corruption. The sum of squared correlations for each variable is 1. The sum of squared correlations for each principal component (PC) is the square of the first row (squared singular value divided by n, the number of cases) and is equal to the part of the variance explained by that PC, out of a total variance of 5. Expressed as percentages, these are the percentages on the PC dimensions. For example, on the first dimension, 100×2.348÷5=47.0%.

Correspondence analysis

Correspondence analysis[52,53] and its constrained version, canonical correspondence analysis[54], are among the most popular techniques for visualizing abundance or presence/absence data in ecology, but are also extensively used in archaeology, linguistics and sociology. By ‘constrained’, we mean 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. 4b is a constrained PCA, in which 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 from ref. [55]; 600 samples were obtained over a period of 6 years, 1999–2004, each obtained by 15 minutes of trawling in the Barents Sea north of Norway; the numbers of up to 66 different fish species are counted in each sample. The sampling was performed at a similar time of 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 correspondence analysis 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 Eq. 4), using a weighting of the squared differences between profile elements by the inverse of the average profile with elements $c_j = n_{+j}/n$, that is, the column sums n_{+j} divided by the total n. The chi-square distance between two rows uses a standardization of the profile data, $(n_{ij}/n_{i+})/\sqrt{c_j}$, followed by the usual Euclidean distance applied to these transformed values.

The final property that distinguishes correspondence analysis from PCA is that the points have weights proportional to their marginal frequencies; the row weights are n_{i+}/n. Correspondence analysis also has the special property that it treats rows and columns symmetrically – it is equivalent to think of the relative frequencies columnwise. The column profiles, which are the points to be approximated in multidimensional space, can be considered 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.

Similar to the genetic study of the child cancers, there is a specific objective in analysing the Barents Sea fish data: to see whether there is a temporal evolution of the relative fish abundances across the 6 years. This is achieved analytically by aggregating the fish abundances into a 6 × 66 matrix, where the rows are the 6 years and the counts are now

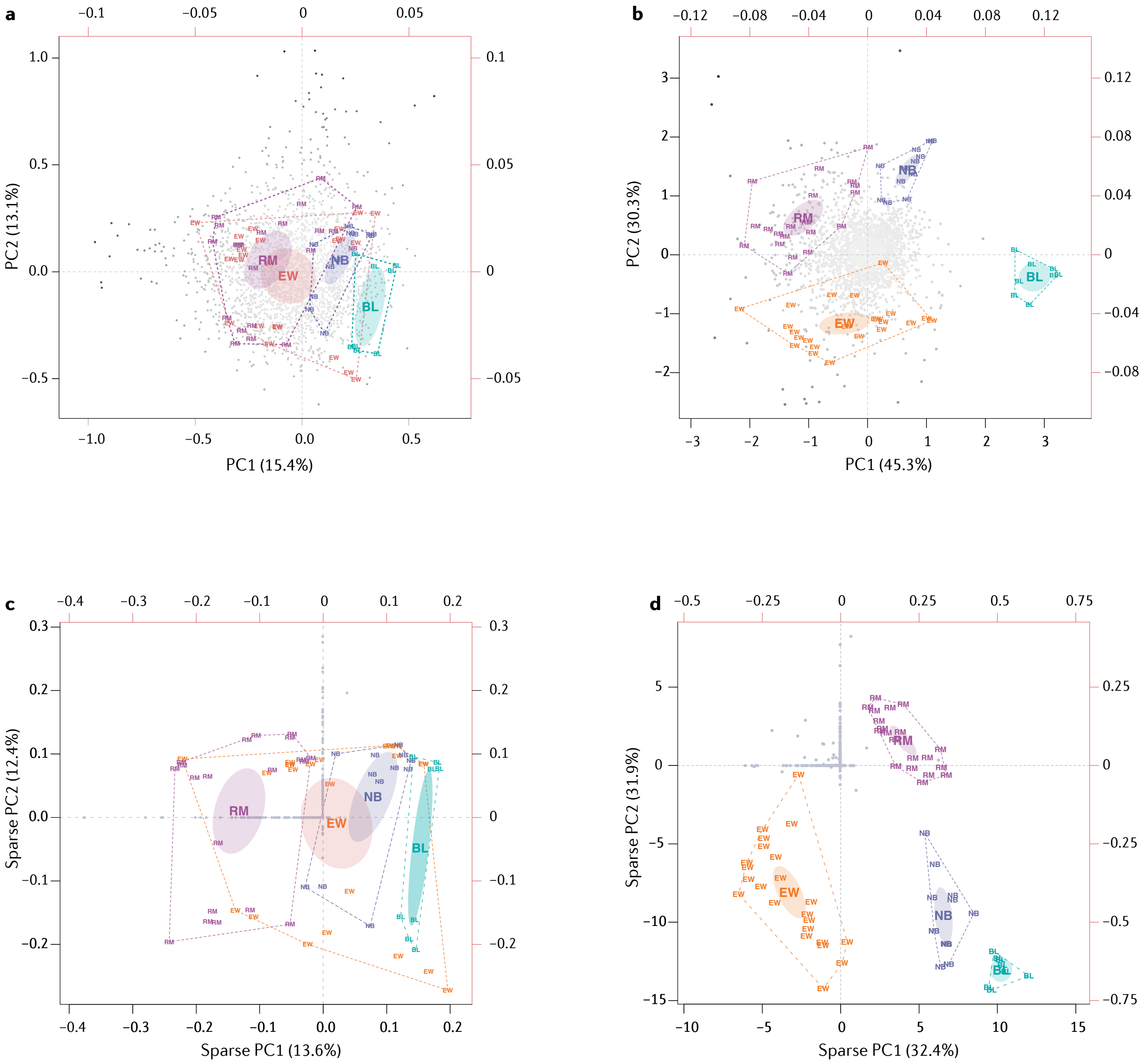


Fig. 4 | PCA of the child cancer data. a, Unsupervised principal component analysis (PCA) of the individual-level data. The four tumour groups — Burkitt lymphoma (BL), Ewing's sarcoma (EW), neuroblastoma (NB) and rhabdomyosarcoma (RM) — are enclosed by convex hulls. 95% confidence ellipses are shown for the group means, which are located at the group label in larger font. The 2,308 genes are displayed as dots, where darker dots indicate higher contributions to the separation of individual tumours. **b**, Supervised PCA of the tumour data, explaining the between-group variance. The four tumour groups are again enclosed by convex hulls, with confidence ellipses for the group means that are now all separated. The darker dots now correspond to genes making higher contributions to the group separation. **c**, Sparse PCA of the tumour data, comparable to the regular PCA in panel **a**. Most of the 2,308 genes are eliminated and the remaining genes are now identified with either the first or second PC, and in a few cases with both PCs. The percentage of explained variance has dropped from 28.5% in the panel **a** solution 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 panel **b** to 71.6%.

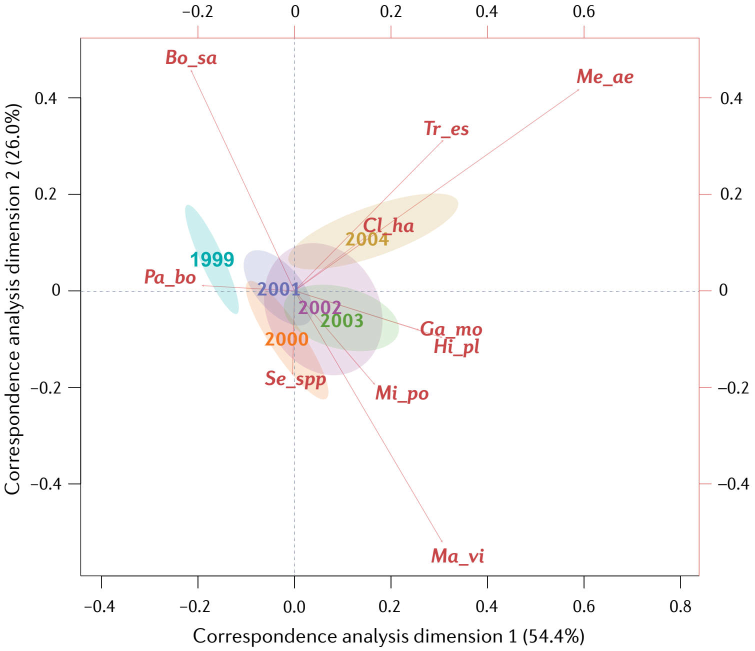


Fig. 5 | Correspondence analysis of the Barents Sea fish data, 1999–2004, explaining the between-year variance. The year means are shown, as well as their 95% confidence ellipses. The 10 species (out of 66) that contribute more than average to this two-dimensional solution are shown. Only species abbreviations 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) and *Se_spp* (redfish). The 600 individual sample points, which show great variation owing to the sparsity of the data, are not shown.

summed for each year. The constraint is achieved by the discrete variable year, with six categories. So the correspondence analysis applied to this aggregated matrix is effectively a canonical correspondence analysis, shown in Fig. 5. 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. These are based on 1,000 bootstrap resamplings of the coordinates of the 600 samples, followed by recomputing the year aggregations for each bootstrap sample and computing the ellipse for each year's set of 1,000 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. *Pa_bo* (*Pandalus borealis*, shrimp) is highest in 1999, while *Me_ae* (*Melanogrammus aeglefinus*, haddock) and *Tr_es* (*Trisopterus esmarkii*, Norway pout) are highest in 2004. These conclusions can be verified in the table of relative abundances. For example, the last two species, *Me_ae* and *Tr_es*, have percentages in 2004 of 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), which has percentages highest (1.2%) in 1999 and 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] that 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 five-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. This random permutation of year labels is 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, gives 3 out of 1,000 in the tail of the permutation distribution, hence the *P* value.

Imposing external constraints

Figures 4b and 5 are examples of PCA and correspondence analysis constrained to the variance between groups of cases, which are cancer types in Fig. 4b and years in Fig. 5. Constraints can be made with respect to categorical variables as well as continuous variables, a strategy that is very common in ecological applications. The data matrix for the PCA (or correspondence analysis) 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, such as sampling year, are coded as dummy variables that also act as constraining variables. Other examples are morphometric measurements on different fish, or microbial compositions, as the multivariate responses and constraining variables could be fish diet[55].

Rather than explain the total variance of the response dataset, the objective is to focus on the part of variance that is directly related to the explanatory variables. This is achieved by projecting the response dataset onto the space defined by the explanatory variables (called the constrained or restricted space), thereby eliminating the biological variance unrelated to the environmental variables. The search for principal components is then performed in the constrained space, called 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 the directions of continuous constraining explanatory variables or points showing the positions of categories of constraining categorical variables, as in Figs. 4b and 5. Canonical correspondence analysis is an analogous constrained method for response data such as frequency counts or presence–absence data and is one of the most widely used methods in quantitative ecology[54,60,61].

Multiple correspondence analysis

A popular variant of correspondence analysis is multiple correspondence analysis[53,62], used for multivariate categorical data, often found in social surveys where respondents choose response categories in a series of questions[63–65]. 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. Multiple correspondence analysis is the application of correspondence analysis to the indicator matrix, generating biplots of the respondents and categories. One advantage of this approach is that association patterns of single categories, such as 'missing value' or 'no opinion' categories, can be investigated[66,67]. In sociological applications it is generally the averages and dispersions of respondents for different demographic categories that are of interest in the multiple correspondence analysis results.

Mixed 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[68–71]. 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[72–74].

Another context related to fuzzy category coding occurs when 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[75–82].

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[83–86]. A composite indicator is formed when individual indicators are compiled into a single index. For example, to investigate public opinion on government measures aimed at reducing carbon dioxide, a survey could be executed that asks 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, giving an approximation of participants' opinion on the government measures. However, this raises questions about whether taking the direct sum is appropriate, and if all questions measure the same concept. PCA can be used as a first exploration. A single large eigenvalue is a strong indication that there is a single dominant scale. Two or more large eigenvalues are indications of the presence of multiple concepts, meaning 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[87]. Multiple correspondence analysis has also been used to construct indices based on categorical data[88,89], since the method assigns quantitative values to categories to maximize explained variance, and these summed quantifications then constitute new scales[90].

Reproducibility and data deposition

Minimal reporting

The results of a PCA are generally reported in the form of a two-dimensional biplot, where the unspoken assumption is that this is an adequate explanation of the dataset. Percentages of variance should be reported for each dimension. Adequate does not necessarily mean 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. 5. For datasets with many variables, less attention needs to be paid to variables that make low contributions to the solution, as in Fig. 4a and b, or in Fig. 5, where the low contributors are de-emphasized or omitted. To avoid distortion, there should be an aspect ratio of 1 in such a plot, since its interpretation is in terms of distances and perpendicular projections.

R and Python implementations

PCA is widely implemented in commercial and open-source statistical packages. In the R language, there are 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 this Primer. Table 2 shows the packages and functions that can be used to implement the methods described in this Primer. Base R functions, sometimes requiring more code, give all the flexibility needed for producing publication quality results.

Limitations and optimizations

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[91], image compression[92], face recognition[93,94], industrial process modelling[95], quantitative finance[96], neuroscience[97], genetics and genomics[98–101], to name a few, the size and the dimensionality of the datasets can be very large, leading to computational issues. At the core of PCA 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 into 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 interest usually being focused on the first few eigenvalues or singular values, and the corresponding eigenvectors or singular vectors; that is, a truncated EVD or SVD. The goal of the power method is to find the largest eigenvalue and associated eigenvector[102], and the Lanczos algorithm is an adaptation to find the leading eigenvalues and vectors[103]. Some of the most enhanced batch EVD methods are variations of the Lanczos algorithm[104,105].

Table 2 | Packages and functions implementing PCA and its variants

Package	Function	Description
R packages		
`stats`	`prcomp, princomp`	These base R functions have minimal output and sometimes conflicting terminology.
`base`	`svd`	Singular value decomposition of a matrix.
`FactoMineR`[140]	`PCA`	The `FactoMineR`, `ade4`, `amap`, `easyCODA` and `PCAtools` packages all have options for weighting rows and columns of the data matrix. Options for supplementary rows and supplementary columns are provided in `PCA` (`FactoMinerR`) and `dudi.pca` (`ade4`). Note that `PCA` (`easyCODA`) has supplementary rows only. The `FactoMineR`, `ade4` and `easyCODA` packages have extensive results in the created objects. The `easyCODA` package is aimed at compositional data analysis but has functions for PCA, correspondence analysis, LRA and RDA. Most of these packages have dedicated plotting functions (in the case of the `ade4` package there is a separate package `adegraphics`[141]).
`ade4`[142]	`dudi.pca`	
`amap`	`acp`	
`easyCODA`[21]	`PCA`	
`PCAtools`	`pca`	
`pca3d`	`pca3d`	Three-dimensional PCA graphics.
`vegan`	`rda`	This function computes RDA, that is, PCA with constraints, but can also perform PCA with no constraints. The same package has function `cca` for correspondence analysis with or without constraints.
`elasticnet`	`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`	`prcomp_irlba`	These fast and memory-efficient functions (prcomp_irlba, svds, rpca, batchpcs and i_pca) are used when the data are too large to fit into memory, or are arriving in streams.
`RSpectra`	`svds`	
`rsvd`[143]	`rpca`	
`onlinePCA`	`batchpca`	
`idm`[144]	`i_pca`	
`symbolicDA`	`PCA.centers.SDA`	PCA for interval-valued data.
`RSDA`	`sym.pca`	
`fdapace`	`FDA`	PCA of functional data, where data are sparse and longitudinal.
`softImpute`	`softImpute`	Imputation of missing values for PCA or matrix completion; can handle very large and sparse matrices.
`missMDA`[145]	`imputePCA`	Imputation of missing values for PCA.
Python libraries		
`scikit-learn`[146]	`sklearn. decomposition. PCA`	PCA, also with truncated SVD for large datasets.
	`sklearn. decomposition. SparsePCA`	Sparse PCA using the lasso penalty.
	`sklearn. decomposition. IncrementalPCA`	Computes solution by processing data in chunks, when dataset is too large to fit into memory.
`NumPy`[147]	`linalg.svd`	SVD of a matrix.

LRA, logratio analysis; PCA, principal component analysis; RDA, redundancy analysis; SVD, singular value decomposition.

An alternative probabilistic approach leads to approximate yet accurate matrix decompositions[106].

Batch methods lead to a substantial reduction of the computational cost, but do not solve cases where the matrix cannot be stored in memory, or when new data are constantly produced as data flows. The general aim of online matrix decomposition methods is to incrementally update an existing EVD or SVD as more data arrive. Several approaches have been proposed in the literature[107–111]. 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[112] and data from recommender systems on e-commerce websites[113]. An example of the continuous arrival of image data is surveillance cameras[114–117], where each image is coded as a single vector, with p given by the number of pixels of that image. If nothing happens, the background corresponds to low-variance singular vectors, whereas any disturbance or intruder, however small, creates a big change.

Missing values using SVD

PCA can be extended to the case when data are partially observed. For example, suppose that 10% of the 149 × 5 = 745 entries in the World Happiness Report dataset were corrupted and, as a result, are indicated as missing. An intuitive way 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, missing values could be replaced by the mean of the corresponding column. For example, missing values for the variable *Life* would be replaced by the average value for all the countries/territories with observed values. Although widely applied in practice, this approach ignores correlation between the variables.

To explain the goal of PCA with missing values, the standard PCA is linked to the low-rank matrix approximation problem. In what follows, it is assumed that $\mathbf{X}$ is a matrix with missing values, which has been pre-centred and pre-scaled using the observed values. Finding the first r

principal components is equivalent to searching for the matrix $\mathbf{X}$ of rank r, denoted by $\mathbf{X}_r$, that minimizes the residual sum-of-squares (RSS) to the original data matrix. For fully observed data, RSS is measured for all matrix elements, but when some data values are missing, RSS is measured 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 2. An example is given in the online R script of simulating 10% missing data, and the results are quite consistent with those using the complete dataset.

Matrix completion

The previous section describes an algorithm to make PCA work on a data matrix with missing data. Attention was not focused on the values used to replace the missing ones. However, in other contexts, the replaced or imputed values are of principal interest. A well known recent example is the Netflix competition[113], in which a huge dataset of 480,189 customers and 17,770 movie titles was supplied to contestants: see a small part of the dataset in Fig. 6a. On average, each customer rated about 200 movies, meaning 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 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.

Box 2

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 residual sum-of-squares (RSS) between completed $\mathbf{X}$
    and $\mathbf{X}_0$ and denote it by RSS$_0$.

Step 2: find solutions for ranks $r$=1, 2, ..., $p$ 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 singular value
        decomposition as follows:
            $$\mathbf{X}_r = \sum_{k=1}^{r} \alpha_k \mathbf{u}_k \mathbf{v}_k^{\mathsf{T}}$$
    (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 (RSS$_{r-1}$–RSS$_r$)/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 $\mathbf{v}_k$
for $k$=1, 2, ..., $r^*$ that form the decomposition of $\mathbf{X}_{r^*}$.
```

Notes

- Because of the pre-centring, 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.

PCA, principal component analysis.

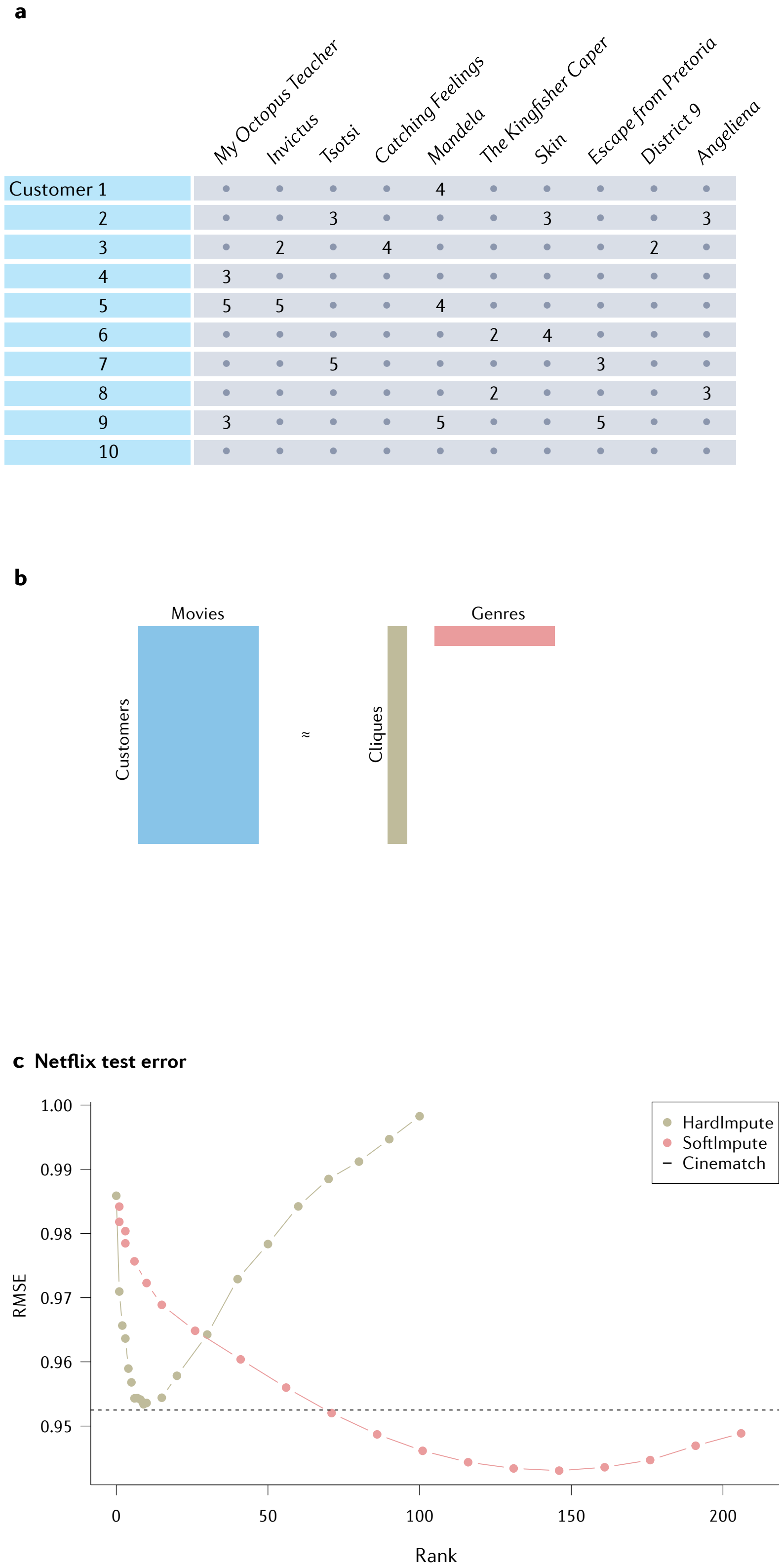
A low-rank matrix approximation of the PCA type is a natural solution to such a problem (Fig. 6b): $\mathbf{M} \approx \mathbf{CG}^{\mathsf{T}}$. Movies can be considered as belonging to r genres – for example, thrillers, romance and more – represented as the rows of the red matrix, while users belong to r cliques – such as those who like thrillers, those who like romance and so on – which are the columns of the green matrix. This translates into a matrix approximation $\widehat{\mathbf{M}}$ of rank r, and the general element of the low-rank approximation $\widehat{\mathbf{M}} = \mathbf{CG}^{\mathsf{T}}$ is $\widehat{m}_{ij} = \sum_{k=1}^{r} c_{ik} g_{jk}$, where the cliques and the genres are combined. As a result, 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 RSS, $\sum (m_{ij} - \widehat{m}_{ij})^2$, which optimizes the fit of the $\widehat{m}_{ij}$ to the m_{ij} by least squares, over the observed values in $\mathbf{M}$ only. The form of the matrix product $\mathbf{CG}^{\mathsf{T}}$ 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 2 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[118]. One trick involves solving the SVD problem, with filled-in values in $\mathbf{M}$, in alternating stages by fixing the genre matrix $\mathbf{G}$, optimizing the fit with respect to $\mathbf{C}$, then fixing the clique matrix $\mathbf{C}$ and optimizing with respect to $\mathbf{G}$. Another trick is to store only the observed elements of the matrix $\mathbf{M}$, which are very few in the Netflix example compared with the elements of the whole matrix, in sparse format and adapting the computations to this format. The basic algorithm has a further adaptation, called the SoftImpute algorithm[118], which applies a shrinkage penalty to the singular values; some become zero in the process and this determines the rank of the solution.

The SoftImpute algorithm is described more fully in refs. [118–121] and has been demonstrated to give improved performance over HardImpute in many applications, see refs. [122,123]. For the massive Netflix example, Fig. 6c 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.

Outlook

PCA has been, and will remain, the workhorse of exploratory data analysis and unsupervised machine learning, while also 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 unexpected areas of research. This section mentions some recent innovations in which PCA and its core algorithm, the SVD, play an important part, especially in the analysis of very large challenging datasets from genetics, ecology, linguistics, business, finance and signal processing. Some of these have already been described, such as sparse PCA and matrix completion. Images, physical objects, and functions are non-standard data objects, to which PCA can be



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

Fig. 6 | Movie recommender system via matrix completion. **a**, A small portion of the large data matrix ***M*** of movie ratings. **b**, The matrix factorization of the data matrix ***M*** approximately into a product of low-rank matrices, ***C*** (cliques) and the transpose of ***G*** (genres). **c**, 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.

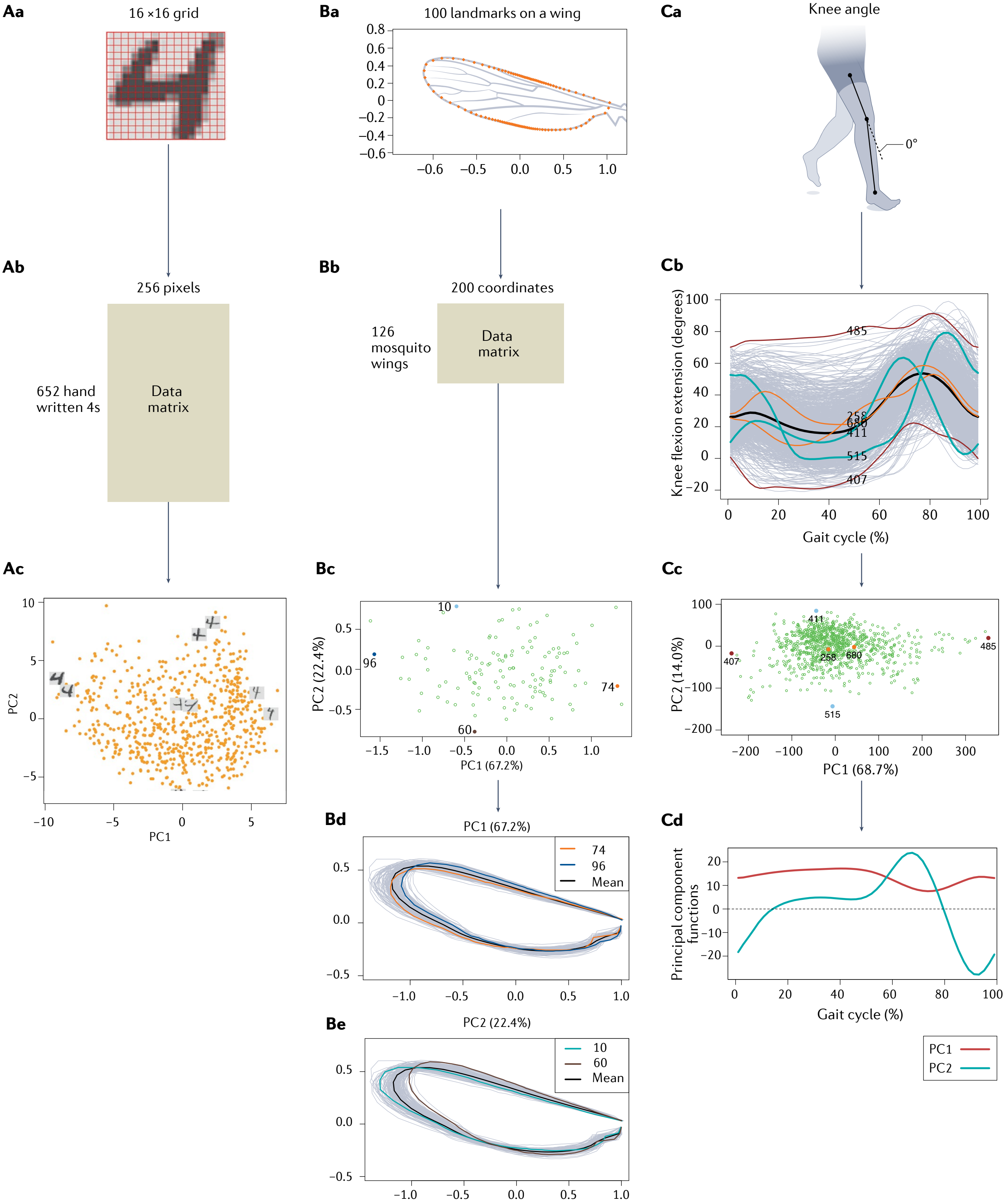

Aa
16 ×16 grid
Ab
256 pixels
652 hand written 4s
Data matrix
Ac
PC2
PC1
10
5
0
−5
−10
−5
0
5
Ba
100 landmarks on a wing
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.6
−0.5
0.0
0.5
1.0
Bb
200 coordinates
126 mosquito wings
Data matrix
Bc
PC2 (22.4%)
PC1 (67.2%)
10
96
74
60
0.5
0.0
−0.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
Bd
PC1 (67.2%)
74
96
Mean
0.5
0.0
−0.5
−1.0
−0.5
0.0
0.5
1.0
Be
PC2 (22.4%)
10
60
Mean
0.5
0.0
−0.5
−1.0
−0.5
0.0
0.5
1.0
Ca
Knee angle
0°
Cb
Knee flexion extension (degrees)
Gait cycle (%)
485
258
680
411
515
407
100
80
60
40
20
0
−20
0
20
40
60
80
100
Cc
PC2 (14.0%)
PC1 (68.7%)
411
407
258
680
485
515
100
0
−100
−200
−200
−100
0
100
200
300
Cd
Principal component functions
Gait cycle (%)
20
10
0
−10
−20
0
20
40
60
80
100
PC1
PC2

Fig. 7 | PCA of visualizable objects: images, shapes and functions. A, Principal component analysis (PCA) of images. **Aa**, 16 × 6 grid on one of 662 handwritten fours, for coding the greyscale image for the 256 cells. **Ab**, Resultant 652 × 256 data matrix for all the fours. **Ac**, PCA plot of the 662 samples, showing some selected images. **B**, PCA of shapes. **Ba**, 100 landmarks, each with an *x*- and *y*-coordinate, shown on 1 of 126 mosquito wings, giving 200 values for each wing. **Bb**, The resultant 126 × 200 data matrix of coordinates. **Bc**, The PCA of the 126 wings, with some emblematic ones labelled on the extremes of the two dimensions, PC1 and PC2. **Bd**, All the wing shapes, showing the mean wing shape and the shapes of the two wings on the extremes of PC1. **Be**, As for PC1 but now showing the mean wing shape and the shapes of the two wings on the extremes of PC2. **C**, PCA of functions. **Ca**, Angle measurement of the knee. **Cb**, 1,000 knee flexion curves, showing the mean curve as the thicker black curve in the middle and some other selected ones, which are both similar and different from the mean. **Cc**, The PCA plot of the 1,000 curves, showing the positions of the ones highlighted in panel **Cb**. **Cd**, The first (red) and second (blue) principal component functions. Data in panel **C** are taken from ref.[148].

applied after using clever ways of coding the data in the form of a data matrix.

PCA of images

Often the observations represented by PCA can be rendered in a recognizable form, such as images. Such images could be birds from closely related species, human faces or retinal scans taken during routine eye exams. An example dataset contains 652 handwritten versions of the number four, scanned from the zip codes on letters posted in New York. Each is represented by a 16 × 16 greyscale image (see the grid in Fig. 7Aa, with pixel values ranging from −1 to +1). Each image of a four can then be coded as a single vector of length 256, defining a point in a 256-dimensional space. Consequently, the data matrix in Fig. 7Ab is 652 × 256. Figure 7Ac shows a plot of the first two principal component scores for these data, where some emblematic examples of the points to interpret the configuration have been added. These examples include two images each that project to the extremes of PC1 and PC2, and two that project near the middle. Their images are included in the plot, to illustrate what components of variation the axes explain. The PC1 axis seems to differentiate fours with stubby tails (negative side) versus long tails (positive side). The PC2 axis (positive side) has fours with stubby upturns in the left part of their horizontal arms, and long right arms, contrasted with the opposite pattern on the negative side.

PCA of shapes

A special case of images is that of shapes. Here, an example is presented from morphometrics, the study of shape, looking at 126 mosquito wings. The plot in Fig. 7Ba shows one mosquito wing, with 100 landmarks indicated along the edge of a wing[124]. Each wing is represented by 100 pairs of (x,y) coordinates, with 200 numbers in total. The data matrix for the 126 mosquito species studied is a 126 × 200 matrix of coordinates (Fig. 7Bb), in which 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, another multivariate method that relies on the singular value decomposition[125,126]. PCA is used to understand the shape variation of the wings. Fig. 7Bc 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. 7Bd shows all the wings in grey, the mean wing shape in black and then the two extreme wings on PC1 coloured the same as the dots in Fig. 7Bc. Fig. 7Be 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.

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[127,128], there have been several developments[129–134]. Suppose that each data feature corresponds to a value of some function evaluated at different points of continuous time. The context presented here is the measurement of the angles of knee flexion, shown in Fig. 7Ca, for a set of 1,000 patients during a gait cycle, 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 1,000 × 100 matrix, and all the functions are represented as a set of curves in Fig. 7Cb, 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. 7Cc.

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, for instance by the mean vector, the 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. Now, each observed curve is approximated by the mean curve plus linear combinations of the two principal component curves. Fig. 7Cc shows the PCA plot of the 1,000 curves. 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. 7Cd with the same horizontal scale as Fig. 7Cb, 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. 7Cc have curves similar to the mean curve in Fig. 7Cb. It can be deduced that PC1 is mostly a size component, in the form of an almost constant vertical knee shift, and PC2 is a shape component in the form of a differential phase shift. Looking back at the emblematic samples in Fig. 7Cb and Cc confirms this interpretation.

The two principal component curves shown in Fig. 7Cd 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 smooth principal component curves are preferred for the solution. In this case, it could be insisted that any solution curve be a linear combination of a small set of smooth functions in the columns of a matrix $\mathbf{S}$. 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 $\mathbf{S}$ 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 $\mathbf{XS}$, where, in this application, $\mathbf{X}$ is the original 1,000 × 100 data matrix[128,135].

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 only a few examples. The art is in coding the appropriate variables, or features, prior to the application of PCA.

Glossary

Active variables
Variables used to construct the principal component analysis solution.

Biplot
Joint representation in principal component analysis of the sampling units (usually the rows of the data matrix) represented as points in a scatterplot, often using the principal components as coordinates and variables (the columns) obtained from the right singular vectors shown as arrows.

Biplot axis
Axis in the direction of the variable arrow in a biplot.

Bootstrap
Process aimed at assessing the statistical variability of a solution by repeatedly creating a bootstrap dataset derived from the original dataset through sampling the cases with replacement and computing the solution each time.

Covariance matrix
Matrix containing the covariances between all pairs of variables.

Dense
In the context of a data matrix, the presence of very few or no zeros; in the context of principal component analysis, the presence of no zeros in the principal component coefficients.

Eigenvalue
In principal component analysis, a value indicating the accounted variance by a principal component.

Eigenvalue decomposition
Reconstruction of any square and symmetric matrix through a sum of rank-one matrices of the outer product of an eigenvector with itself ($\mathbf{vv}^{\mathsf{T}}$) times the corresponding eigenvalue.

Eigenvector
In principal component analysis, this provides the linear combination for a principal component.

Euclidean distance
The measure of distance between two points defined as the length, in the physical sense, of the shortest straight line connecting these points.

Least-squares matrix approximation
Approximation of a data matrix such that the sum over all squared differences is minimized, between values in the data matrix and the corresponding approximated values.

Linear combination
For a set of variables, a sum of scalar coefficients times the variables.

Low-rank matrix approximation
Approximation of a matrix by one of lower rank.

Nonlinear multivariate analysis
General strategy that optimally assigns numerical values to the categories of a categorical variable and, in the context of principal component analysis, this strategy helps to increase the variance accounted for by the principal components.

Passive variables
Variables that are not used to determine the principal component analysis solution and are fitted into the solution afterwards, also called supplementary variables.

Permutation test
General computational method that compares a statistic of observed data with the distribution of the statistic simulated many times using data with the values randomly permuted under a certain null hypothesis.

Principal axis
The same as a dimension in principal component analysis and equivalent to the direction corresponding to maximal variance projections of the sampling units and uncorrelated to other principal axes.

Principal coordinates
The coordinates of the sampling units or variables on a dimension that have average sum of squares equal to the variance accounted for by that dimension.

Regressed
In the context of principal component analysis, using multiple regression to predict a variable from the principal components.

Scree plot
Plot of eigenvalue by dimension often used for selecting the number of principal component analysis dimensions by those above the straight line (scree) that goes approximately through the higher dimensions.

Shrinkage penalty
The addition to the objective function of an additional objective to reduce the absolute value of certain quantities being estimated; for example, the singular values in matrix completion, or the principal component coefficients in sparse principal component analysis.

Singular value
In principal component analysis, the square root of the variance accounted for by a principal component.

Singular value decomposition
Reconstruction of any matrix by the weighted sum of rank-one matrices consisting of the outer product of the left and right singular vectors ($\mathbf{uv}^{\mathsf{T}}$) multiplied by their corresponding positive singular value.

Singular vectors
In principal component analysis (PCA), the vectors of the singular value decomposition that lead to the row and column coordinates in a PCA biplot.

Sparsity
In the context of a data matrix, the presence of many zeros; in the context of principal component analysis, the presence of many zeros in the principal component coefficients.

Standard coordinates
Coordinates in a principal component analysis that are standardized to have the average sum of squares equal to 1.

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[136]. PCA is applied to classify the songs and see their similarities in terms of times of 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[137].

To understand the movement patterns of mice[138], continuous three-dimensional imaging data were subjected to wavelet decomposition and then analysed by PCA. This transformed the data into continuous trajectories through PC space. The first ten PCs, explaining 88% of the variance, were used to build models of the three-dimensional behaviour of mice.

The article in ref. [139] treats the problem of reconstructing images of three-dimensional molecules, using single-particle imaging by X-ray free electron lasers. This paper deals with several methodological aspects of PCA discussed and used in this Primer: alternative ways of standardization for balancing out the contributions of the image features, using the error standard deviation rather than the usual overall standard deviation; the weighting of features; and using shrinkage to determine the number of PCA dimensions.

Concluding remarks

PCA was one of the first multivariate analysis techniques to be proposed in the literature, and has since become an important and universally used tool for understanding and exploring data. This Primer has presented several applications in diverse disciplines, showing how this simple and versatile method can extract 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 arising from this quintessential statistical method are likely, meaning that PCA, along with its many variants and extensions, will remain one of the cornerstones of data science.

Code availability

Several datasets and the R scripts that produce certain results in this Primer can be found on GitHub at: https://github.com/michaelgreenacre/PCA.

Published online: 22 December 2022

References

1. Pearson, K. On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dubl. Phil. Mag. J. Sci.* **2**, 559–572 (2010).
2. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24**, 417–441 (1933).
3. Wold, S., Esbensen, K. & Geladi, P. Principal component analysis. *Chemometr. Intell. Lab. Syst.* **2**, 37–52 (1987).
4. Jackson, J. E. *A User's Guide To Principal Components* (Wiley, 1991).
5. Jolliffe, I. T. *Principal Component Analysis* 2nd edn (Springer, 2002).
Covering all major aspects of theory of PCA and with a wide range of real applications.
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. Meth.* **6**, 2812–2831 (2014).
A tutorial on how to understand, use, and interpret PCA in typical chemometric areas, with a general treatment that is applicable to other fields.
9. Jolliffe, I. T. & Cadima, J. Principal component analysis: a review and recent developments. *Phil. 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. In *World Happiness Report* Ch. 2, 13–56 (2021).
11. Cantril, H. *Pattern Of Human Concerns* (Rutgers Univ. 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, 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).
Comprehensive treatment of biplots, including principal component and correspondence analysis biplots, explained in a pedagogical way and aimed at practitioners.
16. Greenacre, M. Contribution biplots. *J. Comput. Graph. Stat.* **22**, 107–122 (2013).
17. Eckart, C. & Young, G. The approximation of one matrix by another of lower rank. *Psychometrika* **1**, 211–218 (1936).
18. 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).
19. Greenacre, M. Compositional data analysis. *Annu. Rev. Stat. Appl.* **8**, 271–299 (2021).
20. Aitchison, J. & Greenacre, M. Biplots of compositional data. *J. R. Stat. Soc. Ser. C* **51**, 375–392 (2002).
21. Greenacre, M. *Compositional Data Analysis In Practice* (Chapman & Hall/CRC Press, 2018).
22. Cattell, R. B. The scree test for the number of factors. *Multivar. Behav. Res.* **1**, 245–276 (1966).
23. Jackson, D. A. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology* **74**, 2204–2214 (1993).
24. 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).
25. Auer, P. & Gervini, D. Choosing principal components: a new graphical method based on Bayesian model selection. *Commun. Stat. Simul. Comput.* **37**, 962–977 (2008).
26. Cangelosi, R. & Goriely, A. Component retention in principal component analysis with application to cDNA microarray data. *Biol. Direct.* **2**, 2 (2007).
27. 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).
28. Choi, Y., Taylor, J. & Tibshirani, R. Selecting the number of principal components: estimation of the true rank of a noisy matrix. *Ann. Stat.* **45**, 2590–2617 (2017).
29. Wang, M., Kornblau, S. M. & Coombes, K. R. Decomposing the apoptosis pathway into biologically interpretable principal components. *Cancer Inf.* **17**, 1176935118771082 (2018).
30. Greenacre, M. & Degos, L. Correspondence analysis of HLA gene frequency data from 124 population samples. *Am. J. Hum. Genet.* **29**, 60–75 (1977).
31. Borg, I. & Groenen, P. J. F. *Modern Multidimensional Scaling: Theory And Applications* (Springer Science & Business Media, 2005).
32. Khan, J. et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.* **7**, 673–679 (2001).
33. Hastie, T., Tibshirani, R., Friedman, J. H. & Friedman, J. H. *The Elements of Statistical Learning Data Mining, Inference, And Prediction* (Springer, 2009).
34. James, G., Witten, D., Hastie, T. & Tibshirani, R. *Introduction To Statistical Learning* 2nd edn (Springer, 2021).
General text on methodology for data science, with extensive treatment of PCA in its various forms, including matrix completion.
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.* **12**, 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. 30th Int. Conf. on Machine Learning (PMLR)* **28**, 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 edn (Chapman & Hall/CRC Press, 2016).
Comprehensive treatment of correspondence analysis (CA) and its variants, multiple correspondence analysis (MCA) and canonical correspondence analysis (CCA).
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, 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. *Meth. 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. Abdi, H. & Valentin, D. Multiple correspondence analysis. *Encycl. Meas. Stat.* **2**, 651–657 (2007).
63. Richards, G. & van der Ark, L. A. Dimensions of cultural consumption among tourists: multiple correspondence analysis. *Tour. Manag.* **37**, 71–76 (2013).
64. 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).
65. 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).
66. Greenacre, M. & Pardo, R. Subset correspondence analysis: visualizing relationships among a selected set of response categories from a questionnaire survey. *Sociol. Meth. Res.* **35**, 193–218 (2006).
67. 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).
68. Aşan, Z. & Greenacre, M. Biplots of fuzzy coded data. *Fuzzy Sets Syst.* **183**, 57–71 (2011).
69. Vichi, M., Vicari, D. & Kiers, H. A. L. Clustering and dimension reduction for mixed variables. *Behaviormetrika* **46**, 243–269 (2019).
70. van de Velden, M., Iodice D'Enza, A. & Markos, A. Distance-based clustering of mixed data. *Wiley Interdiscip. Rev. Comput. Stat.* **11**, e1456 (2019).
71. Greenacre, M. Use of correspondence analysis in clustering a mixed-scale data set with missing data. *Arch. Data Sci. Ser. B* https://doi.org/10.5445/KSP/1000085952/04 (2019).
72. Gifi, A. *Nonlinear Multivariate Analysis* (Wiley-Blackwell, 1990).
73. Michailidis, G. & de Leeuw, J. The Gifi system of descriptive multivariate analysis. *Stat. Sci.* **13**, 307–336 (1998).
74. Linting, M., Meulman, J. J., Groenen, P. J. F. & van der Koojj, A. J. Nonlinear principal components analysis: introduction and application. *Psychol. Meth.* **12**, 336–358 (2007).
Gentle introduction to nonlinear PCA for data that have categorical or ordinal variables, including an in-depth application to data of early childhood caregiving.
75. 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).
76. Bock, H.-H., Chouakria, A., Cazes, P. & Diday, E. Symbolic factor analysis. In *Analysis of Symbolic Data* (ed. Bock H.-H. & Diday, E.) 200–212 (Springer, 2000).
77. Lauro, C. N. & Palumbo, F. Principal component analysis of interval data: a symbolic data analysis approach. *Comput. Stat.* **15**, 73–87 (2000).
78. Gioia, F. & Lauro, C. N. Principal component analysis on interval data. *Comput. Stat.* **21**, 343–363 (2006).
79. 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).
The application of PCA to non-atomic coded data, that is, interval or fuzzy data.
80. Makosso-Kallyth, S. & Diday, E. Adaptation of interval PCA to symbolic histogram variables. *Adv. Data Anal. Classif.* **6**, 147–159 (2012).
81. 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).
82. Le-Rademacher, J. & Billard, L. Principal component analysis for histogram-valued data. *Adv. Data Anal. Classif.* **11**, 327–351 (2017).
83. Booysen, F. An overview and evaluation of composite indices of development. *Soc. Indic. Res.* **59**, 115–151 (2002).
84. Lai, D. Principal component analysis on human development indicators of China. *Soc. Indic. Res.* **61**, 319–330 (2003).
85. Krishnakumar, J. & Nagar, A. L. On exact statistical properties of multidimensional indices based on principal components, factor analysis, MIMIC and structural equation models. *Soc. Indic. Res.* **86**, 481–496 (2008).
86. Mazziotta, M. & Pareto, A. Use and misuse of PCA for measuring well-being. *Soc. Indic. Res.* **142**, 451–476 (2019).
87. Fabrigar, L. R., Wegener, D. T., MacCallum, R. C. & Strahan, E. J. Evaluating the use of exploratory factor analysis in psychological research. *Psychol. Meth.* **4**, 272–299 (1999).
88. 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).
89. Wabiri, N. & Taffa, N. Socio-economic inequality and HIV in South Africa. *BMC Public. Health* **13**, 1037 (2013).
90. Lazarus, J. Vetal The global NAFLD policy review and preparedness index: are countries ready to address this silent public health challenge? *J. Hepatol.* **76**, 771–780 (2022).
91. Rodarmel, C. & Shan, J. Principal component analysis for hyperspectral image classification. *Surv. Land. Inf. Sci.* **62**, 115–122 (2002).
92. Du, Q. & Fowler, J. E. Hyperspectral image compression using JPEG2000 and principal component analysis. *IEEE Geosci. Remote. Sens. Lett.* **4**, 201–205 (2007).
93. Turk, M. & Pentland, A. Eigenfaces for recognition. *J. Cogn. Neurosci.* **3**, 71–86 (1991).
94. Paul, L. & Suman, A. Face recognition using principal component analysis method. *Int. J. Adv. Res. Comput. Eng. Technol.* **1**, 135–139 (2012).
95. 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).
96. Ghorbani, M. & Chong, E. K. P. Stock price prediction using principal components. *PLoS One* **15**, e0230124 (2020).
97. Pang, R., Lansdell, B. J. & Fairhall, A. L. Dimensionality reduction in neuroscience. *Curr. Biol.* **26**, R656–R660 (2016).
98. Abraham, G. & Inouye, M. Fast principal component analysis of large-scale genome-wide data. *PLoS One* **9**, e93766 (2014).
99. 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).
Application of PCA to gene expression data, proposing the concepts of eigenarrays and eigengenes as representative linear combinations of original arrays and genes.
100. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
101. Tsuyuzaki, K., Sato, H., Sato, K. & Nikaido, I. Benchmarking principal component analysis for large-scale single-cell RNA-sequencing. *Genome Biol.* **21**, 9 (2020).
102. Golub, G. H. & van Loan, C. F. *Matrix Computations* (JHU Press, 2013).
103. Lanczos, C. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Nat. Bureau Standards* **45**, 255–282 (1950).
104. Baglama, J. & Reichel, L. Augmented GMRES-type methods. *Numer. Linear Algebra Appl.* **14**, 337–350 (2007).
105. Wu, K. & Simon, H. Thick-restart Lanczos method for large symmetric eigenvalue problems. *SIAM J. Matrix Anal. Appl.* **22**, 602–616 (2000).
106. 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).
A comprehensive review of randomized algorithms for low-rank approximation in PCA and SVD.
107. Weng, J., Zhang, Y. & Hwang, W.-S. Candid covariance-free incremental principal component analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **25**, 1034–1040 (2003).
108. 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).
Proposal of incremental implementations of PCA for applications to large data sets and data flows.
109. Cardot, H. & Degras, D. Online principal component analysis in high dimension: which algorithm to choose? *Int. Stat. Rev.* **86**, 29–50 (2018).
110. 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, A., Coli, M. & Angulo Ibanez, J.-M.) 453–463 (Springer, 2012).
111. Iodice D'Enza, A., Markos, A. & Palumbo, F. Chunk-wise regularised PCA-based imputation of missing data. *Stat. Meth. Appl.* **31**, 365–386 (2021).
112. Shiokawa, Y. et al. Application of kernel principal component analysis and computational machine learning to exploration of metabolites strongly associated with diet. *Sci. Rep.* **8**, 3426 (2018).
113. Koren, Y., Bell, R. & Volinsky, C. Matrix factorization techniques for recommender systems. *Computer* **42**, 30–37 (2009).
114. Li, Y. On incremental and robust subspace learning. *Pattern Recogn.* **37**, 1509–1518 (2004).
115. Bouwmans, T. Subspace learning for background modeling: a survey. *Recent Pat. Comput. Sci.* **2**, 223–234 (2009).
116. 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, J. Il & Kim, J.) 315–320 (Springer, 2012).
117. 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).
118. Mazumder, R., Hastie, T. & Tibshirani, R. Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.* **11**, 2287–2322 (2010).
119. Josse, J. & Husson, F. Handling missing values in exploratory multivariate data analysis methods. *J. Soc. Fr. Stat.* **153**, 79–99 (2012).
120. Hastie, T., Tibshirani, R. & Wainwright, M. *Statistical Learning With Sparsity: The LASSO And Generalizations* (CRC Press, 2015).
Comprehensive treatment of the concept of sparsity in many different statistical contexts, including PCA and related methods.
121. 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).
122. 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**, 284 (2018).
123. Ioannidis, A. G. et al. Paths and timings of the peopling of Polynesia inferred from genomic networks. *Nature* **597**, 522–526 (2021).
124. 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).
125. Gower, J. C. Generalized Procrustes analysis. *Psychometrika* **40**, 33–51 (1975).
126. Dryden, I. L. & Mardia, K. V. *Statistical Shape Analysis: With Applications In R* 2nd edn, Vol. 995 (John Wiley & Sons, 2016).
127. Ocaña, F. A., Aguilera, A. M. & Valderrama, M. J. Functional principal components analysis by choice of norm. *J. Multivar. Anal.* **71**, 262–276 (1999).
128. Ramsay, J. O. & Silverman, B. W. Principal components analysis for functional data. In *Functional Data Analysis* 147–172 (Springer, 2005).
129. James, G. M., Hastie, T. J. & Sugar, C. A. Principal component models for sparse functional data. *Biometrika* **87**, 587–602 (2000).

130. Yao, F., Müller, H.-G. & Wang, J.-L. Functional data analysis for sparse longitudinal data. *J. Am. Stat. Assoc.* **100**, 577–590 (2005).
131. Hörmann, S., Kidziński, Ł. & Hallin, M. Dynamic functional principal components. *J. R. Stat. Soc. Ser. B* **77**, 319–348 (2015).
132. 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).
133. 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).
134. Song, J. & Li, B. Nonlinear and additive principal component analysis for functional data. *J. Multivar. Anal.* **181**, 104675 (2021).
135. Tuzhilina, E., Hastie, T. J. & Segal, M. R. Principal curve approaches for inferring 3D chromatin architecture. *Biostatistics* **23**, 626–642 (2022).
136. 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).
137. Allen, J. A. et al. Song complexity is maintained during inter-population cultural transmission of humpback whale songs. *Sci. Rep.* **12**, 8999 (2022).
138. Wiltschko, A. B. et al. Mapping sub-second structure in mouse behavior. *Neuron* **88**, 1121–1135 (2015).
139. Liu, L. T., Dobriban, E. & Singer, A. ePCA: high dimensional exponential family PCA. *Ann. Appl. Stat.* **12**, 2121–2150 (2018).
140. Lê, S., Josse, J. & Husson, F. FactoMineR: an R package for multivariate analysis. *J. Stat. Softw.* **25**, 1–18 (2008).
141. 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).
142. Thioulouse, J. et al. *Multivariate Analysis Of Ecological Data With ade4* (Springer, 2018).
143. Erichson, N. B., Voronin, S., Brunton, S. L. & Kutz, J. N. Randomized matrix decompositions using R. *J. Stat. Softw.* **89**, 1–48 (2019).
144. Iodice D'Enza, A., Markos, A. & Buttarazzi, D. The idm package: incremental decomposition methods in R. *J. Stat. Softw.* **86**, 1–24 (2018).
145. Josse, J. & Husson, F. missMDA: a package for handling missing values in multivariate data analysis. *J. Stat. Softw.* **70**, 1–31 (2016).
146. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
147. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
148. Kidziński, Ł. et al. Deep neural networks enable quantitative movement analysis using single-camera videos. *Nat. Commun.* **11**, 4054 (2020).

Acknowledgements

This review is dedicated to the memory of Professor Cas Troskie, who was the head of the Department of Statistics at the University of Cape Town, both teacher and mentor to M.G. and T.H., and who planted the seeds of principal component analysis in them at an early age. T.H. was partially supported by grants DMS2013736 and IIS1837931 from the National Science Foundation, and grant 5R01 EB001988-21 from the National Institutes of Health. E.T. was supported by the Stanford Data Science Institute.

Author contributions

Introduction (M.G. & T.H.); Experimentation (M.G., P.J.F.G. & T.H.); Results (M.G., P.J.F.G., T.H. & E.T.); Applications (M.G., P.J.F.G., T.H. & E.T.); Reproducibility and data deposition (M.G., A.I.D'E. & A.M.); Limitations and optimizations (M.G., T.H., A.I.D'E., A.M. & E.T.); Outlook (M.G., T.H., A.I.D'E., A.M. & E.T.); Overview of the Primer (all authors).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43586-022-00184-w.

Correspondence should be addressed to Michael Greenacre.

Peer review information *Nature Reviews Methods Primers* thanks Age Smilde, Carles Cuadras and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Related links

amap: https://CRAN.R-project.org/package=amap
elasticnet: https://CRAN.R-project.org/package=elasticnet
fdapace: https://CRAN.R-project.org/package=fdapace
irlba: https://CRAN.R-project.org/package=irlba
Musical illustration of the SVD: https://www.youtube.com/watch?v=JEYLfIVvR9I
onlinePCA: https://CRAN.R-project.org/package=onlinePCA
PCAtools: https://github.com/kevinblighe/PCAtools
pca3d: https://CRAN.R-project.org/package=pca3d
RSDA: https://CRAN.R-project.org/package=RSDA
RSpectra: https://CRAN.R-project.org/package=RSpectra
softImpute: https://CRAN.R-project.org/package=softImpute
stats: https://www.R-project.org/
symbolicDA: https://CRAN.R-project.org/package=symbolicDA
vegan: https://CRAN.R-project.org/package=vegan