POINTS OF SIGNIFICANCE

Principal component analysis

PCA helps you interpret your data, but it will not always find the important patterns.

Principal component analysis (PCA) simplifies the complexity in high-dimensional data while retaining trends and patterns. It does this by transforming the data into fewer dimensions, which act as summaries of features. High-dimensional data are very common in biology and arise when multiple features, such as expression of many genes, are measured for each sample. This type of data presents several challenges that PCA mitigates: computational expense and an increased error rate due to multiple test correction when testing each feature for association with an outcome. PCA is an unsupervised learning method and is similar to clustering[1]—it finds patterns without reference to prior knowledge about whether the samples come from different treatment groups or have phenotypic differences.

PCA reduces data by geometrically projecting them onto lower dimensions called principal components (PCs), with the goal of finding the best summary of the data using a limited number of PCs. The first PC is chosen to minimize the total distance between the data and their projection onto the PC (**Fig. 1a**). By minimizing this distance, we also maximize the variance of the projected points, σ^2 (**Fig. 1b**). The second (and subsequent) PCs are selected similarly, with the additional requirement that they be uncorrelated with all previous PCs. For example, projection onto PC1 is uncorrelated with projection onto PC2, and we can think of the PCs as geometrically orthogonal. This requirement of no correlation means that the maximum number of PCs possible is either the number of samples or the number of features, whichever is smaller. The PC selection process has the effect of maximizing the correlation (r^2) (ref. 2) between data and their projection and is equivalent to carrying out multiple linear regression[3,4] on the projected data against each variable of the original data. For example, the projection onto PC2 has maximum r^2 when used in multiple regression with PC1.

The PCs are defined as a linear combination of the data's original variables, and in our two-dimensional (2D) example, PC1 = $x/\sqrt{2} + y/\sqrt{2}$ (**Fig. 1c**). These coefficients are stored in a 'PCA loading matrix', which can be interpreted as a rotation matrix that rotates data such that the projection with greatest variance goes along the first axis. At first glance, PC1 closely resembles the linear regression line[3] of y versus x or x versus y (**Fig. 1c**). However, PCA differs from linear regression in that PCA minimizes the perpendicular distance between a data point and the principal component, whereas linear regression minimizes the distance between the response variable and its predicted value.

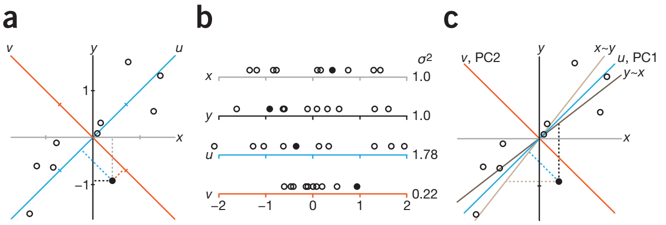

Figure 1 | PCA geometrically projects data onto a lower-dimensional space. (**a**) Projection is illustrated with 2D points projected onto 1D lines along a path perpendicular to the line (illustrated for the solid circle). (**b**) The projections of points in **a** onto each line. σ^2 for projected points can vary (e.g., high for u and low for v). (**c**) PC1 maximizes the σ^2 of the projection and is the line u from **a**. The second (v, PC2) is perpendicular to PC1. Note that PC1 is not the same as linear regression of y vs. x (y~x, dark brown) or x vs. y (x~y, light brown). Dashed lines indicates distances being minimized.

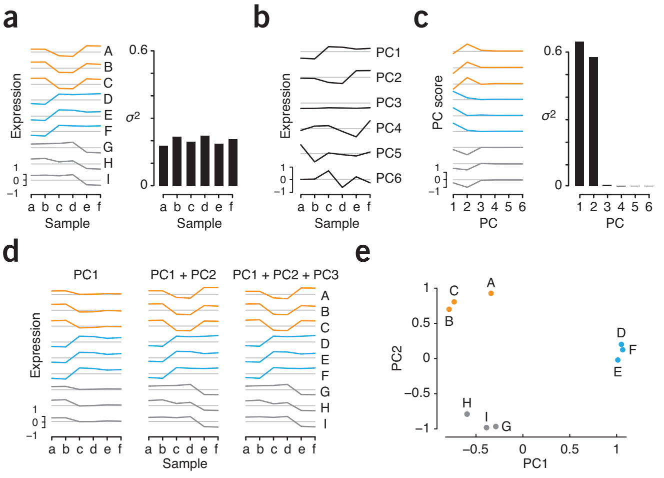

Figure 2 | PCA reduction of nine expression profiles from six to two dimensions. (**a**) Expression profiles for nine genes (A–I) across six samples (a–f), coded by color on the basis of shape similarity, and the expression variance of each sample. (**b**) PC1–PC6 of the profiles in **a**. PC1 and PC2 reflect clearly visible trends, and the remaining capture only small fluctuations. (**c**) Transformed profiles, expressed as PC scores and σ^2 of each component score. (**d**) The profiles reconstructed using PC1–PC3. (**e**) The 2D coordinates of each profile based on the scores of the first two PCs.

To illustrate PCA on biological data, we simulated expression profiles for nine genes that fall into one of three patterns across six samples (**Fig. 2a**). We find that the variance is fairly similar across samples (**Fig. 2a**), which tells us that no single sample captures the patterns in the data appreciably more than another. In other words, we need all six sample dimensions to express the data fully.

Let's now use PCA to see whether a smaller number of combinations of samples can capture the patterns. We start by finding the six PCs (PC1–PC6), which become our new axes (**Fig. 2b**). We next transform the profiles so that they are expressed as linear combinations of PCs—each profile is now a set of coordinates on the PC axes—and calculate the variance (**Fig. 2c**). As expected, PC1 has the largest variance, with 52.6% captured by PC1 and 47.0% captured by PC2. A useful interpretation of PCA is that r^2 of the regression is the percent variance (of all the data) explained by the PCs. As additional PCs are added to the prediction, the difference in r^2 corresponds to the variance explained by that PC. However, all the PCs are not typically used because the majority of variance, and hence patterns in the data, will be limited to the first few PCs. In our example, we can ignore PC3–PC6, which contribute little (0.4%) to explaining the variance, and express the data in two dimensions instead of six.

Figure 2d verifies visually that we can faithfully reproduce the profiles using only PC1 and PC2. For example, the root mean square (r.m.s.) distances of the original profile A from its 1D, 2D and 3D reconstructions are 0.29, 0.03 and 0.01, respectively. Approximations using two or three PCs are useful, because we

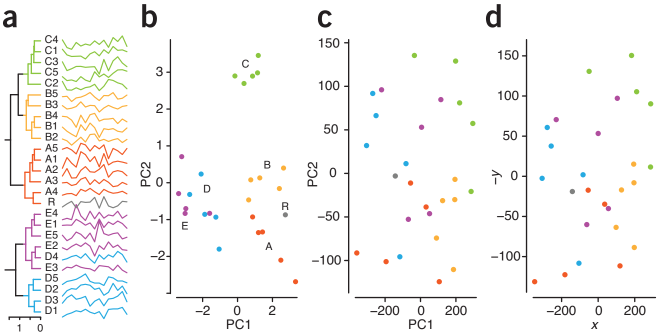

Figure 3 | PCA can help identify clusters in the data. (**a**) Complete linkage hierarchical clustering of previously described expression profiles[1] with the expression of all 26 genes (listed vertically) represented with lines across 15 samples (horizontally). (**b**) When shown as coefficients of the first two PCs, profiles group in a similar manner to the hierarchical clustering—groups D and E are still difficult to separate. (**c**) PCA is not scale invariant. Shown are the first two PC components of profiles whose first and second variable (subject) were scaled by 300 and 200, respectively. A grouping very different from that in **b** is obtained. (**d**) The plot of the two scaled variables in each profile, ignoring the remaining 13 variables. The grouping of points is very similar to that in **c**, because PCA puts more weight on variables with larger absolute magnitude.

can summarize the data as a scatter plot. In our case, this plot easily identifies that the profiles fall into three patterns (**Fig. 2e**). Moreover, the projected data in such plots often appear less noisy, which enhances pattern recognition and data summary.

Such PCA plots are often used to find potential clusters. To relate PCA to clustering, we return to the 26 expression profiles across 15 subjects from a previous column[1], which we grouped using hierarchical clustering (**Fig. 3a**). It turns out that we can recover these clusters using only two PCs (**Fig. 3b**), reducing the dimensionality from 15 (the number of subjects) to 2.

Scale matters with PCA. We illustrate this by showing PC1 and PC2 coefficients of each profile after artificially scaling up the expression in the first two subjects in every profile by factors of 300 and 200 so that they are dominant (**Fig. 3c**). This scenario might arise if expression in the first two subjects was measured using a different technique, resulting in dramatically different variance. In fact, when a small set of variables has a much larger magnitude than others, the components in the PCA analysis are heavily weighted along those variables, while other variables are ignored. As a consequence, the PCA simply recovers the values of these high-magnitude variables (**Fig. 3d**).

If the variance is dramatically different across variables (e.g., expression across patients in the scaled data in **Fig. 3c**), or if the variables use different scales, such as expression and phenotype data, it may be appropriate to standardize them such that each variable has unit variance. However, if the variables are already on the same scale, standardization is not normally appropriate, as it may actually distort the data. For instance, after standardization, gene expression that varies dramatically owing to biological function may look similar to gene expression that varies only owing to noise.

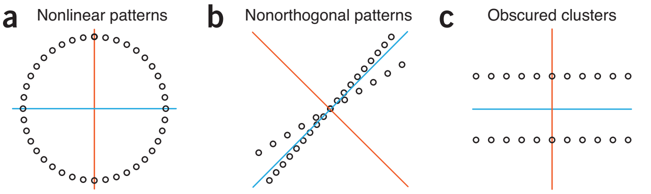

Figure 4 | The assumptions of PCA place limitations on its use. (**a**–**c**) Limitations of PCA are that it may miss nonlinear data patterns (**a**); structure that is not orthogonal to previous PCs may not be well characterized (**b**); and PC1 (blue) may not split two obvious clusters (**c**). PC2 is shown in orange.

PCA is a good data summary when the interesting patterns increase the variance of projections onto orthogonal components. But PCA also has limitations that must be considered when interpreting the output: the underlying structure of the data must be linear (**Fig. 4a**), patterns that are highly correlated may be unresolved because all PCs are uncorrelated (**Fig. 4b**), and the goal is to maximize variance and not necessarily to find clusters (**Fig. 4c**).

Conclusions made with PCA must take these limitations into account. As with all statistical methods, PCA can be misused. The scaling of variables can cause different PCA results, and it is very important that the scaling is not adjusted to match prior knowledge of the data. If different scalings are tried, they should be described. PCA is a tool for identifying the main axes of variance within a data set and allows for easy data exploration to understand the key variables in the data and spot outliers. Properly applied, it is one of the most powerful tools in the data analysis tool kit.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Jake Lever, Martin Krzywinski & Naomi Altman

1. Altman, N. & Krzywinski, M. *Nat. Methods* **14**, 545–546 (2017).
2. Altman, N. & Krzywinski, M. *Nat. Methods* **12**, 899–900 (2015).
3. Altman, N. & Krzywinski, M. *Nat. Methods* **12**, 999–1000 (2015).
4. Krzywinski, M. & Altman, N. *Nat. Methods* **12**, 1103–1104 (2015).

Jake Lever is a PhD candidate at Canada's Michael Smith Genome Sciences Centre. Martin Krzywinski is a staff scientist at Canada's Michael Smith Genome Sciences Centre. Naomi Altman is a Professor of Statistics at The Pennsylvania State University.