Ukrainian Catholic University

Applied Sciences Faculty

Data Science Master Programme

Kernel Principal Component Analysis and its Applications

Project report

Authors:
Irynei Baran
Hanna Pylieva

February 16, 2018

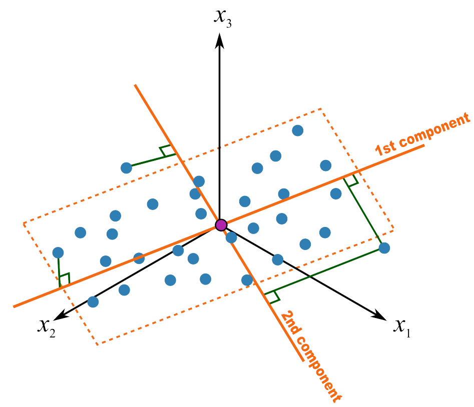

Figure 1: Principal components

Abstract

Principal component analysis (PCA) is a popular tool for linear dimensionality reduction and feature extraction. Kernel PCA is the nonlinear form of PCA, which better exploits the complicated spatial structure of high-dimensional features. In this project, we first review the basic ideas of PCA and kernel PCA. Then we show some experimental results to compare the performance of kernel PCA and standard PCA for classification problems. We also provide an overview of PCA and kPCA applications.

revise Abstract after finishing

1 Introduction

In this section, we briefly review the principal component analysis method, it's applications and limitations.

rewrite the introduction

2 Principal Component Analysis

Principal component analysis, or PCA, is a mathematical procedure which is widely used for dimensionality reduction and feature selection. Those applications are achieved by projecting the data orthogonally onto a space with lower dimension, known as the principal subspace or feature space, such that the variance of projected data is maximal (Bishop , 2006).

Consider a data set X containing N observations of D features ($D < N$). In order to visualize data or diminish number of features for modeling we want to reduce dimensionality of feature space to $M < D$. Whereas we are interested in most influential features to minimize data losses, that is why our goal is to maximize variance of projected data. The directions on which the data is projected called principal components. They are orthogonal and form coordinate system of subspace M (see on Figure 1).

2.1 Finding principal components

To begin with, consider the projection on M when $dim(M) = 1$. Let a unit vector $u_1 \in D$ be the direction of M. Then projection of an observation $x_n \in X$ onto M is $u_1^T x_n$ and the variance of projected data is

$$\frac{1}{N}\sum_{n=1}^{N}\{u_1^T x_n - u_1^T \bar{x}\} = u_1^T S u_1 \tag{1}$$

do we need to provide formula of covariance matrix explicitly? I don't feel so..

where $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_i$ is the mean of sample set and S is the covariance matrix of data set X.

Maximization of (1) is kept in a unit circle as we chose u_1 s.t. $||u_1|| = u_1^T u_1 = 1$. So we need to find maximum of the next Lagrange function:

$$L(X, \lambda_1) = u_1^T S u_1 + \lambda_1(1 - u_1^T u_1) \tag{2}$$

By setting the derivative with respect to u_1 equal to zero we find that in stationary point u_1 needs to be an eigenvector of S:

$$Su_1 = \lambda_1 u_1 \tag{3}$$

Now when we left-multiply by u_1^T and make use of $u_1^T u_1 = 1$ we find out that the variance is given by

$$u_1^T S u_1 = \lambda_1 \tag{4}$$

and so the variance will be a maximum when we set u_1 equal to the eigenvector having the largest eigenvalue λ_1. This eigenvector id called the first principal component (Bishop , 2006).

Next principal components can be found following the same procedure and choosing each new direction such that it maximizes the projected variance amongst all possible directions orthogonal to those already considered.

Do we need better explain the further procedure of it's clear enough?

2.2 Limitations of standard PCA

Although PCA is very useful from practical prospective it has some limitations.

1. Assumptions of linear dependency. PCA projects data orthogonally to reduce dimensionality. This works only if data has linear dependent variables: $y = kx + \epsilon$. Otherwise the method doesn't identify the direction where projected data has highest possible variance which can be seen on Figure 2 .

2. Spread maximization. PCA searches for a subspace where projected data has the maximal spread. However, this is not always the best way to represent data in lower-dimension subspace. A common example is the task of separating and counting pancakes from an image (see Figure 3). Important information for the task (number of pancakes) is located along Z axis which has the lowest variance. So Z axis will be the last principal component identified by PCA which is not what required.

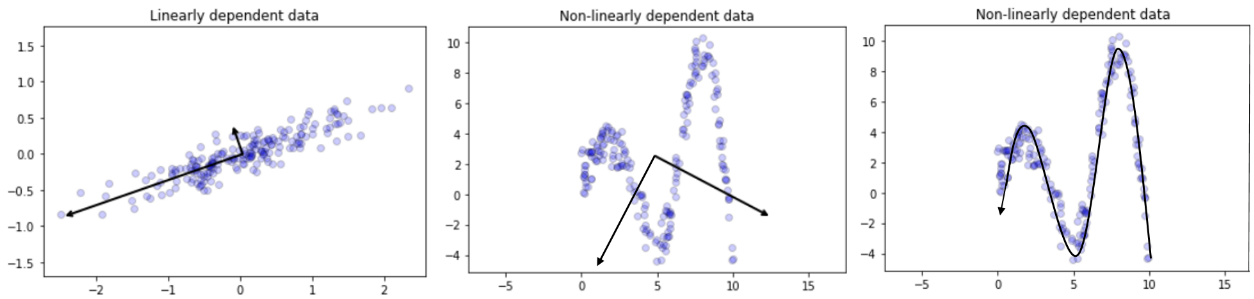

Figure 2: Demonstration of PCA's low performance for non-linearly dependent data

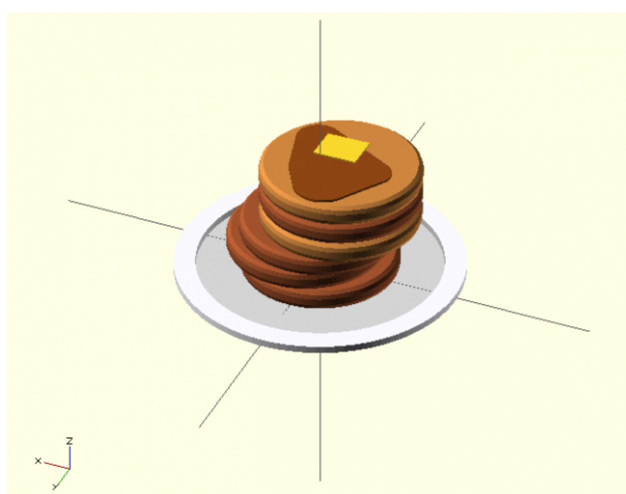

Figure 3: Image from 3D Parametric Pancakes

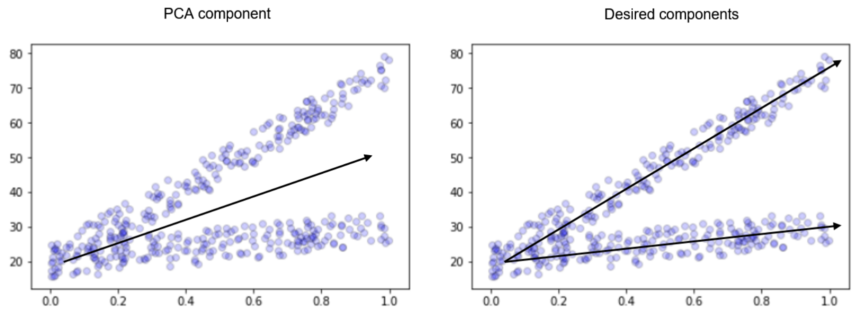

Figure 4: Demonstration of PCA's low performance for non-linearly dependent data

3. Principal components are hard to interpret. PCA reveals implicit dependencies in data which is non-trivial to interpret in case of high-dimensional space reduction. The leverage for this issue is deep understanding of domain.

4. Requirement of orthogonality. PCA comes up with orthogonal principal components, which do not overlap in the space. For some tasks such functionality will produce wrong results (see Figure 4).

2.3 Modifications of PCA

References

Bishop, Christopher M. *Pattern Recognition and Machine Learning.* Springer, Cambridge, U.K., 2006.