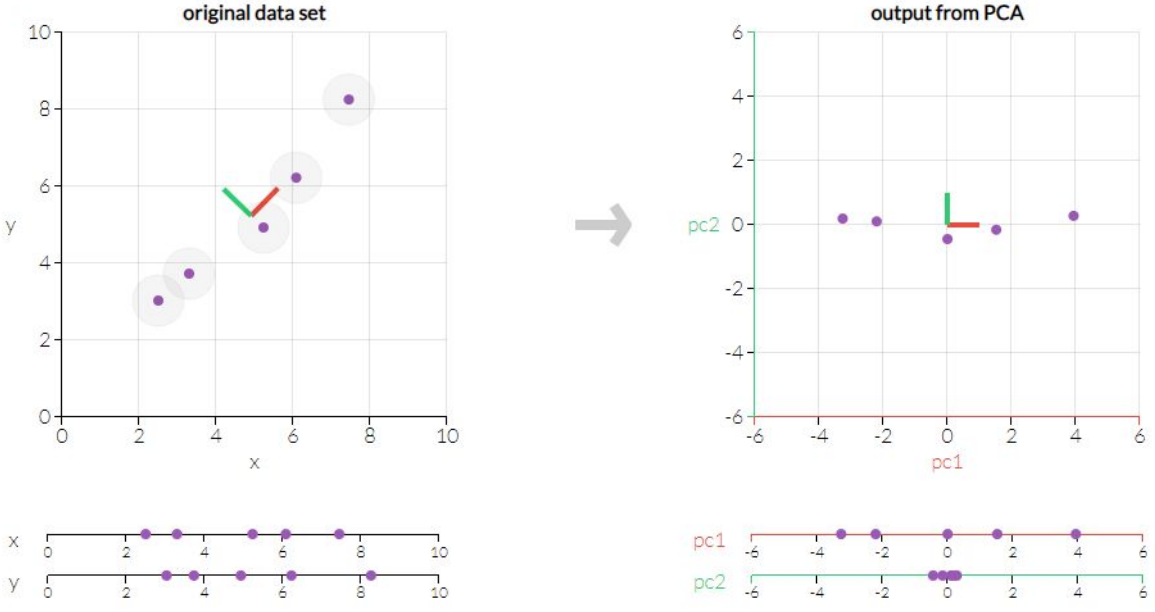
Principal Component Analysis (PCA)

PCA is a method that uses matrix operations from linear algebra and statistics to calculate a projection of the original data into the same number or fewer dimensions. Mathematically, the principal components are the eigenvectors of the covariance matrix of the original dataset. Because the covariance matrix is symmetric, the eigenvectors are orthogonal. Conceptually PCA is trying to find the axes with maximum variances where the data is most spread - within a class, since PCA treats the whole data set as one class.

PCA steps

1. Take the whole dataset consisting of d-dimensional samples ignoring the class labels
2. Compute the d-dimensional mean vector (i.e., the means for every dimension of the whole dataset)
3. Scale the variables to unit variance
4. Subtract the mean vectors from the samples X to center the data at the coordinate system's origin
5. Compute the covariance matrix of the whole dataset
6. Compute eigenvectors and corresponding eigenvalues
7. Sort the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues to form a $d \times k$ dimensional matrix W (where every column represents an eigenvector)
8. Use this $d \times k$ eigenvector matrix to transform the samples onto the new subspace.
   $y = W^T \times x$
   where x is a $d \times 1$-dimensional vector representing one sample, and y is the transformed $k + 1$-dimensional sample in the new subspace.)

If $k < d$ then there is a dimensional reduction.



We don't lose much information by dropping PC2 since it contributes the least to the variation in the data set.

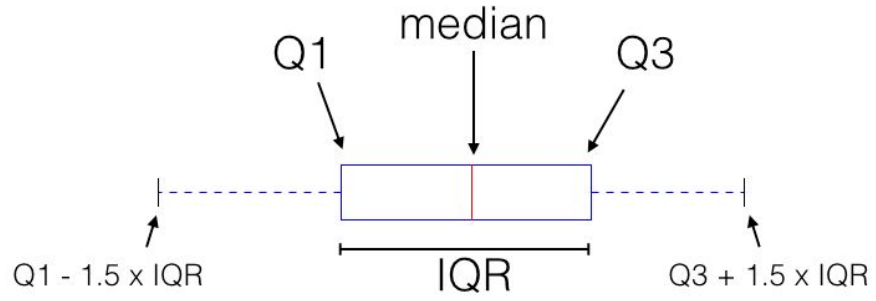
Covariance matrix

The covariance matrix is a $d \times d$ matrix where each element represents the covariance between two features. Covariance provides a measure of the strength of the correlation.

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^T \quad \text{where} \quad \bar{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i$$

Tukey's Method for identifying outliers

An outlier is a data point that lies an abnormal distance from other data points of the same feature. Tukey's rule says that the outliers are values more than 1.5 times the interquartile range (IQR) from the quartiles — either below Q1 − 1.5IQR, or above Q3 + 1.5IQR.



A data point outside the range $[Q1 - 1.5 \times IQR;\ Q3 + 1.5 \times IQR]$ is considered abnormal, where Q1 is the 25th percentile and Q3 the 75th percentile of the data.

Clustering

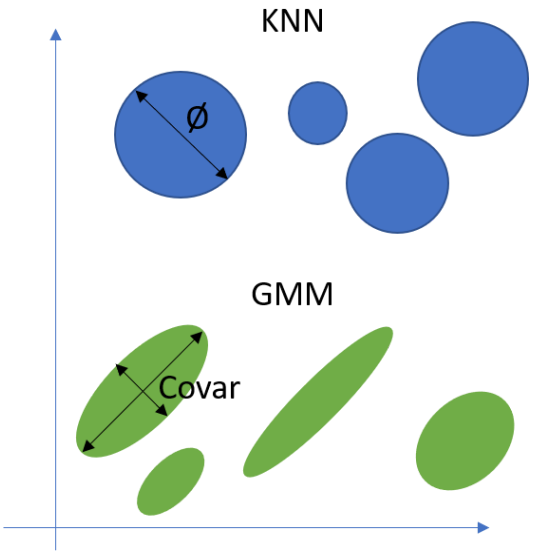
There are two clustering methods:

- K-Means clustering: Each data point is assigned to the closest cluster in terms of Euclidian distance (hard assignment)
- Gaussian Mixture Model (GMM) clustering: Each data point is assigned a probability for each cluster, indicating how likely it belongs to the cluster (soft assignment)

GMMs can fit more complex cluster shapes since each mixture component can freely fit its covariance matrix. K-means treats distances in all directions equally.

The advantage of k-means compared to GMMs is that it is much faster. For GMMs, many parameters must be fitted to the data (quadratic in the number of features) while k-means only maintains cluster centers (linear in number of features). Therefore k-means will be much quicker to train.

K-means can be seen as a special (limit) case of GMMs, specifically as GMM with diagonal, equal and small covariance matrices. Hence, if you constrain the covariance matrices a lot, you will get results similar to k-means.



Silhouette coefficient

The silhouette coefficient for a data point measures how similar it is to its assigned cluster from -1 (dissimilar) to 1 (similar). Calculating the mean silhouette coefficient provides for a simple scoring method of a given clustering.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \qquad s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)}, & a(i) < b(i) \\ 0, & a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1, & a(i) > b(i) \end{cases}$$

where

- $a(i)$: Average distance between object i and the objects in its cluster A
- $b(i)$: Average distance between object i and the objects in its second closest cluster B
- $s(i)$: Silhouette of i measures how good the assignment of i to its cluster is
    - -1 : bad, on average closer to members of B
    - 0 : in between A and B
    - +1 : good assignment of i to its cluster A
- Silhouette Coefficient = Average silhouette of all objects
    - 0.7 < Silhouette Coefficient <= 1.0 : *Strong* structure
    - 0.5 < Silhouette Coefficient <= 0.7 : *Medium* structure
    - 0.25 < Silhouette Coefficient <= 0.5 : *Weak* structure
    - Silhouette Coefficient <= 0.25 : *No* structure

Depending on the problem, the number of clusters that you expect to be in the data may already be known. When the number of clusters is not known a priori, there is no guarantee that a given number of clusters best segments the data, since it is unclear what structure exists in the data. Calculating the mean silhouette coefficient provides for a simple scoring method of a given clustering.