MLCC 2015
Dimensionality Reduction and PCA

Lorenzo Rosasco
UNIGE-MIT-IIT

June 25, 2015

Outline

Dimensionality Reduction

In many practical applications it is of interest to reduce the dimensionality of the data:

- data visualization
- data *exploration*: for investigating the "effective" dimensionality of the data

Dimensionality Reduction (cont.)

This problem of dimensionality reduction can be seen as the problem of defining a map

$$M : X = \mathbb{R}^D \to \mathbb{R}^k, \quad k \ll D,$$

according to some *suitable criterion*.

Dimensionality Reduction (cont.)

This problem of dimensionality reduction can be seen as the problem of defining a map

$$M : X = \mathbb{R}^D \to \mathbb{R}^k, \quad k \ll D,$$

according to some *suitable criterion*.

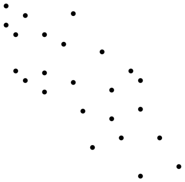
In the following data **reconstruction** will be our guiding principle.

Principal Component Analysis

PCA is arguably the most popular dimensionality reduction procedure.

Principal Component Analysis

PCA is arguably the most popular dimensionality reduction procedure.



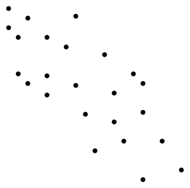
It is a data driven procedure that given an **unsupervised** sample

$$S = (x_1, \ldots, x_n)$$

derive a dimensionality reduction defined by a linear map M.

Principal Component Analysis

PCA is arguably the most popular dimensionality reduction procedure.



It is a data driven procedure that given an **unsupervised** sample

$$S = (x_1, \dots, x_n)$$

derive a dimensionality reduction defined by a linear map M.

PCA can be derived from several prospective and here we give a **geometric** derivation.

Dimensionality Reduction by Reconstruction

Recall that, if

$$w \in \mathbb{R}^D, \quad \|w\| = 1,$$

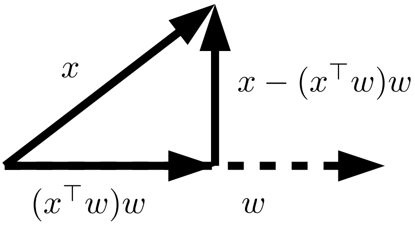
then $(w^T x)w$ is the **orthogonal projection** of x on w

Dimensionality Reduction by Reconstruction

Recall that, if

$$w \in \mathbb{R}^D, \quad \|w\| = 1,$$

then $(w^T x)w$ is the **orthogonal projection** of x on w



Dimensionality Reduction by Reconstruction (cont.)

First, consider $k = 1$. The associated **reconstruction error** is

$$\|x - (w^T x)w\|^2$$

(that is how much we lose by projecting x along the direction w)

Dimensionality Reduction by Reconstruction (cont.)

First, consider $k = 1$. The associated **reconstruction error** is

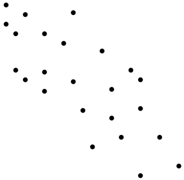
$$\|x - (w^T x)w\|^2$$

(that is how much we lose by projecting x along the direction w)

Problem:

Find the direction p allowing the best reconstruction of the training set

Dimensionality Reduction by Reconstruction (cont.)



Let $\mathbb{S}^{D-1} = \{w \in \mathbb{R}^D \mid \|w\| = 1\}$ is the sphere in D dimensions. Consider the **empirical reconstruction** minimization problem,

$$\min_{w \in \mathbb{S}^{D-1}} \frac{1}{n} \sum_{i=1}^{n} \|x_i - (w^T x_i) w\|^2.$$

The solution p to the above problem is called the **first principal component** of the data

An Equivalent Formulation

A direct computation shows that $\|x_i - (w^T x_i)w\|^2 = \|x_i\| - (w^T x_i)^2$

An Equivalent Formulation

A direct computation shows that $\|x_i - (w^T x_i)w\|^2 = \|x_i\| - (w^T x_i)^2$

Then, problem

$$\min_{w \in \mathbb{S}^{D-1}} \frac{1}{n} \sum_{i=1}^{n} \|x_i - (w^T x_i)w\|^2$$

is equivalent to

$$\max_{w \in \mathbb{S}^{D-1}} \frac{1}{n} \sum_{i=1}^{n} (w^T x_i)^2$$

Outline

Reconstruction and Variance

Assume the data to be centered, $\bar{x} = \frac{1}{n} x_i = 0$, then we can interpret the term

$$(w^T x)^2$$

as the **variance** of x in the direction w.

Reconstruction and Variance

Assume the data to be centered, $\bar{x} = \frac{1}{n} x_i = 0$, then we can interpret the term

$$(w^T x)^2$$

as the **variance** of x in the direction w.

The first PC can be seen as the direction along which the data have maximum variance.

$$\max_{w \in \mathbb{S}^{D-1}} \frac{1}{n} \sum_{i=1}^{n} (w^T x_i)^2$$

Centering

If the data are not centered, we should consider

$$\max_{w \in \mathbb{S}^{D-1}} \frac{1}{n} \sum_{i=1}^{n} (w^T (x_i - \bar{x}))^2 \tag{1}$$

equivalent to

$$\max_{w \in \mathbb{S}^{D-1}} \frac{1}{n} \sum_{i=1}^{n} (w^T x_i^c)^2$$

with $x^c = x - \bar{x}$.

Centering and Reconstruction

If we consider the effect of centering to reconstruction it is easy to see that we get

$$\min_{w,b \in \mathbb{S}^{D-1}} \frac{1}{n} \sum_{i=1}^{n} \|x_i - ((w^T(x_i - b))w + b)\|^2$$

where

$$((w^T(x_i - b))w + b$$

is an affine (rather than an orthogonal) projection

Outline

PCA as an Eigenproblem

A further manipulation shows that PCA corresponds to an eigenvalue problem.

PCA as an Eigenproblem

A further manipulation shows that PCA corresponds to an eigenvalue problem.

Using the symmetry of the inner product,

$$\frac{1}{n}\sum_{i=1}^{n}(w^T x_i)^2 = \frac{1}{n}\sum_{i=1}^{n} w^T x_i w^T x_i = \frac{1}{n}\sum_{i=1}^{n} w^T x_i x_i^T w = w^T(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T)w$$

PCA as an Eigenproblem

A further manipulation shows that PCA corresponds to an eigenvalue problem.

Using the symmetry of the inner product,

$$\frac{1}{n}\sum_{i=1}^{n}(w^T x_i)^2 = \frac{1}{n}\sum_{i=1}^{n} w^T x_i w^T x_i = \frac{1}{n}\sum_{i=1}^{n} w^T x_i x_i^T w = w^T(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T)w$$

Then, we can consider the problem

$$\max_{w \in \mathbb{S}^{D-1}} w^T C_n w, \quad C_n = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$$

PCA as an Eigenproblem (cont.)

We make two observations:

- The ("covariance") matrix $C_n = \frac{1}{n}\sum_{i=1}^{n} X_n^T X_n$ is symmetric and positive semi-definite.

PCA as an Eigenproblem (cont.)

We make two observations:

- The ("covariance") matrix $C_n = \frac{1}{n}\sum_{i=1}^{n} X_n^T X_n$ is symmetric and positive semi-definite.

- The objective function of PCA can be written as

$$\frac{w^T C_n w}{w^T w}$$

  the so called *Rayleigh* quotient.

PCA as an Eigenproblem (cont.)

We make two observations:

- The ("covariance") matrix $C_n = \frac{1}{n}\sum_{i=1}^{n} X_n^T X_n$ is symmetric and positive semi-definite.

- The objective function of PCA can be written as

$$\frac{w^T C_n w}{w^T w}$$

  the so called *Rayleigh* quotient.

Note that, if $C_n u = \lambda u$ then $\frac{u^T C_n u}{u^T u} = \lambda$, since u is normalized.

PCA as an Eigenproblem (cont.)

We make two observations:

- The ("covariance") matrix $C_n = \frac{1}{n}\sum_{i=1}^{n} X_n^T X_n$ is symmetric and positive semi-definite.

- The objective function of PCA can be written as

$$\frac{w^T C_n w}{w^T w}$$

  the so called *Rayleigh* quotient.

Note that, if $C_n u = \lambda u$ then $\frac{u^T C_n u}{u^T u} = \lambda$, since u is normalized.

Indeed, it is possible to show that the Rayleigh quotient achieves its maximum at a vector corresponding to the maximum eigenvalue of C_n

PCA as an Eigenproblem (cont.)

Computing the first principal component of the data reduces to computing the biggest eigenvalue of the covariance and the corresponding eigenvector.

$$C_n u = \lambda u, \quad C_n = \frac{1}{n} \sum_{i=1}^{n} X_n^T X_n$$

Outline

Beyond the First Principal Component

We discuss how to consider more than one principle component $(k > 1)$

$$M : X = \mathbb{R}^D \to \mathbb{R}^k, \quad k \ll D$$

The idea is simply to iterate the previous reasoning

Residual Reconstruction

The idea is to consider the one dimensional projection that can best reconstruct the residuals

$$r_i = x_i - (p^T x_i)p_i$$

Residual Reconstruction

The idea is to consider the one dimensional projection that can best reconstruct the residuals

$$r_i = x_i - (p^T x_i) p_i$$

An associated minimization problem is given by

$$\min_{w \in \mathbb{S}^{D-1}, w \perp p} \frac{1}{n} \sum_{i=1}^{n} \| r_i - (w^T r_i) w \|^2.$$

(note: the constraint $w \perp p$)

Residual Reconstruction (cont.)

Note that for all $i = 1, \ldots, n$,

$$\|r_i - (w^T r_i)w\|^2 = \|r_i\|^2 - (w^T r_i)^2 = \|r_i\|^2 - (w^T x_i)^2$$

since $w \perp p$

Residual Reconstruction (cont.)

Note that for all $i = 1, \ldots, n$,

$$\|r_i - (w^T r_i)w\|^2 = \|r_i\|^2 - (w^T r_i)^2 = \|r_i\|^2 - (w^T x_i)^2$$

since $w \perp p$

Then, we can consider the following equivalent problem

$$\max_{w \in \mathbb{S}^{D-1}, w \perp p} \frac{1}{n} \sum_{i=1}^{n} (w^T x_i)^2 = w^T C_n w.$$

PCA as an Eigenproblem

$$\max_{w \in \mathbb{S}^{D-1}, w \perp p} \frac{1}{n} \sum_{i=1}^{n} (w^T x_i)^2 = w^T C_n w.$$

Again, we have to minimize the Rayleigh quotient of the covariance matrix with the extra constraint $w \perp p$

PCA as an Eigenproblem

$$\max_{w \in \mathbb{S}^{D-1}, w \perp p} \frac{1}{n} \sum_{i=1}^{n} (w^T x_i)^2 = w^T C_n w.$$

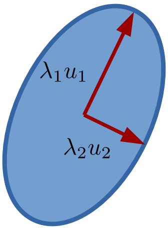
Again, we have to minimize the Rayleigh quotient of the covariance matrix with the extra constraint $w \perp p$

Similarly to before, it can be proved that the solution of the above problem is given by the second eigenvector of C_n, and the corresponding eigenvalue.

PCA as an Eigenproblem (cont.)

$$C_n u = \lambda u, \quad C_n = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$$

The reasoning generalizes to more than two components:
computation of k principal components reduces to finding k eigenvalues and eigenvectors of C_n.



Remarks

- Computational complexity roughly $O(kD^2)$ (complexity of forming C_n is $O(nD^2)$). If we have n points in D dimensions and $n \ll D$ can we compute PCA in less than $O(nD^2)$?

Remarks

- Computational complexity roughly $O(kD^2)$ (complexity of forming C_n is $O(nD^2)$). If we have n points in D dimensions and $n \ll D$ can we compute PCA in less than $O(nD^2)$?

- The dimensionality reduction induced by PCA is a linear projection. Can we generalize PCA to non linear dimensionality reduction?

Outline

Singular Value Decomposition

Consider the data matrix X_n, its singular value decomposition is given by

$$X_n = U\Sigma V^T$$

where:

- U is a n by k orthogonal matrix,
- V is a D by k orthogonal matrix,
- Σ is a diagonal matrix such that $\Sigma_{i,i} = \sqrt{\lambda_i}$, $i = 1, \ldots, k$ and $k \leq \min\{n, D\}$.

The columns of U and the columns of V are the left and right singular vectors and the diagonal entries of Σ the singular values.

Singular Value Decomposition (cont.)

The SVD can be equivalently described by the equations

$$C_n p_j = \lambda_j p_j, \quad \frac{1}{n} K_n u_j = \lambda_j u_j,$$
$$X_n p_j = \sqrt{\lambda_j} u_j, \quad \frac{1}{n} X_n^T u_j = \sqrt{\lambda_j} p_j,$$

for $j = 1, \dots, d$ and where $C_n = \frac{1}{n} X_n^T X_n$ and $\frac{1}{n} K_n = \frac{1}{n} X_n X_n^T$

PCA and Singular Value Decomposition

If $n \ll p$ we can consider the following procedure:

- form the matrix K_n, which is $O(Dn^2)$
- find the first k eigenvectors of K_n, which is $O(kn^2)$
- compute the principal components using

$$p_j = \frac{1}{\sqrt{\lambda_j}} X_n^T u_j = \frac{1}{\sqrt{\lambda_j}} \sum_{i=1}^{n} x_i u_j^i, \quad j = 1, \ldots, d$$

  where $u = (u^1, \ldots, u^n)$, This is $O(knD)$ if we consider k principal components.

Outline

Beyond Linear Dimensionality Reduction?

By considering PCA we are implicitly assuming the data to lie on a linear subspace....

Beyond Linear Dimensionality Reduction?

By considering PCA we are implicitly assuming the data to lie on a linear subspace....

...it is easy to think of situations where this assumption might violated.

Beyond Linear Dimensionality Reduction?

By considering PCA we are implicitly assuming the data to lie on a linear subspace....

...it is easy to think of situations where this assumption might violated.

Can we use kernels to obtain non linear generalization of PCA?

From SVD to KPCA

Using SVD the projection of a point x on a principal component p_j, for $j = 1, \dots, d$, is

$$(M(x))^j = x^T p_j = \frac{1}{\sqrt{\lambda_j}} x^T X_n^T u_j = \frac{1}{\sqrt{\lambda_j}} \sum_{i=1}^n x^T x_i u_j^i,$$

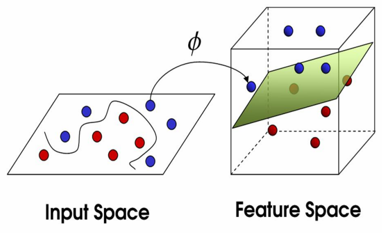
Recall

$$C_n p_j = \lambda_j p_j, \quad \frac{1}{n} K_n u_j = \lambda_j u_j,$$
$$X_n p_j = \sqrt{\lambda_j} u_j, \quad \frac{1}{n} X_n^T u_j = \sqrt{\lambda_j} p_j,$$

PCA and Feature Maps

$$(M(x))^j = \frac{1}{\sqrt{\lambda_j}} \sum_{i=1}^{n} x^T x_i u_i^i,$$

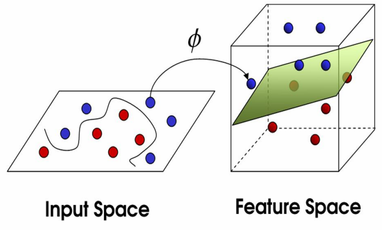
What if consider a non linear feature-map $\Phi : X \to F$, before performing PCA?



PCA and Feature Maps

$$(M(x))^j = \frac{1}{\sqrt{\lambda_j}} \sum_{i=1}^{n} x^T x_i u_i^i,$$

What if consider a non linear feature-map $\Phi : X \to F$, before performing PCA?



$$(M(x))^j = \Phi(x)^T p_j = \frac{1}{\sqrt{\lambda_j}} \sum_{i=1}^{n} \Phi(x)^T \Phi(x_i) u_j^i,$$

where $K_n \sigma_j = \sigma_j u_j$ and $(K_n)_{i,j} = \Phi(x)^T \Phi(x_j)$.

Kernel PCA

$$(M(x))^j = \Phi(x)^T p_j = \frac{1}{\sqrt{\lambda_j}} \sum_{i=1}^{n} \Phi(x)^T \Phi(x_i) u_j^i,$$

If the feature map is defined by a positive definite kernel $K : X \times X \to \mathbb{R}$, then

$$(M(x))^j = \frac{1}{\sqrt{\lambda_j}} \sum_{i=1}^{n} K(x, x_i) u_j^i,$$

where $K_n \sigma_j = \sigma_j u_j$ and $(K_n)_{i,j} = K(x_i, x_j)$.

Wrapping Up

In this class we introduced PCA as a basic tool for dimensionality reduction. We discussed computational aspect and extensions to non linear dimensionality reduction (KPCA)

Next Class

In the next class, beyond dimensionality reduction, we ask how we can devise interpretable data models, and discuss a class of methods based on the concept of sparsity.