# Mathematical Methods in Machine Learning

Wojciech Czaja

UMD, Spring 2016

# Outline

# Introduction

Let's start by paraphrasing some of the principles we laid out in the first lecture:

- There **has always been** an abundance of data available.
- This data **was always** too large, too high-dimensional, too noisy, and too complex.
- Typical problems associated with such data **were always** to cluster, classify, or segment it; and to detect anomalies or embedded targets.

# Plan

Our plan for today is as follows:

- We shall start by describing a class of data-dependent techniques that will be of interest to us. These are the so called Kernel Eigenmap Methods.
- We will briefly mention some possible generalizations of these methods, both from algorithmic and mathematical perspective.
- And then we will discuss the oldest example that fits into our category of interest: the PCA.

# Data Organization and Manifold Learning

- There are many techniques for Data Organization and Manifold Learning, e.g., Principal Component Analysis (PCA), Locally Linear Embedding (LLE), Isomap, genetic algorithms, and neural networks.
- We are interested in a subfamily of these techniques known as *Kernel Eigenmap Methods*. These include Kernel PCA, LLE, Hessian LLE (HLLE), and Laplacian Eigenmaps, Diffusion Maps, Diffusion Wavelets (Output Normalized Methods)).
- Kernel eigenmap methods require two steps. Given data space $X$ of $N$ vectors in $\mathbb{R}^D$.
  1. Construction of an $N \times N$ symmetric, positive semi-definite kernel, $K$, from these $N$ data points in $\mathbb{R}^D$.
  2. Diagonalization of $K$, and then choosing $d \leq D$ *significant* eigenmaps of $K$. These become our new coordinates, and accomplish dimensionality reduction.

Norbert Wiener Center for Harmonic Analysis and Applications

# Kernel Eigenmap Methods for Dimension Reduction - Kernel Construction

- Kernel eigenmap methods were introduced to address complexities not resolvable by linear methods.
- The idea behind *kernel methods* is to express correlations or similarities between vectors in the data space $X$ in terms of a symmetric, positive semi-definite kernel function $K : X \times X \to \mathbb{R}$. Generally, there exists a Hilbert space $\mathbb{K}$ and a mapping $\Phi : X \to \mathbb{K}$ such that

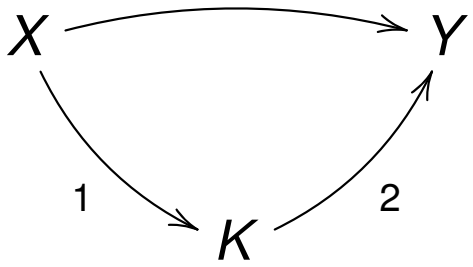
$$K(x, y) = \langle \Phi(x), \Phi(y) \rangle.$$

  Then, diagonalize by the spectral theorem and choose significant eigenmaps to obtain dimensionality reduction.
- Kernels can be constructed by many kernel eigenmap methods. These include Kernel PCA, LLE, HLLE, and Laplacian Eigenmaps.

# Kernel Eigenmap Methods for Dimension Reduction - Kernel Diagonalization

- The second step in kernel eigenmap methods is the diagonalization of the kernel.
- Let $e_j$, $j = 1, \ldots, N$, be the set of eigenvectors of the kernel matrix $K$, with eigenvalues $\lambda_j$.
- Order the eigenvalues monotonically.
- Choose the top $d << D$ significant eigenvectors to map the original data points $x_i \in \mathbb{R}^D$ to $(e_1(i), \ldots, e_d(i)) \in \mathbb{R}^d$, $i = 1, \ldots, N$.

# Data Organization




There are other alternative interpretations for the steps of our diagram:

1. Constructions of kernels K may be independent from data and based on principles.
2. Redundant representations, such as frames, can be used to replace orthonormal eigendecompositions.

We need not select the target dimensionality to be lower than the dimension of the input. This leads, to data expansion, or data organization, rather then dimensionality reduction.

# Operator Theory on Graphs

- Presented approach leads to analysis of operators on data-dependent structures, such as graphs or manifolds.
- Locally Linear Embedding, Diffusion Maps, Diffusion Wavelets, Laplacian Eigenmaps, Schroedinger Eigenmaps.
- Mathematical core:
  - Pick a positive symmetric operator $A$ as the infinitesimal generator of a semigroup of operators, $e^{tA}$, $t > 0$.
  - The semigroup can be identified with the Markov processes of diffusion or random walks, as is the case, e.g., with Diffusion Maps and Diffusion Wavelets of R. R. Coifman and M. Maggioni.
  - The infinitesimal generator and the semigroup share the common representation, e.g., eigenbasis.

# First step: PCA - Principal Component Analysis

- K. Pearson, *On lines and planes of closest fit to systems of points in space,* Philosophical Magazine, vol. 2 (1901), pp. 559–572
- H. Hoteling, *Analysis of a complex of statistical variables into principal components,* Journal of Education Psychology, vol. 24 (1933), pp. 417–44
- K. Karhunen, *Zur Spektraltheorie stochastischer Prozesse,* Ann. Acad. Sci. Fennicae, vol. 34 (1946)
- M. Loève, *Fonctions aléatoire du second ordre,* in Processus stochastiques et mouvement Brownien, p. 299, Paris (1948)

# PCA from Data perspective

We present a data-inspired model for PCA.

- Assume we have $D$ observed (measured) variables: $y = [y_1, \ldots, y_D]^T$. This is our data.
- Assume we know that our data is obtained by a linear transformation $W$ from $d$ unknown variables $x = [x_1, \ldots, x_d]^T$:

$$y = W(x).$$

  Typically we assume $d < D$.
- Assume moreover that the $D \times d$ matrix $W$ is a change of a coordinate system, i.e., columns of $W$ (or towns of $W^T$) are orthonormal to each other:

$$W^T W = Id_d.$$

  Note that $WW^T$ need not be an identity.

# PCA

Given the above assumptions the problem of PCA can be stated as follows:

*How can we find the transformation $W$*
*and the dimension $d$ from a finite number of measurements $y$?*

We shall need 2 additional assumptions:

- Assume that the unknown variables are Gaussian;
- Assume that both the unknown variables and the observations have mean zero (this is easily guaranteed by subtracting the mean, or the sample mean).

# Criteria for leading to PCA

We can represent the principle behind PCA ver generally, as an optimization criterion. In this regard with have:

- $A : \mathbb{R}^D \to \mathbb{R}^d$, s.t. $y \mapsto x = A(y)$;
- $S : \mathbb{R}^d \to \mathbb{R}^D$, s.t., $x \mapsto y = S(x)$.

And we want to minimize the function:

$$E_y \left( \| y - S(A(y)) \| \right).$$

We already have matrix $W$, which in view of the above, will be the synthesis/decoding mapping, $S = W$.
For the analysis/coding map, Pearson proposed to use $A = W^+$, the pseudoinverse of $W$.

# PCA from minimizing the reconstruction error

For a noninvertible matrix, we have its pseudoinverse defined as

$$W^+ = (W^T W)^{-1} W^T$$

In our case, $W^+ = W^T$, Thus, if $y = Wx$, we have

$$WW^T y = WW^T Wx = WId_d x = y,$$

or, equivalently,

$$y - WW^T y = 0.$$

With the presence of noise, we cannot assume anymore the perfect reconstruction, hence, we shall minimize the reconstruction error defined as

$$E_y(\|y - WW^T y\|_2^2).$$

It is not difficult to see that

$$E_y(\|y - WW^T y\|_2^2) = E_y(y^T y) - E_y(y^T WW^T y).$$

# Optimization

Indeed,

$$\begin{aligned} E_y(\|y - WW^T y\|_2^2) &= E_y((y - WW^T y)^T (y - WW^T y)) \\ &= E_y(y^T y - 2y^T WW^T y + y^T WW^T WW^T y) \\ &= E_y(y^T y - 2y^T WW^T y + y^T WW^T y) \\ &= E_y(y^T y - y^T WW^T y) \\ &= E_y(y^T y) - E_y(y^T WW^T y). \end{aligned}$$

# PCA from minimizing the reconstruction error

As $E_y(y^T y)$ is constant, our minimization of error reconstruction turns into a maximization of $E_y(y^T WW^T y)$. In reality, we know little about $y$, so we have to rely on the measurements $y(k)$, $k = 1, \ldots, N$. Then,

$$E_y(y^T WW^T y) \sim \frac{1}{N}\sum_{n=1}^{N}(y(n))^T WW^T(y(n)) \sim \frac{1}{N} tr(Y^T WW^T Y),$$

where $Y$ is the matrix whose columns are the measurements $y(n)$ (hence $Y$ is a $D \times N$ matrix).
Using SVD for $Y$: $Y = V\Sigma U^T$, we obtain:

$$E_y(y^T WW^T y) \sim \frac{1}{N} tr(U\Sigma^T V^T WW^T V\Sigma U^T).$$

Therefore, after some computations we obtain:

$$argmax_W E_y(y^T WW^T y) \sim V\, Id_{D\times d},$$

and so $x \sim Id_{d\times D} V^T y$.

# Conclusions

- We have presented today the first data dependent representation: the Principal Component Analysis.
- We have introduced PCA from the perspective of reconstruction error minimization. This is however not the only possible approach.
- Next time, we shall look at the PCA from the perspective of maximizing variant and decorrelation. Our immediate goal to establish a set of design principles which can be generalized to include other types of transformations.