Seminar: Manifold Learning – Modern Approaches for Dimensionality Reduction

Topic: Generalized PCA

Talk: Alexander Pohl, Review: Roxana Shaikh, Supervisor: Fabian Scheipl

Author: Alexander Pohl,
Matrikelnummer: 12180809

December 20, 2020

Abstract

In this essay, which belongs to a series of talks contributed to the seminar on Manifold Learning: Modern Approaches for Dimensionality Reduction, we discuss generalizations of the in multivariate statistics well known and frequently applied principal component analysis (PCA) method in the context of manifold learning techniques.

A very simplistic definition of manifold learning describes a group of algorithms with the objective to recover a low-dimensional manifold embedding of a high-dimensional ambient space. Thereby reducing the count of co-variables of a high-dimensional dataset towards a small number of essential variables containing most of the information concerning the structure and interrelation of a given dataset. Criteria for an optimal solution can be the distance between points in the high-dimensional space as well as the amount of the total variance of the dataset explained by the manifold embedded. The mahalanobis distance is, beside others, a frequently used example for the measurement of distance and variance serves in the context of information theory as a measure of entropy (loss of information in a transmitted signal or message) of given data.

This essay starts with a short recap of the fundamental PCA method and a summary of some basics of linear algebra, necessary for understanding the principals. A theoretical assertion of the method from a statistical point of view and an overview of the scope of applications is given. The focus of this paper though lies on generalizations of the PCA method.

Missing data is a topic which is often only barely discussed in introductive lectures although the problem is ubiquitous in praxis. Many popular algorithms though can not be extended to the incomplete data case in a trivial manner. In a first extension we look at algorithms capable of handling the incomplete data case, meaning a technique to recover a low-dimensional subspace from missing data. Specifically we explain the ”NIPALS” algorithm (one of the earliest methods to tackle the problem of missing data) which uses an alternating weighted least squares method, where two weighted simple linear regressions are alternated to obtain the first principal component and principal axes of an incomplete dataset. Applying the same method to the residual matrix then results in the following dimensions.

As this pioneering method comes with some drawbacks a modern approach, known as iterative PCA algorithm, which estimates parameters and missing values simultaneously is introduced. Iterative PCA is based on iterative imputation of the missing elements during the estimation of the parameters, and can be seen as a particular EM algorithm. We use toy datasets to challenge the two algorithms on varying degrees of the percentage of missing data. A comparison of the results will be given and a generalized version of the iterative PCA is elaborated which is intended to tackle the often observed overfitting problem.

We note that the results of the above mentioned algorithms can only be trusted for data completely missing at random (CMAR) or missing at random (MAR). An explanation of the different natures of missing data is not covered in this paper and we refer the interested reader to the respective literature on that topic.

The last part of this paper focuses on nonlinear extensions of the PCA algorithm. In practical implementations a linear or affine subspace is often not able to capture vital nonlinear structures in the dataset. To avoid this problem the idea of nonlinear principal component analysis (NLPCA) in a nutshell can be described as follows. If for a given dataset the embedded data structure is not captured by a linear or affine subspace, there may exist a nonlinear mapping of the data into a higher dimensional space such that the embedding of the data lies approximately in a linear or affine subspace of that higher dimensional space. So instead of performing PCA directly on a given dataset the individual data points are first mapped via a nonlinear transformation into a higher dimension, the so called feature space, and the PCA is then applied on the embedded data points (features). A practical drawback of this method though is, that the dimension of the feature space quite regularly becomes very high and therefore the required computations become costly if not prohibitive.

This is why in practice instead of executing the NLPCA algorithm in the feature space the computations are reduced to calculations in the input space of the data points. It will be shown that this is actually possible by introducing what in literature is often referred to as the kernel trick. We will work out that using kernel functions, which are serving as a shortcut computation, enable us to perform nonlinear PCA even for mappings into a infinite feature space. In literature this technique is often described as kernel PCA (KPCA). This essay gives insights to the most often used kernel functions and explains in detail the theory of KPCA. An application is given using toy datasets and the results are summarized and visualized.