1.4 Principle Component Analysis

Principle component analysis (PCA) is a mainstream approach of modern data analysis, often a black box that is widely used but poorly understood. Intuitively speaking, it is a simple, non-parametric method of extracting relevant information from confusing data sets, and thereby provides guidances on how to reduce a complex data set to a lower dimension to reveal the sometimes hidden, simplified dynamics that underlie the data set.

1.4.1 A 2D Example

Assume we have a n samples of a vector with $p = 2$ variables and plot them as shown in Fig 1.3 (left). If we pass a vector through the long axis of the point cloud and a second vector at right angles to the first vector (with both vectors passing through the centroid of the data), we could find the coordinates of all the data points relative to these two *perpendicular* vectors and re-plot the data, as shown in Fig 1.3 (right).

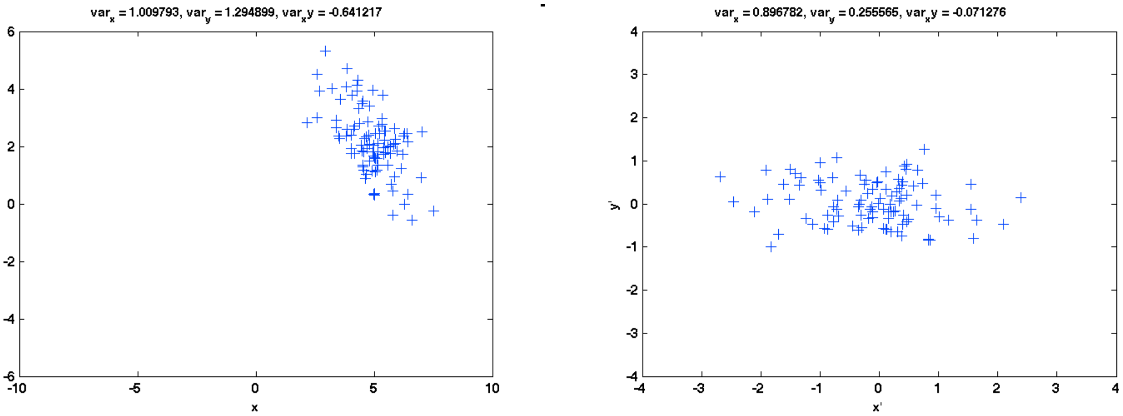


Figure 1.3: A simple 2D example of PCA.

What do we get from this new representation of the data set (the new reference frame)?

- The spatial relationships of the data points are **unchanged**. The above process is merely a rotation of the data set.

- Variance is greater along axis 1 than it is along axis 2. (Compared to the original representation, the variance is roughly the same along the 2 variables.

- The two new axes are perpendicular (uncorrelated). (Compared to the original representation, the two variables are highly correlated).

The two new axes (principle components) might have particular explanations of the data sets, and they generally will not coincide exactly with the original variables. While this 2D example is trivial, these relationships may not be as obvious when one is dealing with many variables. The intuitive interpretation of PCA is to **rotate** the data set in a way such that each successive axis displays a decreasing amount of variance. Based on this, one can reduce the dimensionality of a data set by ignoring the principle components with small variances.

1.4.2 Change of Basis

Now we can state the PCA as the means to compute the most meaningful **basis** to rerepresent a noisy, garbled data set so that the new basis will filter out the noise and reveal hidden dynamics. In a more rigid way, *Is there another basis, which is a liner combination of the original basis, that best represent the data set?*

Assume we have $p \times n$ data matrices $\mathbf{X}$ and $\mathbf{Y}$ related by a linear transformation $\mathbf{P}^T$ (Please bear with the use of a transpose matrix here, the reason of using which will be evident later). $\mathbf{X}$ is the original data set with n samples of the vector $\mathbf{x}_i$ with p variables, and $\mathbf{Y}$ is the re-representation of the data set.

$$\mathbf{P}^T\mathbf{X} = \mathbf{Y}, \quad \text{where} \quad \mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \cdots \mathbf{p}_p] \tag{1.26}$$

This can be expressed as:

$$\mathbf{Y} = \left[\begin{array}{c} \mathbf{y}_1, \cdots \mathbf{y}_n \end{array}\right] = \mathbf{P}^T\mathbf{X} = \left[\begin{array}{c} \mathbf{p}_1^T \\ \vdots \\ \mathbf{p}_p^T \end{array}\right] \left[\begin{array}{c} \mathbf{x}_1 \cdots \mathbf{x}_n \end{array}\right]$$

In other words,

$$\mathbf{y}_i = \left[\begin{array}{c} \mathbf{p}_1^T\mathbf{x}_i \\ \vdots \\ \mathbf{p}_p^T\mathbf{x}_i \end{array}\right]$$

where each coefficient of $\mathbf{y}_i$ is a dot-product of the original vector $\mathbf{x}_i$ with the corresponding row in $\mathbf{P}^T$. In other words, $\mathbf{y}_i$ is a projection of $\mathbf{x}_i$ on to the basis of $\{\mathbf{p}_1^T, \cdots, \mathbf{p}_p^T\}$. Therefore, the rows of $\mathbf{P}^T$ (the columns of $\mathbf{P}$) are a new set of basis for representing $\mathbf{x}_i$.

In this way the problem reduces to find the appropriate **change of basis** and the row vectors in $\mathbf{P}^T$ will become the principal component of $\mathbf{X}$. Now the questions are:

- What is the best way to re-represent $\mathbf{X}$? In other words, what features we would like $\mathbf{Y}$ to exhibit?

1.4.3 Variance: Noise and Redundancy

To answer the question of *what is the best expression of the data*, we look at two important properties of the data that we want to interpret.

• Noise.

We must assume that noise in our dat set are low, i.e., that our measurement devices are reasonably good (Otherwise no matter how advanced the analysis technique is, no information about the system can be extracted!). A common measure is the signal-to-noise ration (SNR)

$$SNR = \frac{\sigma^2_{signal}}{\sigma^2_{noise}}$$

SNR $>> 1$ indicates that the data is very precise, low SNR indicates that the data is contaminated by noise. PCA assumes that we have data with reasonable SNR, and thus determine successive PCs by decreasing variances. Dimensionality reduction of PCs somehow assumes that PCs associated with low variances represent noise.

• Redundancy.

Assume we have two separate features r_1 and r_2 as plotted in Fig 1.4.

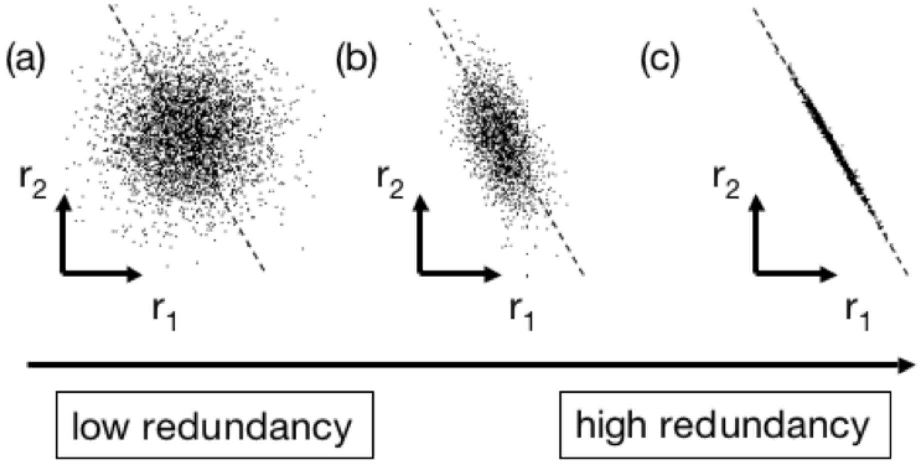


Figure 1.4: Examples of possible redundancies in data from two separate measurement types r_1 and r_2.

Fig 1.4 (a) depicts two variables with no redundancy (uncorrelated). In both Fig 1.4 (b) and (c), the recordings r_1 and r_2 appear to be strongly correlated. Especially for situations in (c), it would be more meaningful to just record a single variable, the linear combination $r_2 - kr_1$, instead of two variables separately. It would express the data more concisely and reduce the number of sensor recordings. This is the very idea behind PCA and dimensionality reduction.

• Covariance matrix.

A simple way to quantify the redundancy between individual measurement types (variables) is to consider the covariance (spread) between the variables. Consider a vector of two variable (two features) $\mathbf{x}_i = [a_i, b_i]^T$, and n sets of measurements are made $i = 1, \cdots, n$. Let:

$$A = \{a_1, a_2, \cdots, a_n\}, B = \{b_1, b_2, \cdots, b_n\}$$

The variance of A and B are individually defined as:

$$\sigma_A^2 = \frac{1}{n-1}\sum_{i=1}^{n}(a_i - \bar{A})^2, \sigma_B^2 = \frac{1}{n-1}\sum_{i=1}^{n}(b_i - \bar{B})^2$$

where $\bar{A}$ and $\bar{B}$ are the mean of each variable:

$$\bar{A} = \frac{1}{n}\sum_{i=1}^{n} a_i, \bar{B} = \frac{1}{n}\sum_{i=1}^{n} b_i$$

The covariance between A and B is:

$$\sigma_{AB}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})$$

Two important properties about the covariance:

- $\sigma_{AB}^2 = 0$ if and only if A and B are entirely uncorrelated.
- $\sigma_{AB}^2 = \sigma_{BA}^2$.

If we write the sets of A and B into row vectors:

$$\mathbf{a} = [a_1 - \bar{A}, a_2 - \bar{A}, , \cdots, a_n - \bar{A},], \mathbf{b} = [b_1 - \bar{B}, b_2 - \bar{B}, \cdots, b_n - \bar{B},]$$

The covariance σ_{AB}^2 can be written as:

$$\sigma_{AB}^2 = \frac{1}{n-1}\mathbf{a}\mathbf{b}^T$$

The covariance matrix of the data set

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] = \left[\begin{array}{c} \mathbf{a} \\ \mathbf{b} \end{array}\right]$$

can be expressed as

$$\mathbf{P_X} = \frac{1}{n-1}\begin{bmatrix} \mathbf{a}\mathbf{a}^T, \mathbf{a}\mathbf{b}^T \\ \mathbf{b}\mathbf{a}^T, \mathbf{b}\mathbf{b}^T \end{bmatrix} = \frac{1}{n-1}\mathbf{X}\mathbf{X}^T$$

Now we can generalize this definition to a $p \times n$ data set $\mathbf{X}$ that represents n samples of the vector $\mathbf{x}$ of p variables. In other words, each column of $\mathbf{X}$ corresponds to the set of features (variables) from one particular trial, and each row of $\mathbf{X}$ correspond to all measurement of one particular feature (variable). The covariance matrix of the data set is:

$$\mathbf{P_X} = \frac{1}{n-1}\mathbf{X}\mathbf{X}^T \tag{1.27}$$

- $\mathbf{P_X}$ is a square, symmetric $p \times p$ matrix
- The diagonal terms of $\mathbf{P_X}$ are the variance of the particular measurement types.
- The off-diagonal terms of $\mathbf{P_X}$ are the covariance between measurement types.

Computing $\mathbf{P_X}$ thus quantifies the correlations between all possible pairs of measurements. Between one pair of measurements, a large covariance corresponds to a situation like Fig 1.4 (c), while zero covariance corresponds to entirely uncorrelated data as in Fig 1.4 (a).

Because of the special property of $\mathbf{P_X}$ to describe the redundancy between pairs of the measurement types (variables), now the question is: *what features do we want to optimize in* $\mathbf{P_Y}$*?*

1.4.4 Solving PCA

The goal of PCA is to reduce redundancy, namely, we would like the covariances between separate measurement types (variables) in $\mathbf{y}_i$ to be zero. Relating to covariance matrix $\mathbf{P_Y}$, it means that we want all off-diagonal terms in $\mathbf{P_Y}$ to be zero. In other words, **removing redundancy diagnolizes $\mathbf{P_Y}$**.

There are many methods for diagonalizing $\mathbf{P_Y}$. For PCA,

- First, PCA assumes that all basis vectors $\{\mathbf{p}_1, \cdots, \mathbf{p}_p\}$ are orthonormal ($\mathbf{p}_i \cdot \mathbf{p}_j = \delta_{ij}$). Recall that $\mathbf{p}_i$ is the row vector in the matrix $\mathbf{P}$ that transforms $\mathbf{X}$ to $\mathbf{Y}$, PCA assumes $\mathbf{P}$ is an orthonormal matrix.
- Second, PCA assumes the directions with the largest variances are the most *important*, or, *principal*

Intuitively, PCA first selects a normalized direction in p-dimensional space, along which the variance in $\mathbf{X}$ is maximized. This is saved as $\mathbf{p}_1$. Again it finds another direction along which the variance is maximized and, because of the orthonormality condition, the search of directions is restricted to all directions perpendicular to all previously selected directions. This continue until p directions are selected. The resulting ordered set of $\{\mathbf{p}_1, \cdots, \mathbf{p}_p\}$ are the principal components.

In the following we will derive the solution to PCA using the basic linear algebra discussed in the earlier part of the class.

• **Goal:** find some orthonormal matrix $\mathbf{P}$ where $\mathbf{Y} = \mathbf{P^T X}$, such that $\mathbf{P_Y} = \frac{1}{n-1}\mathbf{YY}^T$ is diagonalized. The rows of $\mathbf{P}$ are the principal components of $\mathbf{X}$.

• **Relation to eigen-decomposition**:

If we write $\mathbf{P_Y}$ in terms of our target $\mathbf{P}$:

$$\mathbf{P_Y} = \frac{1}{n-1}\mathbf{YY}^T = \frac{1}{n-1}(\mathbf{P}^T\mathbf{X})(\mathbf{P}^T\mathbf{X})^T = \frac{1}{n-1}\mathbf{P}^T(\mathbf{XX}^T)\mathbf{P}^T = \frac{1}{n-1}\mathbf{P}^T\mathbf{AP} \tag{1.28}$$

where $\mathbf{A} = (\mathbf{XX})^T$ is a symmetrix matrix.

Now recall the relationship between eigen-decomposition of a matrix and matrix diagonalization (section 2.3 and 2.4), we know that a real symmetric matrix $\mathbf{A}$ is always diagonalizable (1.19). In other words, there exists an orthogonal matrix $\mathbf{K}$ such that:

$$\mathbf{K^T AK} = \mathbf{\Lambda} \tag{1.29}$$

where Λ is the diagonal matrix consisting of the eigenvalues of $\mathbf{A}$, and each column of $\mathbf{K}$ consists of the corresponding, linearly independent eigenvectors of $\mathbf{A}$.

The relationship between (1.28) and (1.29) is evident.

- The principal component of $\mathbf{X}$, or the rows of $\mathbf{P}^T$, are the eigenvectors of $\mathbf{A} = \mathbf{XX}^T$.
- The $i-$th diagonal value of $\mathbf{P_Y}$ is the variance of $\mathbf{X}$ along the direction $\mathbf{p}_i$, also the scaled $i-$th eigenvalue λ_i of $\mathbf{A}$.

• **Relation to SVD**:

Recall that for an arbitrary $p \times n$ $\mathbf{X}$ with rank r, there exists unitary matrices $\mathbf{U}(p \times p)$ and $\mathbf{V}(n \times n)$ such that:

$$\mathbf{X} = \mathbf{U\Sigma V}^T \tag{1.30}$$

where Σ is an $p \times n$ matrix with entries that are singular values of $\mathbf{X}$. Each individual singular value σ_i equals to the square root of the eigenvalue of matrix $\mathbf{X}\mathbf{X}^T$.

We can rewrite (1.30) into:

$$\mathbf{U^T X} = \mathbf{\Sigma V}^T$$

If we define $\mathbf{Z} = \mathbf{\Sigma V}^T$, $\mathbf{U}^T$ is a change of basis from $\mathbf{X}$ to $\mathbf{Z}$ so that the covariance matrix of $\mathbf{Z}, \mathbf{P_Z} = \frac{1}{n-1}\mathbf{Z}\mathbf{Z}^T = \frac{1}{n-1}\mathbf{\Sigma V^T V \Sigma^T} = \frac{1}{n-1}\mathbf{\Sigma\Sigma^T} = \frac{1}{n-1}\mathbf{\Lambda}$. In other words, the principal component of $\mathbf{X}$, or the rows of $\mathbf{P}^T$, are the left singular vectors $\mathbf{U}$ of $\mathbf{X}$ (or, $\mathbf{P} = \mathbf{U}$).

1.4.5 PCA & Dimensionality Reduction

Performing PCA in practice is quite simple:

1. Organize a data set as an $p \times n$ matrix, where p is the number of measurement types (variables) and n is the number of trials.

2. Substrate off the mean for each variable.

3. Calculate the SVD on $\mathbf{X}$ or the eigen-decomposition on $\mathbf{X}\mathbf{X}^T$.

One benefit of PCA is that we can examine the variances $\mathbf{P_Y}$ associated with the principal components $\mathbf{p}_i$. In practice, we often find the first $k < p$ principal components associated with the largest variances, and then drop off the remaining components. We can conclude that most interesting dynamics occur only in the first k dimensions.

In practice, one common criteria is to ignore principal components at the point at which the next PC offers little increase in the total variance explained. A second criteria is to include all those PCs up to a predetermined total percent variance explained, such as 90%. A third standard is to ignore components whose variance explained is less than the average variance explained, with the idea being that such a PC offers less than one variable's worth of information. A fourth standard is to ignore the last PCs whose variance explained is all roughly equal.

Example 1: Implement the routine for PCA, and apply it on the set of 5842 handwritten 4's (can be downloaded in Mycourses). Examine:

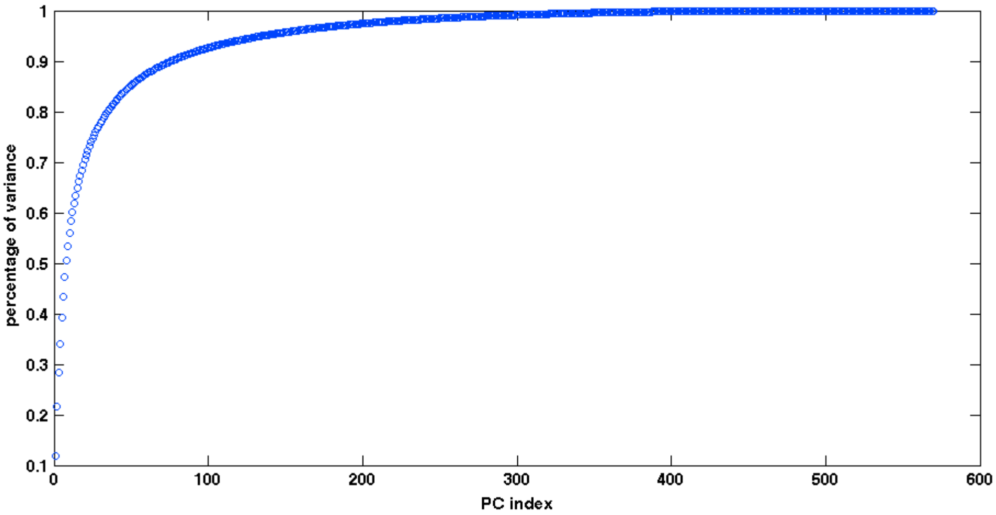
- The decrease of variance accounted by each PC. How many PC's are sufficient for accounting for 90% of the variance?

- How do the lead PCs look like as an image?

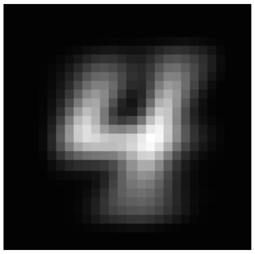
- The linear expansion of a digit image in the PC basis.

You should expect results like this:

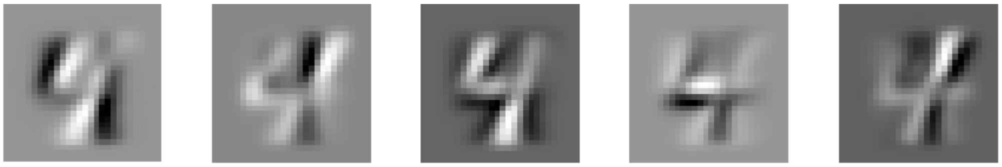
- The first 10 PCs account for about 56% of the variances. 90% of the variance can be accounted for by the first 80 PCs.



- The mean of the dataset:

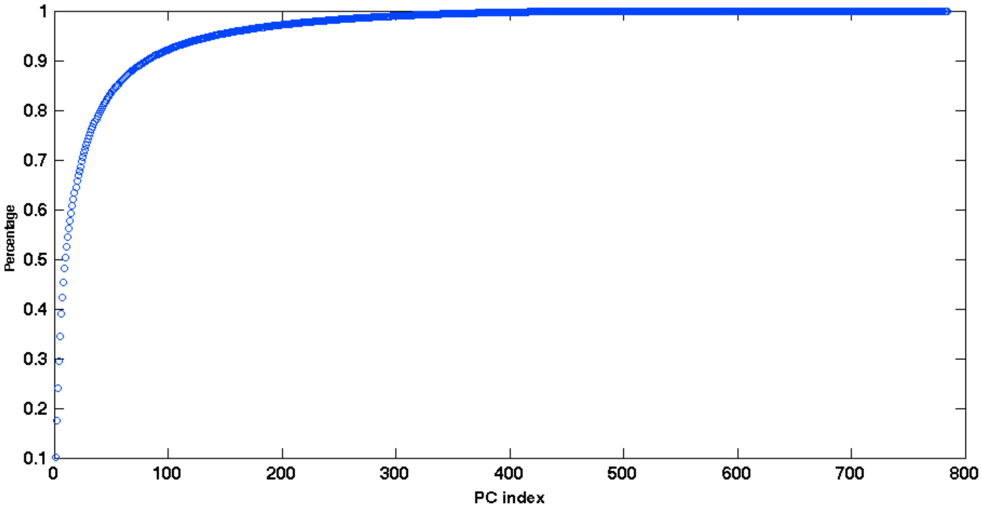


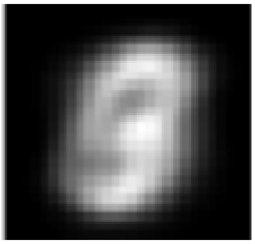
- The first 5 PC's of the dataset:



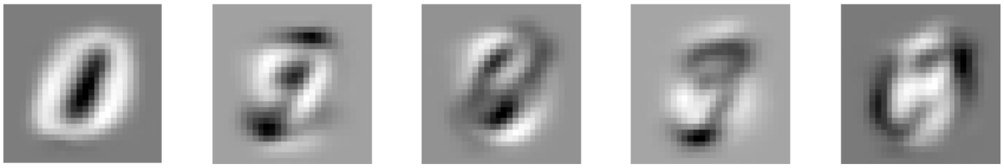
Example 2: Apply it on the set of 2000 handwritten digits that consist of 0 - 9's (can be downloaded in Mycourses) and carry out similar analyses. You should expect results like this:

- The first 10 PCs account for about 50% of the variances. 90% of the variance can be accounted for by the first 82 PCs.





- The mean of the dataset:
- The first 5 PC's of the dataset:



1.4.6 Assumptions, Limitations and Extensions

Both the strength and weakness of PCA is that it is a non-parametric analysis. One only needs to make the assumptions outlined below. There are no parameters to tweak and no coefficients to adjust. The answer is unique and independent of the user. However, it also means that one can not uses *a priori* features of the dynamics of the system.

The limits of PCA are caused by the fundamental assumptions PCA relies on.

- **Linearity**. Linearity frames the problem as a change of basis. Sometimes we can apply a nonlinearity prior to performing PCA: to first convert the data to the appropriately centered polar coordinates and then compute PCA. This prior non-linear transformation is sometimes termed a **kernel** transformation and the entire parametric algorithm is termed **kernel PCA**. This procedure is parametric because the user must incorporate prior knowledge of the dynamics in the selection of the kernel but it is also more optimal in the sense that the dynamics are more concisely described.

- **Gaussian distribution**. PCA assumes mean and covariance are sufficient statistics. The only zero-mean probability distribution that is fully described by the variance is the Gaussian distribution. Therefore in order for this assumption to hold, the probability distribution of $\mathbf{x}_i$ must be Gaussian, wchich also guarantees that the SNR and the covariance matrix fully characterize the noise and redundancies. Deviations from a Gaussian could invaildate this assumption. For example, Fig 1.5 contains a 2D exponentially distributed data set. The largest variances do not correspond to the meaning axes thus PCA fails.

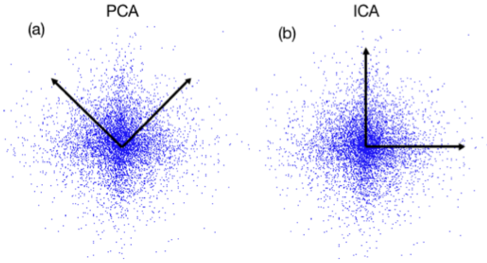


Figure 1.5: Non-Gaussian distributed data causes PCA to fail.

- **Large variance have important dynamics**. This assumption is based on the belief that data has high SNR, so that principal components with larger associated variances represent interesting dynamics, while those with lower variances represent noise.

- **The principal components are orthogonal**.

While PCA is widely applied and easy to use in practice, we should keep in mind the above assumptions it relies on, and therefore the situations it might not apply in practice.

1.5 Independent Component Analysis

$*$ **Reference reading**: Independent Component Analysis (ICA) solves the less constrained set of problems: Find some orthonormal matrix $\mathbf{P}$ where $\mathbf{Y} = \mathbf{P}\mathbf{X}$, such that $\mathbf{P}_{\mathbf{Y}} = \frac{1}{n-1}\mathbf{Y}\mathbf{Y}^T$ is diagonalized.

ICA abandons all assumptions except linearity, and attempt to find axes that satisfy the most formal form of redundancy reduction - **statistical independence**. Intuitively, lack of correlation determines that second-order moments (covariances), while statistical independence determines all of the cross-moments. ICA find a basis such that the joint probability distribution can be factorized:

$$P(\mathbf{y}_i, \mathbf{y}_j) = P(\mathbf{y}_i)P(\mathbf{y}_j)$$

for all i and $j, i \neq j$. ICA is a form of nonlinear optimization, there the solution is difficult to calculate in practice and potentially not unique.