SD-TSIA204 : PCA

Ekhine Irurozki
Télécom Paris, IP Paris

Motivation

What is it ?

- Unsupervised learning technique
- We use is as a prepocessing for the OLS (aka PCA before OLS, aka PCRegression, ...)

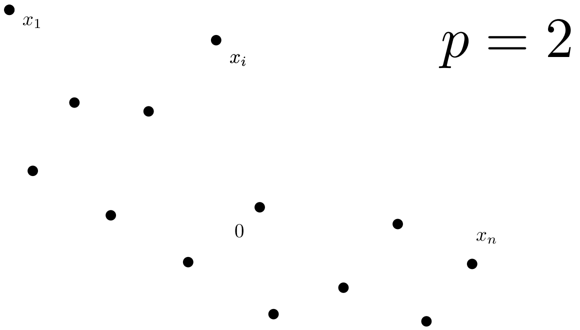
High level idea : find a low dimensional representation of the data X that keeps the variance

- Super-collinearity
- Close to 0 variance features

Graphical representation (not to be confused with OLS)

PCA

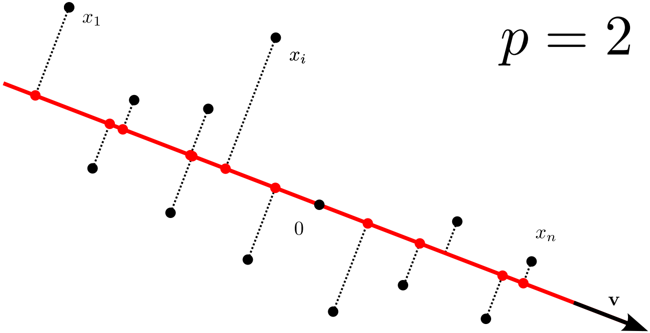
We observe n points $x_1, \ldots, x_n$, i.e., $X = [x_1, \ldots, x_n]^\top \in \mathbb{R}^{n \times p}$, n observations (rows), p *features* (columns)



Rem: we have to center the points so that they have a zero average $X \leftarrow [x_1 - \overline{x}_n, \ldots, x_n - \overline{x}_n]^\top = X - \mathbf{1}_n \overline{x}_n^\top$ (we can also scale to have a similar standard deviation by *feature*)

PCA

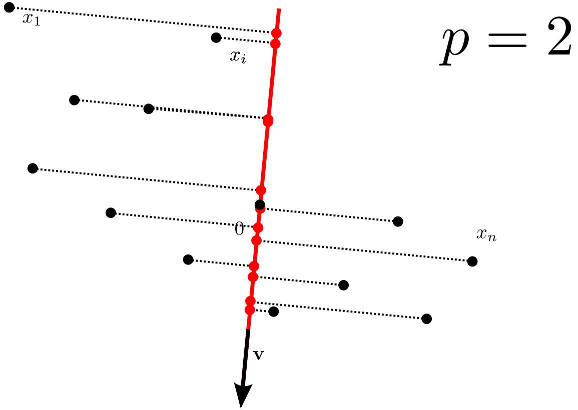
We observe n points $x_1, \ldots, x_n$, i.e., $X = [x_1, \ldots, x_n]^\top \in \mathbb{R}^{n \times p}$, n observations (rows), p *features* (columns)



Rem: we have to center the points so that they have a zero average $X \leftarrow [x_1 - \overline{x}_n, \ldots, x_n - \overline{x}_n]^\top = X - \mathbf{1}_n \overline{x}_n^\top$ (we can also scale to have a similar standard deviation by *feature*)

PCA

We observe n points $x_1, \ldots, x_n$, i.e., $X = [x_1, \ldots, x_n]^\top \in \mathbb{R}^{n \times p}$, n observations (rows), p *features* (columns)



Rem: we have to center the points so that they have a zero average $X \leftarrow [x_1 - \overline{x}_n, \ldots, x_n - \overline{x}_n]^\top = X - \mathbf{1}_n \overline{x}_n^\top$ (we can also scale to have a similar standard deviation by *feature*)

Principal Component Analysis, PCA

Parameter k : number of axes to represent a cloud of n points $(x_1, \ldots, x_n)$, represented by the lines of $X \in \mathbb{R}^{n \times p}$.

This method compresses the point cloud of dimension p into a cloud of dimension k.

The PCA (of level k) consists in performing the SVD of X, and keeping only the k principal axes to represent the cloud.

$$X = \sum_{i=1}^{r} s_i \mathbf{u}_i \mathbf{v}_i^\top \longrightarrow \sum_{i=1}^{k} s_i \mathbf{u}_i \mathbf{v}_i^\top$$

We call **principal axes** the k vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$, and in general $k \ll p$ (*e.g.*,$k = 2$, for a planar display)

Nouvelle représentation des données

- The axes (of direction) $\mathbf{v}_1, \ldots, \mathbf{v}_p \in \mathbb{R}^p$ are called **principal axes** or **factor axes**, the new variables $\mathbf{c}_j = X\mathbf{v}_j, j = 1, \ldots, p$ are called **principal constituents**

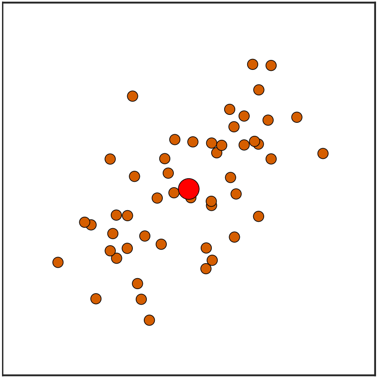
New representation (order k) :

- The matrix XV_k (with $V_k = [\mathbf{v}_1, \ldots, \mathbf{v}_k]$) is the matrix representing the data in the base of the first k eigenvectors

Reconstruction in the original space (debruiter) :

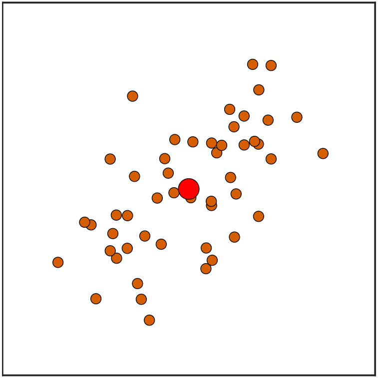
- "Perfect" reconstruction for $\mathbf{x} \in \mathbb{R}^p$: $\mathbf{x} = \sum_{j=1}^{p}(\mathbf{x}^\top \mathbf{v}_j)\mathbf{v}_j$
- Reconstruction with loss of information : $\hat{\mathbf{x}} = \sum_{j=1}^{k}(\mathbf{x}^\top \mathbf{v}_j)\mathbf{v}_j$

Main axis : variance maximization



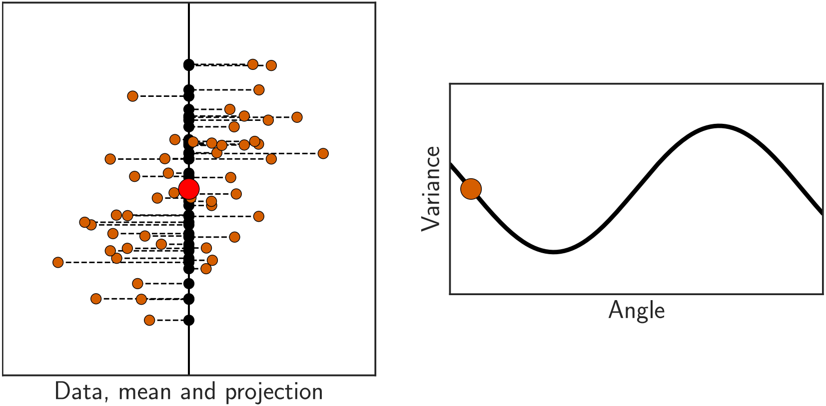
Data and mean

Main axis : variance maximization

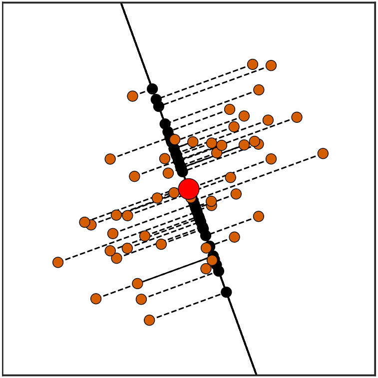


Data and mean

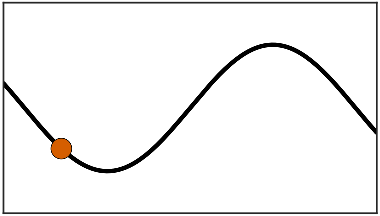
Main axis : variance maximization



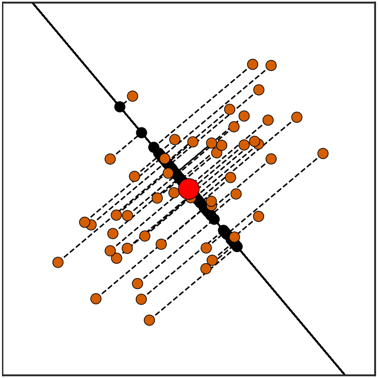
Main axis : variance maximization



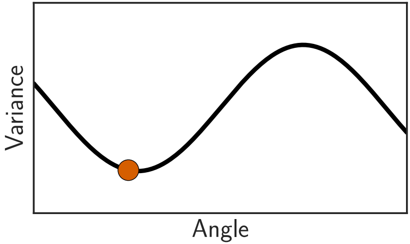
Data, mean and projection



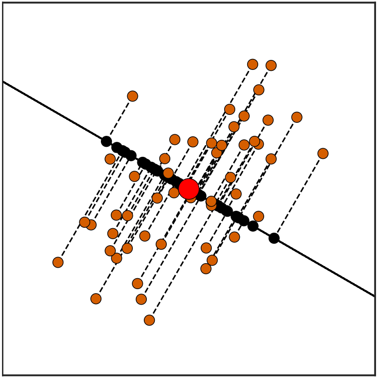
Main axis : variance maximization



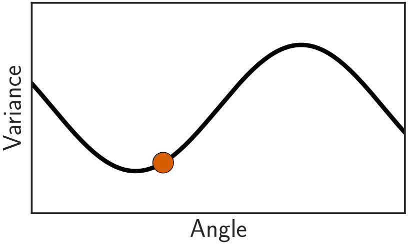
Data, mean and projection



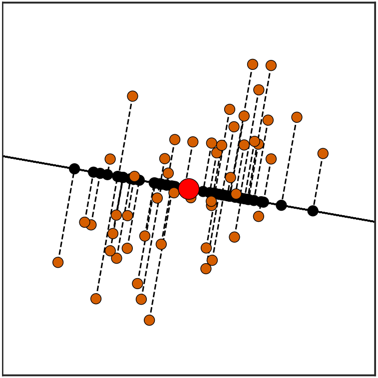
Main axis : variance maximization



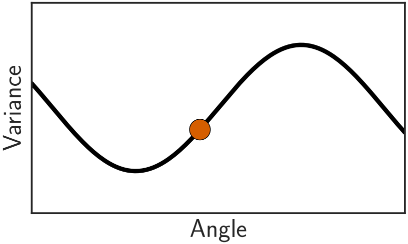
Data, mean and projection



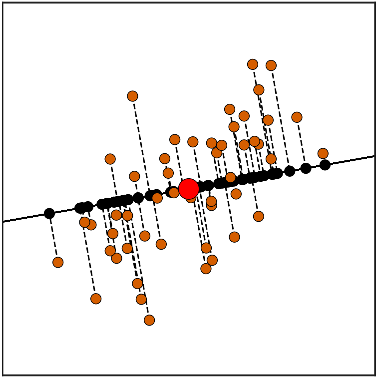
Main axis : variance maximization



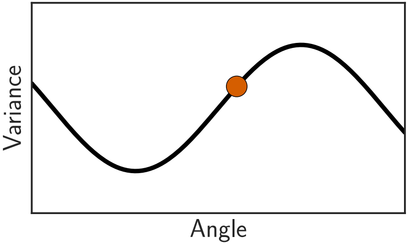
Data, mean and projection



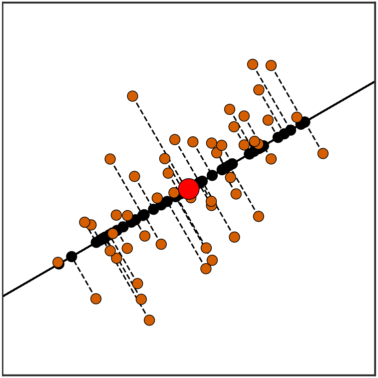
Main axis : variance maximization



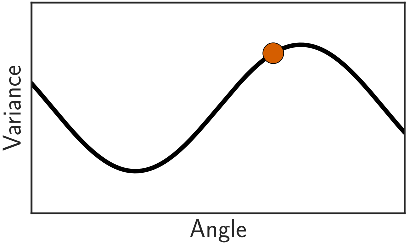
Data, mean and projection



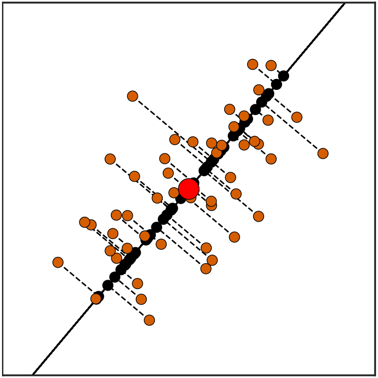
Main axis : variance maximization



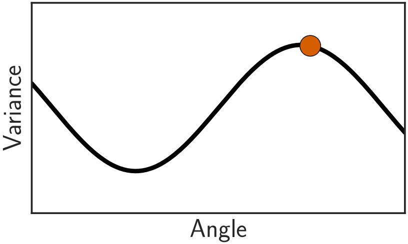
Data, mean and projection



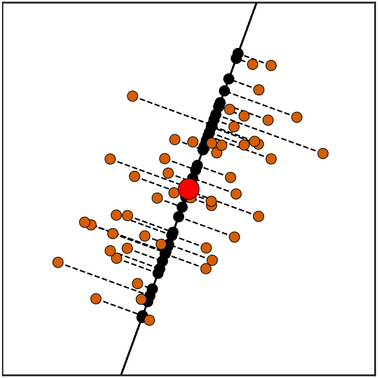
Main axis : variance maximization



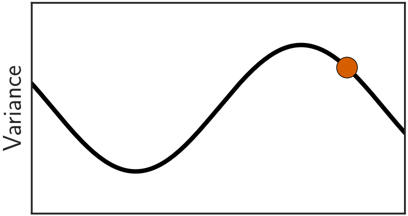
Data, mean and projection



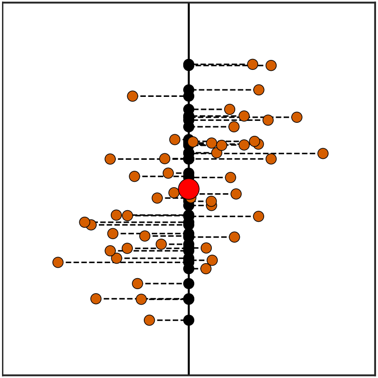
Main axis : variance maximization

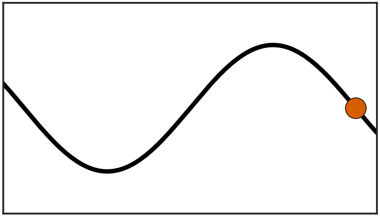


Data, mean and projection

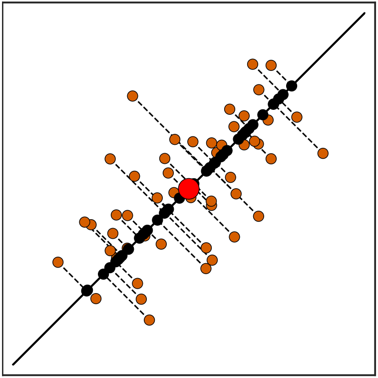


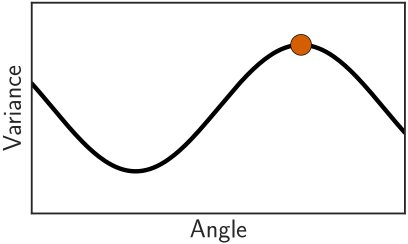
Main axis : variance maximization





Main axis : variance maximization





Problem statement

PCA sketch

- data is centered and standardized
- Direction $v_1 \in \mathbb{R}^p$ is a linear combination of the original dimensions of X
- The distance from the origin to the projection of x_i onto v_1 is $x_i^\top v_1$
- The variance along v_i of the projections is $\sum_{i=1}^n (x_i^\top v_1)^2 = \|Xv_1\|^2 = v_1^\top X^\top X v_1$
- Gram matrix : $G = (n-1)^{-1} X^\top X$, a symmetric covariance matrix
- We rewrite the variance $\sum_{i=1}^n (x_i^\top v_1)^2 = v_1^\top G v_1$
- Optimization problem

$$\underset{v_1 \in \mathbb{R}^p, \|v_1\|=1}{\arg\max} \sum_{i=1}^n (x_i^\top v_1)^2 = \underset{v_1 \in \mathbb{R}^p, \|v_1\|=1}{\arg\max} v_1^\top G v_1$$

Solution in the first direction v_1

By the method of Lagrange multipliers we have that the solution of $\arg\max_{\mathbf{v}\in\mathbb{R}^p,\|\mathbf{v}\|=1} v_1^\top G v_1$

- $G v_1 = \lambda_1 v_1$
- λ_1, v_1 are the eigenvalue/vector
- λ_1 is also the variance

After, find v_2, a direction $\perp v_1$ that maximizes the variance.

Let λ_i, v_i the i-th largest eigenvalue and its associated eigenvector. Then $v_i \perp v_{i-1}$ for $i > 1$ and maximizes the variance

<u>Exercise</u> Show that the i-th singular value of X, σ_i, and the i-th eigenvalue of $X^\top X$, λ_i, are related as follows $\lambda_i = (n-1)^{-1}\sigma_i^2$

PCA before OLS

Algorithme : PCA before OLS

Entrées : $X \in \mathbb{R}^{n \times p}$, itérations K

$\lambda_i, v_i \leftarrow i$-th largest eigenvalue and assoc eigenvector

$Z = XV$ is the new (projected) dataset

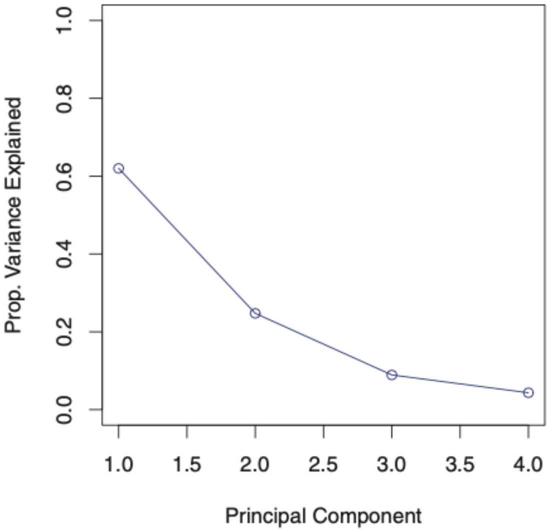
OLS in Z

Understanding the projection/direction, dataset USArrests

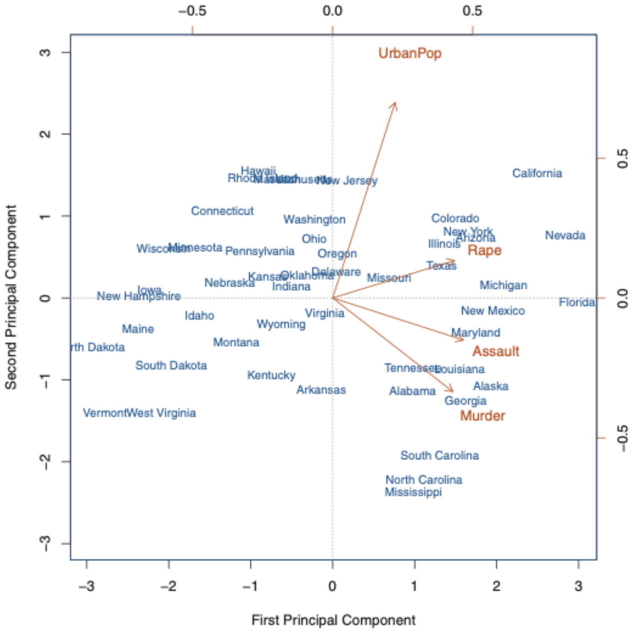
		Murder	Assault	UrbanPop	Rape
0	Alabama	13.2	236	58	21.2
1	Alaska	10.0	263	48	44.5
2	Arizona	8.1	294	80	31.0
3	Arkansas	8.8	190	50	19.5
4	California	9.0	276	91	40.6

...

Percentage of variance explained



Principal components



Conclusions

- PCA is an unsupervised technique
- Dimensionality reduction (more than a feature subset selection method)
- When the target $\mathbf{y}$ is correlated with the variance directions then its useful
- Interpretation of the proportion of variance explained
- Projection to low dimensions
- No interpretability on lower dimensions