Principal Component Analysis

PCA, which stands for Principal Component Analysis, is a dimensionality-reduction method used for reducing the dimensionality of large data sets and is the most popular algorithm. This is done by transforming a large set of variables into a smaller one, but it still contains most of the information from the larger set. The cost of reducing the data set is that the data set becomes less accurate, but what is gained is that it becomes simpler and therefore easier to explore, visualise and analysing the data becomes much easier and faster for machine learning algorithms. Principal components are new variables that are constructed as linear combinations of the initial variables. The combinations are done so that the new variables are uncorrelated and most of the information from the original variables is compressed into the first components. Compressing the information allows one to reduce the dimensionality without losing much information. Principal components represent the data directions that explains a maximal amount of variance, meaning the lines that capture the most information of the data. The larger the variance carried by a line, the larger the dispersion of the data points along it, and the larger the dispersion the more information it holds. The first principal component always accounts for the largest possible variance. The second principal component is always uncorrelated to the first principal component and accounts for the second highest variance. This continues until all needed principal components have been calculated and has to match the number of original variables. It is possible to do two things when talking about the number of components needed. One can either just take two and use this going forward or one can calculate the best possible number of components needed and continue using those.

PCA can be broken down into different stages:

1. Standardize the range of continuous initial variables.
    - This step is done before PCA. This is done so that each variable contributes equally to the analysis and because PCA is quite sensitive to variances of the variables from the beginning. Meaning, that if a variable has a large range it will have a bigger impact on the result whereas a variable with a small range will have very little impact on the result. When the standardization is done, all variables will be transformed to the same scale. The standardization can be done mathematically by subtracting the mean and dividing it by the standard deviation for each value of each variable.

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

2. Compute the covariance matrix to identify correlations.
    - Here the goal is to see if there is a relationship between the variables of the input data set and how it varies from the mean. To identify these correlations between the pairs of variables we have to compute the covariance matrix. If the covariance is positive, then the two variables increase or decrease together, meaning that they are correlated. If the covariance is negative, then it means that when one increases the other one decreases and therefore, they are inversely correlated. The calculation for the covariance matrix is as follows:

$$C = \frac{1}{n-1} Z * Z^T$$

3. Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components.
   - The eigenvectors and eigenvalues are computed from the covariance matrix in order to determine the principal components of the data. Every eigenvector has an eigenvalue, meaning they are always in pairs. Their number is equal to the number of dimensions in the data set. *"The eigenvectors of the covariance matrix are the directions of the axes where there is the most variance".* Eigenvalues are the coefficients attached to eigenvectors, giving the amount of variance carried in each principal component. The way to find the components order of significance is to rank the eigenvectors in order of their eigenvalues from the highest to the lowest. After the percentage of variance accounted by each component is computed, the eigenvalue of each component is divided by the total sum of eigenvalues.

4. Create a feature vector to decide which principal component to keep.
   - In this stage one has to decide if the components of lesser significance, meaning low eigenvalue, are to be discarded or if all of the components are kept. Afterwards a matrix of vectors is formed with the values, which is called *Feature vector*. The feature vector is a matrix where the kept eigenvectors are set as the columns, making it the first step to dimensionality reduction. If the dimensionality is reduced because of discarding, then the result at the end will be affected by this. Here it has to be decided when the eigenvector holds to much information to be discarded. If it is for example only 4%, then it does not have a big impact as you still keep the remaining 96% of information.

5. Recast the data along the principal component axes.
   - The aim here is to use the feature vector to reorient the data from the original axes to the ones represented by the principal components. This is done by multiplying the transpose of the feature vector.

$$FinalDataSet = FeatureVector^T * StandardizedOriginalDataSet^T$$

The data set can for example be measurements describing properties of production samples, chemical compounds or reactions, process time points of a continuous process or batches from a batch process. An actual example for when to use PCA could be to help identify correlations between data points, meaning seeing if there in Nordic countries is a correlation between consumption of foods like frozen fish and crisp bread.

To summarize it, PCA and the idea behind it is to reduce the number of variables of a given data set, but still preserving as much information as possible as it has an effect on the final result.

Sources:

https://builtin.com/data-science/step-step-explanation-principal-component-analysis

https://www.sartorius.com/en/knowledge/science-snippets/what-is-principal-component-analysis-pca-and-how-it-is-used-507186

https://songxia-sophia.medium.com/principle-components-analysis-pca-essence-and-case-study-with-python-43556234d321