1 Principal Component Analysis

1.1 How it works

Goal of PCA. The main idea of PCA is to reduce the dimensionality of a dataset while preserving as much information as possible. The uses are

1. Visualise high dimensional data and spot patterns or clusters
2.

This texts aims to explain how PCA works using a geometric approach as visualisation helps best to form an understanding on this matter.

Setup. In the following, the setup is briefly laid out.

- Let $\boldsymbol{X} \in \mathbb{R}^{N \times K}$ denote the dataset where there are $N \in \mathbb{N}$ observations and each observation has $K \in \mathbb{N}$ features with $K << N$.
- The n-th observation, denoted by $\boldsymbol{x}'(n)$, is the n-th row of $\boldsymbol{X}$.
- The consistent estimator for the covariance matrix of $\boldsymbol{x}$ is denoted by $\Sigma := \frac{1}{N-1}\boldsymbol{X}'\boldsymbol{X}$.

Additionally, for all Figures there are $K = 2$ features. Obviously, dimensionality reduction is redundant for such low-dimensional data, but this helps to illustrate how PCA works.

Main Mechanism. Consider the data displayed in Figure 1. PCA reduces the dimensionality by considering linear combinations (transformations) of the data that display as much variance as possible. Variance, the measure for dispersion, is maximised because this implies that as much variability and thus information of the data as possible is preserved. For example, as can be seen in Figure 1, the data is most spread out in the x_1 direction, i.e. most information is contained in the x_1 direction.

Consequently, an idea to reduce the dimension would be to simply throw out the second dimension and consider the linearly transformed data $\boldsymbol{\alpha}'\boldsymbol{x}(n)$ with $\boldsymbol{\alpha} = (1, 0)'$. In other words, an orthogonal projection of the data onto the x_1 axis is considered as depicted in Figure 2. The data reduced to its first dimension is depicted in Figure 3.

To illustrate the importance of variance, consider Figure 4 where the data is solely projected onto the x_2 axis. As can be seen, the data is squeezed much more together so that more information is lost compared to Figure 3. In other words, Figure 4 overestimates how close observations in the higher two-dimensional space are whereas Figure 3 more accurately reflects the distance between data points in the two-dimensional space.

Problem of the simple approach. Although of course it is possible to simply take some subset $k_1 < K$ features with the highest variance to reduce the dimension, this is not a good general rule and can lead to bad results in the case of correlated data. To see this, consider Figure 5.

Simply only taking the x_1 component will not be ideal because the data is most stretched out diagonally, that is in a direction that involves *both* x_1 and x_2.

PCA. Therefore, the goal of PCA is to maximise the variance of a linear combination of the features, i.e. $\boldsymbol{\alpha}'\boldsymbol{x}$. These linear combinations of the features are called *Principal Components*. This is captured in the following Definition.

Definition 1.1: First Principal Component

Let $\boldsymbol{\alpha} \in \mathbb{R}^K$ such that $\boldsymbol{\alpha}'\boldsymbol{\alpha} = 1$. $\boldsymbol{\alpha}$ is called the first principal component and solves the following problem

$$\max_{\boldsymbol{\alpha}:\ \boldsymbol{\alpha}'\boldsymbol{\alpha}=1} V(\boldsymbol{\alpha}'\boldsymbol{x}) = \max_{\boldsymbol{\alpha}:\ \boldsymbol{\alpha}'\boldsymbol{\alpha}=1} \boldsymbol{\alpha}'\Sigma\boldsymbol{\alpha}.$$

The restriction $\boldsymbol{\alpha}'\boldsymbol{\alpha} = 1$ is required because otherwise the trivial solution would be to set $\boldsymbol{\alpha} \to \infty$. Moreover, this way the vector $\boldsymbol{\alpha}$ has length 1. This is important because the vector $\boldsymbol{\alpha}$ does not induce any stretching on its own. This will become clearer later.

Proposition 1.1: Solution First Principal Component

The first principal component $\boldsymbol{\alpha}_1$ solves

$$\Sigma\boldsymbol{\alpha}_1 = \lambda_1\boldsymbol{\alpha}_1$$

where λ_1 is the largest eigenvalue (in absolute terms) of Σ.

This is easily derived as can be seen in Jolliffe (2002), p.5[1].

Interpretation. Why is the first principal component the eigenvector corresponding to the largest eigenvalue of Σ? Let's consider a simple case with the following covariance matrix of the data

$$\Sigma = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}.$$

In this case, the features x_1 & x_2 are uncorrelated (consider Figure 1). Indeed, the largest eigenvalue is 3 with the unit-length eigenvector $(1,0)'$. Thus, the linear combination $\boldsymbol{\alpha}_1 = (1,0)'$ as depicted in Figure 3 is actually the best linear combination.

This is the case because the eigenvector $\boldsymbol{\alpha}_1$ with the largest eigenvalue λ_1 is the vector (linear combination of x_1 and x_2) that is most stretched by Σ. The matrix Σ is the relevant matrix because it tells us how x_1 and x_2 stretch individually *and* their co-movement ("co-stretching"). Thus, we want the linear combination of the features that is stretched out the most in order to maintain as much variability as possible.

More Principal Components. For K features we can find K linearly independent combinations of the features to span the entire feature space. Thus, after finding the first linear combination $\boldsymbol{\alpha}_1'\boldsymbol{x}$, we can find a new linear combination $\boldsymbol{\alpha}_2'\boldsymbol{x}$. This will become clearer in the subsequent Figures.

However, this second linear combination should be uncorrelated with the first one because it otherwise includes information already included in $\boldsymbol{\alpha}_1\boldsymbol{x}$. Thus, for the two to be uncorrelated we need to enforce that

$$Cov(\boldsymbol{\alpha}_1'\boldsymbol{x}, \boldsymbol{\alpha}_2'\boldsymbol{x}) = \boldsymbol{\alpha}_1'\Sigma\boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_2'\Sigma\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2'\lambda_1\boldsymbol{\alpha}_1 \stackrel{!}{=} 0 \qquad \Rightarrow \qquad \boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2 = 0$$

Using $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2 = 0$ along with the same no-stretching condition $\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_2 = 1$, one can show that $\boldsymbol{\alpha_2}$ is the eigenvector with the next largest eigenvalue of Σ. This holds for a general number of K features. This is captured in the following proposition.

Proposition 1.2: Solution k-th Principal Component

The k-th principal component $\boldsymbol{\alpha}_K$ satisfies $\boldsymbol{\alpha}_k'\boldsymbol{\alpha}_k = 1$ and $\boldsymbol{\alpha}_k'\boldsymbol{\alpha}_i = 0 \; \forall i \in \{1, 2, \ldots, k\}$ and

$$\Sigma\boldsymbol{\alpha}_k = \lambda_k\boldsymbol{\alpha}_k$$

where λ_k is the k-th largest eigenvalue (in absolute terms) of Σ, i.e. $\lambda_1 > \lambda_2 > \ldots > \lambda_k > \lambda_{k+1} > \ldots > \lambda_K$.

Proof. The proof is simple and can be found in Jolliffe (2002) p.5 and 6.

From this Proposition, the following follows immediately

[1] Notice that for a symmetric positive definite matrix Σ the eigenvalues and corresponding vectors are always real and distinct, so that the solution is well-defined.

Proposition 1.3: Orthonormal Basis

All K principal components form an orthonormal basis of $\mathbb{R}^K$.

Proof. This follows immediately from the fact that all principal components have unit length and are uncorrelated with each other.

Furthermore, we can see how much variance a principal component explains due to the following Proposition.

Proposition 1.4: Variance Principal Component

For the k-th principal component it holds that

$$V(\boldsymbol{\alpha}_k \boldsymbol{x}) = \lambda_k$$

Proof. This is immediately seen by computing

$$V(\boldsymbol{\alpha}_1' \boldsymbol{x}) = \boldsymbol{\alpha}_1' \Sigma \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_1 \lambda_1 \boldsymbol{\alpha}_1 = \lambda_1 \qquad \square$$

Thus, the first principal component captures the largest amount of variance as it is the eigenvector with the largest eigenvalue of Σ. To obtain a relative statement, we can rank the importance of a principal component by considering

$$\frac{\lambda_k}{\sum_{i=1}^{K} \lambda_i}.$$

This expression tells us the fraction of variance explained by the k-th principal component.

Visualisation. In the following, PCA is visualised for the data depicted in Figure 5. The covariance matrix of the data is given by

$$\Sigma = \begin{pmatrix} 3.3 & 1.6 \\ 1.6 & 0.9 \end{pmatrix}$$

The largest eigenvalue is $\lambda_1 = 4.1$ with eigenvector $\boldsymbol{\alpha}_1 = (0.9, 0.45)$.

- Notice that the eigenvector's direction is not unique, i.e. $-\boldsymbol{\alpha}_1$ is also stretched by a factor of 4.1.

The second eigenvalue is $\lambda_2 = 0.08$ with $\boldsymbol{\alpha}_2 = (-0.45, 0.9)$. Thus, using the linear combination $\boldsymbol{\alpha}_1' \boldsymbol{x} = 0.9x_1 + 0.45x_2$ captures the largest amount of the variance of the two-dimensional data.

In fact, the first principal component captures

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} \approx 98\%$$

of the variance.

Figure 6 plots the data along with the principal components $\boldsymbol{\alpha}_1$ (orange) and $\boldsymbol{\alpha}_2$ (green). To enhance visibility, the principal components are scaled up in the plot *beyond unit length*. Additionally, the principal components are drawn from the origin because if we project the data onto the components then $\boldsymbol{\alpha}_i' \mathbf{0} = 0$, but technically they can be drawn anywhere as the only aspect that matters is the linear combination of x_1 and x_2, i.e. the direction of the orange and green vector.

Furthermore, notice in Figure 6 that the principal components are perpendicular (orthogonal) to each other as they have no covariance due to $\boldsymbol{\alpha}_1' \boldsymbol{\alpha}_2 = 0$. Therefore, it is clear that the two principal components span the entire $\mathbb{R}^2$ and thus form an orthonormal basis. so that the data $\boldsymbol{X}$ has exactly two principal components. In other words, any data point $\boldsymbol{x}(n)$ can be reached by a linear combination of the two principal components.

PCA Visualisation - Reducing the Dimension. To reduce the dimension, we consider the orthogonal projection of the data onto the first principal component. To visualise how this works, we can rotate the plane in Figure 6 such that the principal components form the new basis vectors. To do so, define the following matrix

Definition 1.2: Eigenvector Matrix

Define the matrix A of dimension 2×2 to be

$$A \equiv [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2]$$

such that the i-th column of A is the i-th principal component.

It is immediate so see the following

Proposition 1.5: Inverse Eigenvector Matrix

It holds that

$$A^{-1} = A'$$

Proof. This follows from the fact that $\boldsymbol{\alpha}_i'\boldsymbol{\alpha}_i = 1$ and $\boldsymbol{\alpha}_i'\boldsymbol{\alpha}_j = 0$ and can be quickly verified by computing $A'A$, i.e.

$$\begin{bmatrix} \boldsymbol{\alpha}_1' \\ \boldsymbol{\alpha}_2' \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_1 & \boldsymbol{\alpha}_2 \end{bmatrix} = \begin{pmatrix} \boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 & \boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2 \\ \boldsymbol{\alpha}_2'\boldsymbol{\alpha}_1 & \boldsymbol{\alpha}_2'\boldsymbol{\alpha}_2 \end{pmatrix} = I \qquad \square$$

The above Proposition always holds if the matrix A consist of orthonormal vectors which the principal components are by design.

It is instructive to rotate the plane to see how the data is projected onto the principal components. To transform the plane such that the principal components form the new basis vectors, we want the first principal component $\boldsymbol{\alpha}_1$ to be the first basis vector $\boldsymbol{e}_1 = (1, 0)'$ and the analogous for the second principal component. This is easy to achieve since

$$A\boldsymbol{e}_j = \boldsymbol{\alpha}_j \qquad \Leftrightarrow \qquad A^{-1}\boldsymbol{\alpha}_j = \boldsymbol{e}_j.$$

Thus, we most rotate the entire plane by A^{-1}. Notice that the transformation A^{-1} is indeed only a rotation due to the following.

The columns of A^{-1} have length one which is why A^{-1} is purely a rotation. To see this, consider the matrix A. Since the j-th column of any matrix tells us where the basis vector $\boldsymbol{e}_j$ will point to, the basis vectors are not stretched when transformed by A, so that the matrix A only rotates the plane. This is also exactly what we want as due to $\boldsymbol{\alpha}_i'\boldsymbol{\alpha}_i = 1$ we specified that the principal components do not stretch the data on their own. Therefore, when applying the inverse of A, geometrically it should only undo the rotation that A induced so that the application of A^{-1} will not stretch the plane in any way. This is captured in the following Proposition.

Proposition 1.6: Length Columns of A^{-1}

The columns of A^{-1} have unit length. Moreover, the columns of A^{-1} form an orthonormal basis.

Proof. We know that $A^{-1} = A'$. Thus, the columns of A^{-1} are the rows of A. Therefore, we must only prove that the rows of A have unit length. We can then simply compute AA' where the i-th diagonal element is the dot product of the i-th row of A with itself. Since $A' = A^{-1}$, it follows that $AA' = I$ so that indeed each diagonal element has length one. Furtheremore, since the dot product of two different rows always equals 0, the proof is complete.

Figure 7 depicts the rotation induced A^{-1} to the plane (and thus rotating each data point by $\boldsymbol{x}_r(n) = A^{-1}\boldsymbol{x}(n)$).

As can be seen, the new basis vectors are the principal components. The old basis vectors are kept in dotted lines to highlight the rotation A^{-1} induced. After this transformation it can clearly be seen that the first principal component captures the most variance of the data.

As in Figure 2 we can now orthogonally project the data onto the first principal component which is portrayed in Figure 8. When rotating the image back by applying A, Figure 9 depicts how the final image looks like if the original data is orthogonally projected onto the first principal component.

Notice that the projection onto the first principal component as depicted in Figure 9 is also be achieved by computing the usual formula for an orthogonal projection, i.e.

$$\boldsymbol{x}_{p1}(n) = \frac{\boldsymbol{\alpha}_1'\boldsymbol{x}(n)}{\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1}\boldsymbol{\alpha}_1.$$

Since the principal components form an orthonormal basis, we indeed have

$$\boldsymbol{x}(n) = \frac{\boldsymbol{\alpha}_1'\boldsymbol{x}(n)}{\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1}\boldsymbol{\alpha}_1 + \frac{\boldsymbol{\alpha}_2'\boldsymbol{x}(n)}{\boldsymbol{\alpha}_2'\boldsymbol{\alpha}_2}\boldsymbol{\alpha}_2.$$

If the principal components were correlated, i.e. $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2 \neq 0$, then this simple projection formula would not apply.

Figure 1: Simple Data

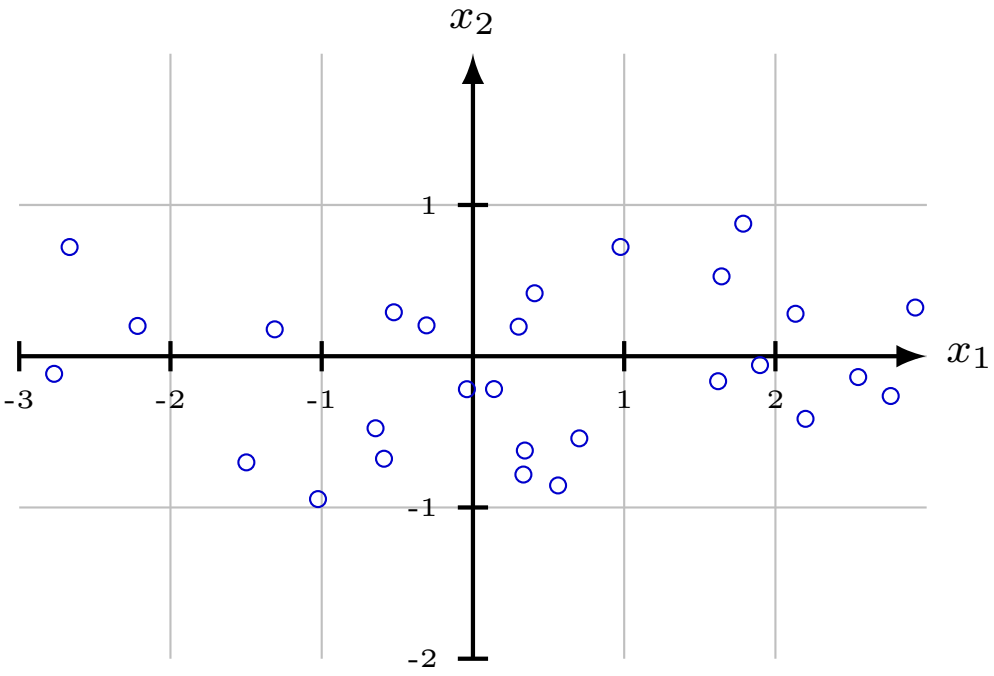


Figure 2: Projection

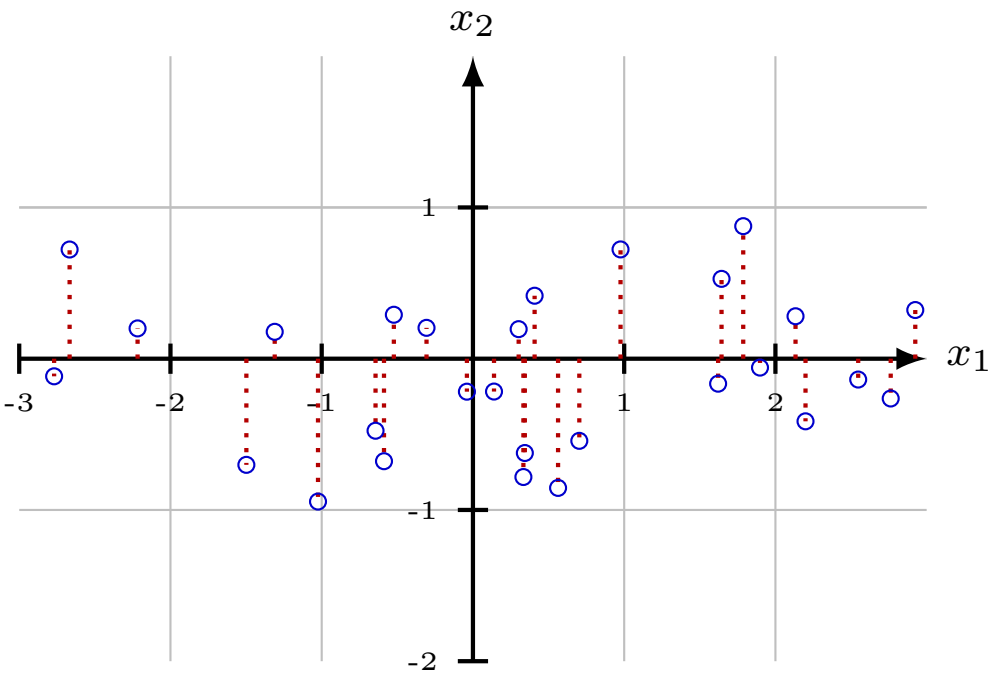


Figure 3: Dimension reduced data onto x_1

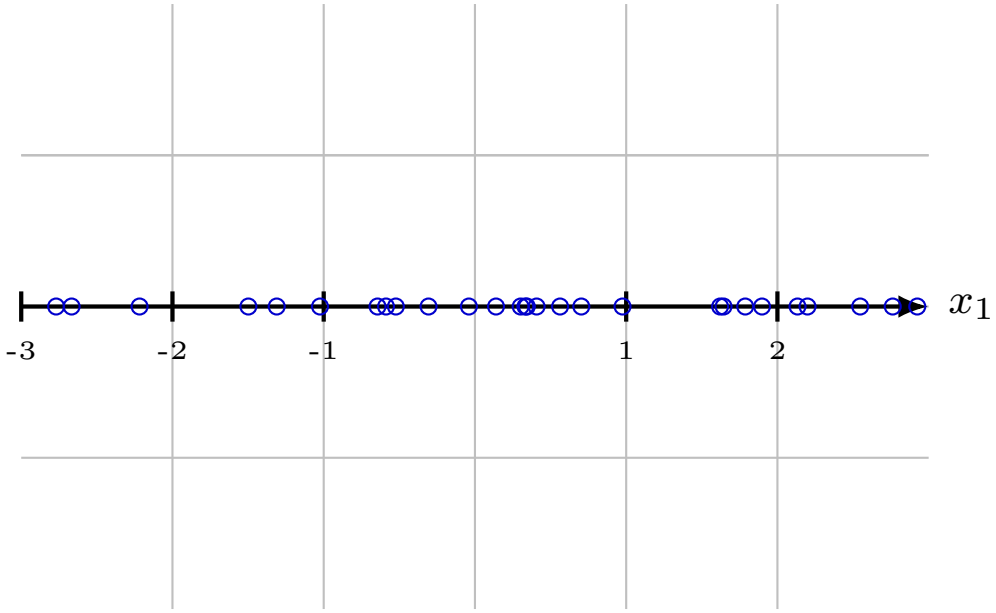


Figure 4: Dimension reduced data onto x_2 (higher information loss)

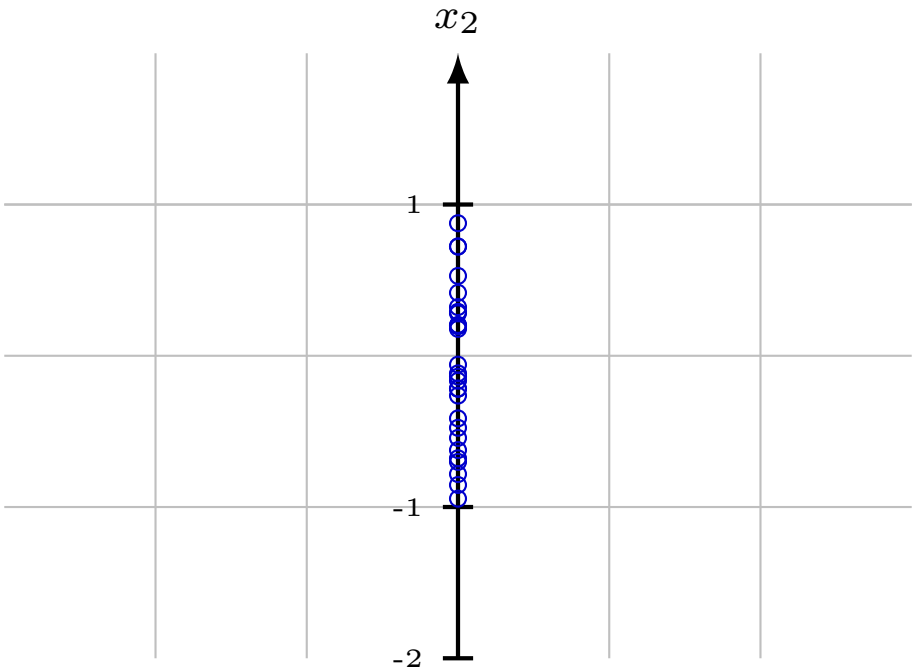


Figure 5: Illustration complex data

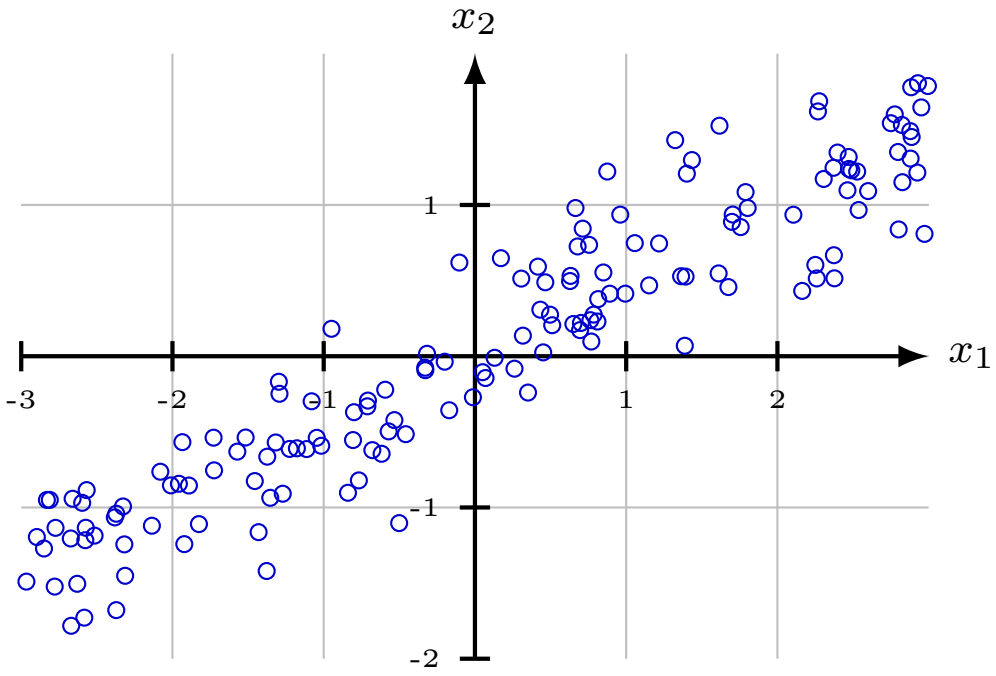


Figure 6: Illustration PCs

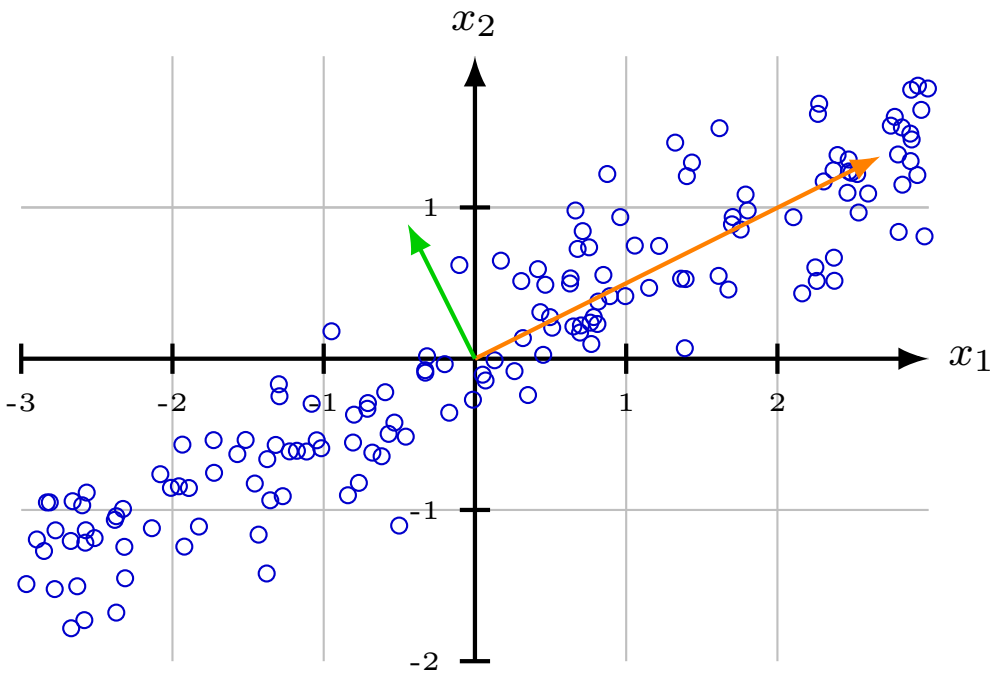


Figure 7: PCA Coordinate Transformation

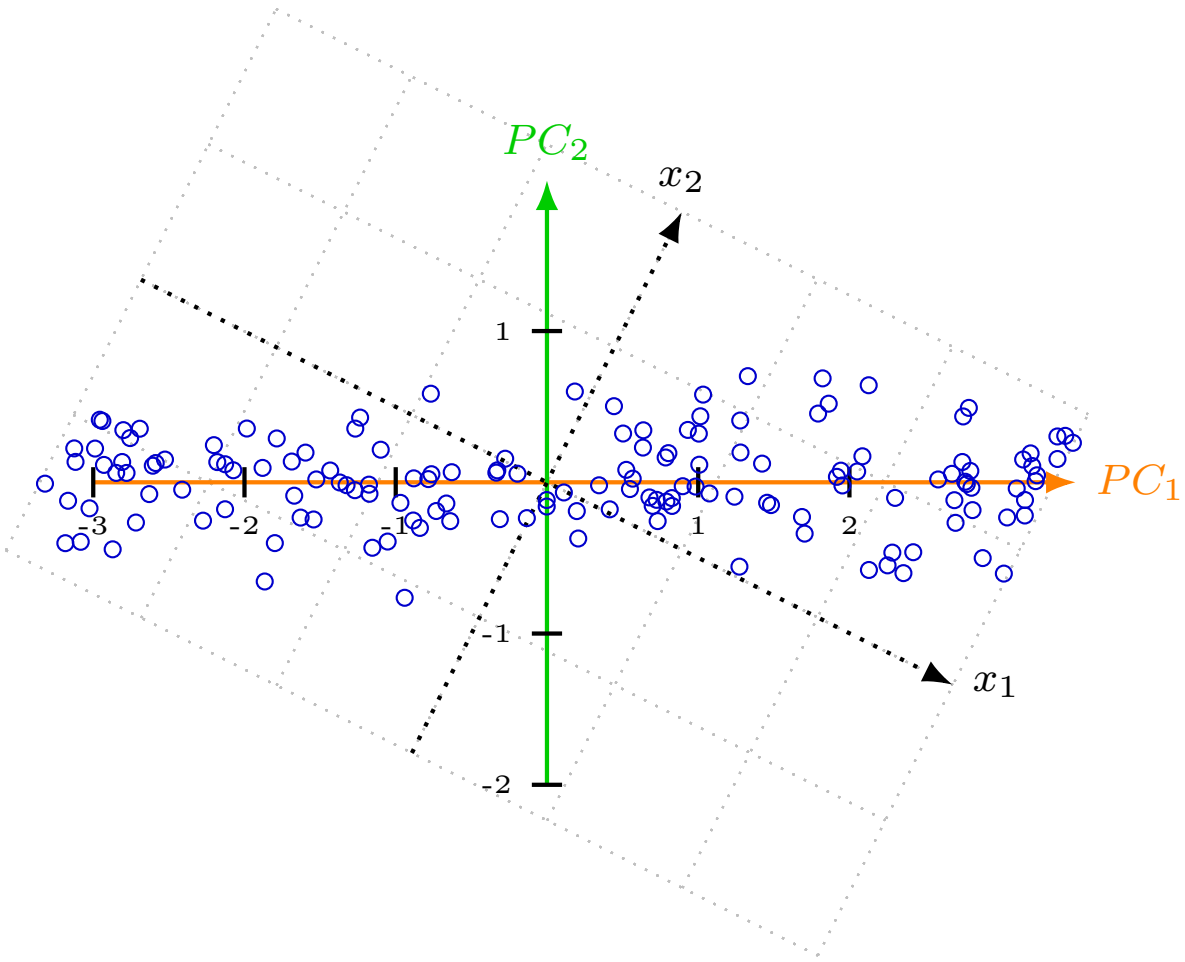


Figure 8: PCA Rotation Orthogonal Projection

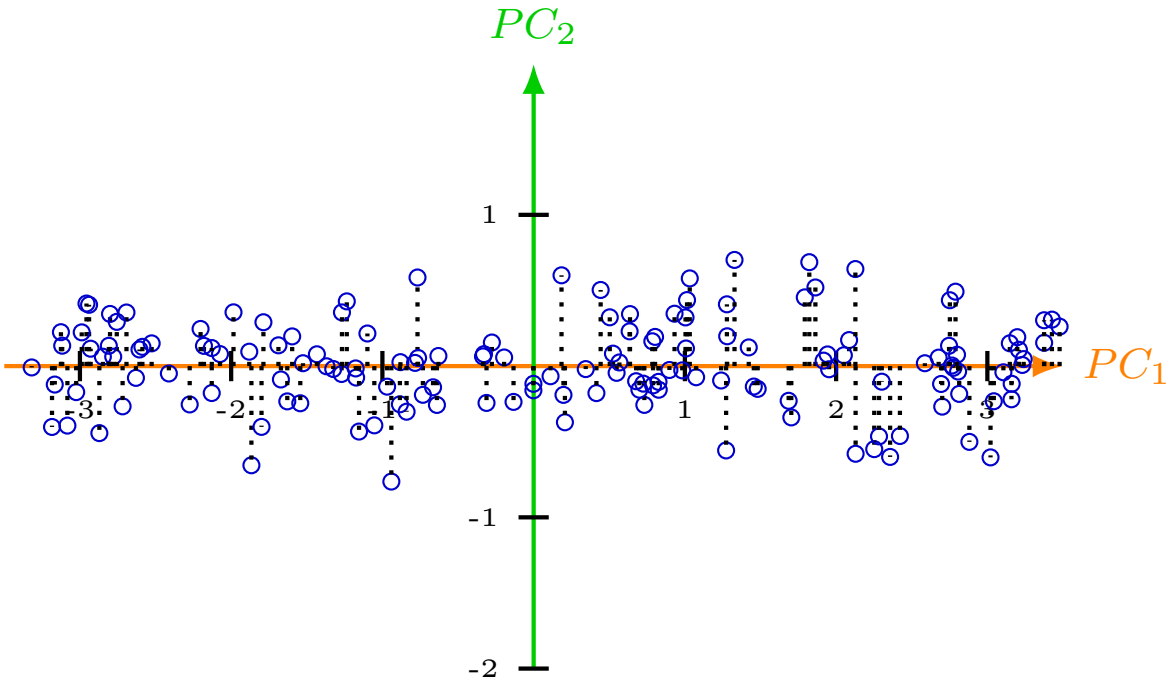
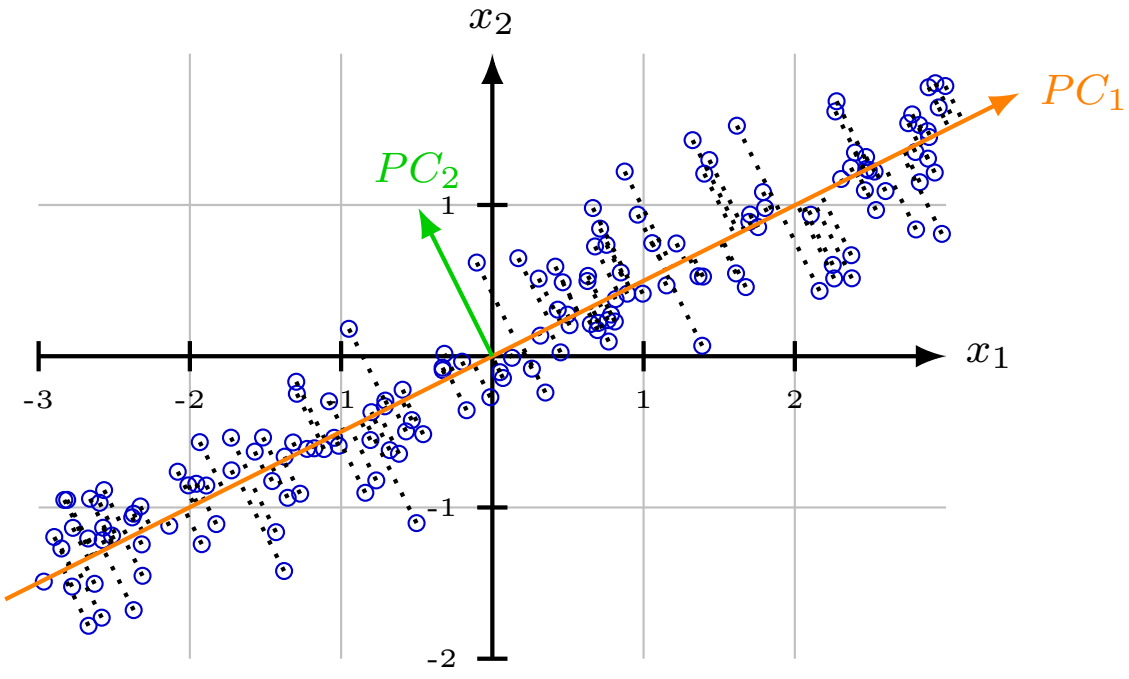


Figure 9: PCA Orthogonal Projection of Data



References

Jolliffe, I. (2002), *Principal Component Analysis*, Springer New York.