Principal Component Analysis

Maciej Staniszewski and Adam Foster

Executive Summary

Principal Component Analysis (PCA) is a technique used for reducing the dimensionality of a dataset allowing for a more intuitive data exploration and analysis.

Table of Contents

Problem Description

In finance, similarly to other domains of science which are heavily reliant on concepts of data analysis and exploration, we are often facing large sets of data. These can be often difficult to analyse manually, even when plotted graphically. Noticing recurring patterns is highly impractical when the number of variables in the data exceeds reasonable bounds.

Interpreting data on aggregate level is especially difficult. Estimating joint distributions is an elementary step when determining the data generating process behind the data set. This in turn is a necessary step in inference-based thinking and causal discovery in data. Dimensionality curse is resticting us to three easily observable dimensions. When analysing complex data sets consisting of hundreds of variables, it is crucial that a dimensionality reducing operation is performed, enabling scientists to consider these joint distributions of every variable in a simplified form.

A very well known method of representing variables for such analysis is the Principal Component Analysis, commonly known as PCA[1]. The main idea behind it is to create new variables from existing ones, that maximize the ratio of original observed variance to the variance captured by them. To do that, n-dimensional dataset needs to be reduced to k-dimensional one, where k is smaller than n. The coordinates of that new system are the namesake principal components. They represent directions in n-dimensional space which retain the most of the original variance when casting the data points onto them. The first vector created this way is said to explain the biggest portion of original variance and is thus called the first principal component.

Overview of the Data

Methodology Used

Results

1 Principal Component Analysis - Wikipedia

Summary and Recommendations