Statistical Learning with High-dimensional Data

Pr. Charles BOUVEYRON

Professor of Statistics
Chair of the Institut 3IA Côte d'Azur
Université Côte d'Azur & Inria

charles.bouveyron@univ-cotedazur.fr
@cbouveyron

Outline

Dimension reduction

A common phantasm about dimension reduction:

- believe that dimension reduction helps for classification,
- this is not true because, most of the time, dimension reduction implies an information loss which would be discriminative.

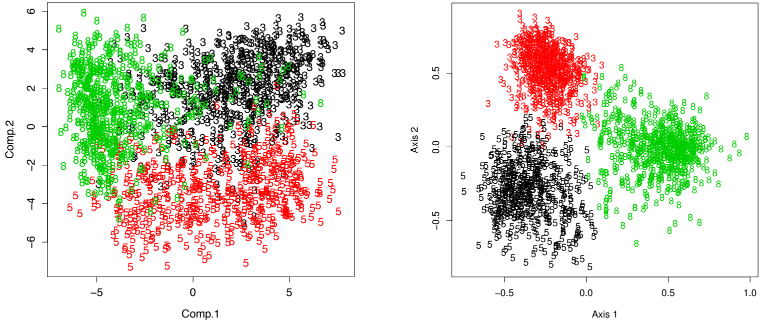
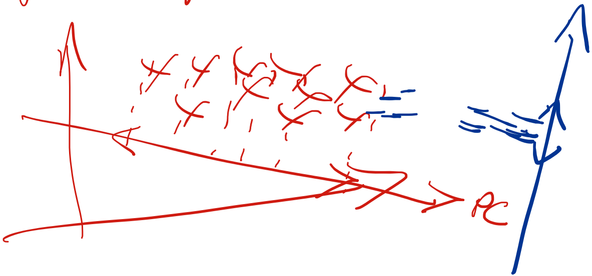


Figure: Projection of the 256-dimensional USPS data with PCA (left, unsupervised) and FDA (right, supervised).

PCA: principal component analysis

The goal of PCA is to create d new variables, which are linear combinations of the original var, such that the variance of the projected data is maximum.



PCA: the principle

$$\max_U \; U^t \underbrace{\bar{X}^t \bar{X}}_{} U$$

co-variance matrix of the centered data

The solution: $\hat{U}$ is a matrix made of the d eigenvectors associated to the largest eigenvalues

$$\bar{X}^t \bar{X} = S = Q^t \Delta Q \quad \text{where } \Delta = \mathrm{diag}(\lambda_1, \dots, \lambda$$

SPD matrix

PCA: the principle

The reciepe :

(i) Compute $S = \bar{X}^t \bar{X}$

(ii) eigendecomposition of S

(iii) keep the vectors q_j such that λ_j, $j = 1 \ldots d$ are the largest eigenvalues.

PCA: projection

$$\underset{n \times d}{Y} = \underset{n \times p}{X} \times \underset{p \times d}{U}$$

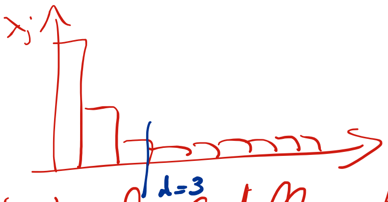
scores.

PCA: how many axes?

(i) the 90% rule: we keep d variables such that

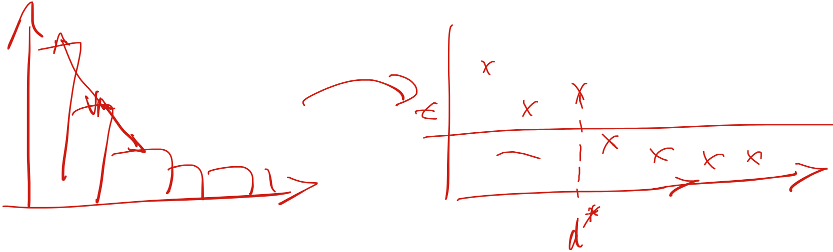
$$\sum_{j=1}^{d} \lambda_j \Big/ \sum_{j=1}^{p} \lambda_j \geq 0.9$$

(ii) the knee test: looking a the eigenvalue scree



(iii) the scree-test of Catell: looking at the differences between eigenvalues, we retain d such that all diff after d are smaller than a threshold

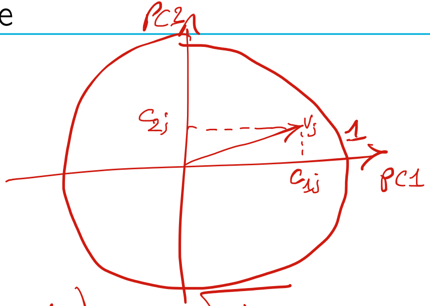
PCA: how many axes?



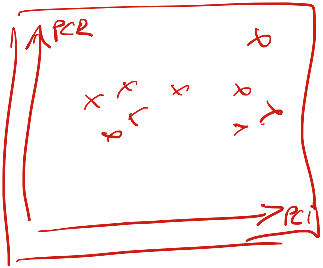
$\epsilon = 0.1 \times \max(\text{diff})$

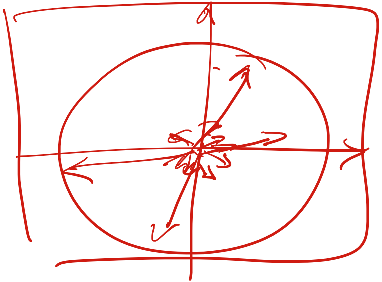
(iv) the other solutions (statistical tests, Bayesian app, ...) are clearly more complex and perform only slightly better.

PCA: correlation circle



$$c_{ij} = \mathrm{Corr}(PC_i, V_j) = \underline{\sqrt{\lambda_i}\, q_{ij}}$$





PCA: analysis of the marathon data

See the R code ...

PPCA: a probabilistic version of PCA

PPCA was proposed by Tipping & Bishop (1996) to better establish the theory of PCA under a Gaussian distribution:

$x \in \mathbb{R}^p$

- $X = U^t Y + \varepsilon$
- $Y \sim \mathcal{N}(\mu, I_d)$

$\varepsilon \sim \mathcal{N}(0, \sigma^2 I_p)$

is called the latent variable $Y \in \mathbb{R}^d$

$$\Longrightarrow \quad p(X|\theta) = \mathcal{N}(U^t\mu, U^tU + \sigma^2 I_p) \quad \text{with } \theta = \{\mu, U, \sigma^2\}$$

PPCA: a probabilistic version of PCA

In this case, doing a (P)PCA is equivalent to estimate the parameters $\theta = \{\mu, \underline{U}, \sigma^2\}$ from the data.

Max Likelihood $\Rightarrow$ $\hat{U}_{ML}$ = the eigenvectors of $\overline{X}^t \overline{X}$ associated to the largest eigen values.

$$\mu = \overline{X}$$

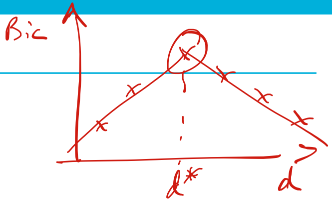
PPCA: why is it interesting?

(i) justification of a very old result

(ii) it allows to use model selection and all statistical tools within PCA

(iii) it allows to propose new models based on this PPCA model

PPCA in practice...



PPCA = PCA in practice

except that we can use AIC or BIC for choosing d, ...

AIC, BIC (...) are model selection criteria which are based on penalized likelihood

$$\mathrm{BIC}(\mathcal{M}) = \log \mathcal{L}(\hat{\theta}) - \frac{\gamma(\mathcal{M})}{2} \log(n)$$

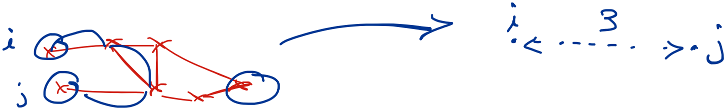
where $\gamma(\mathcal{M})$ is the nb of parameters in $\mathcal{M}$.

MDS: multi-dimensional scaling

The idea of MDS is to find a low-dim representation of the data that keeps the original topology of the data.

$$X \in \left\{ \begin{array}{l} \mathbb{R}^p \\ \text{Graph} \\ \ldots \end{array} \right. \longrightarrow Y \in \mathbb{R}^d$$

where it is possible to compute a distance



MDS: the principle

Given the distances between points in the original space, let say d_{ij} $\forall i,j = 1 \dots n$

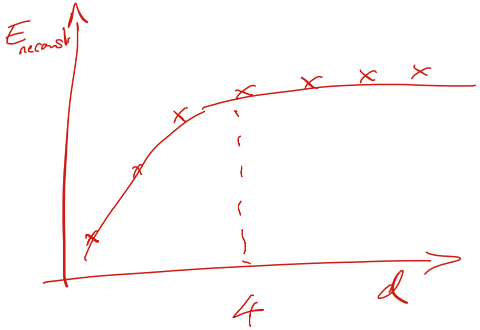
$$\min_{\underline{y}} \sum_{i,j=1}^{n} \| \underline{d_{ij}} - \underline{\delta_{ij}} \|^2$$

where $\delta_{ij} = \| \underline{y}_i - \underline{y}_j \|^2$

MDS in practice...

in R : cmdscale (dist(X) , d)

it remains to choose d



t-SNE: t-distributed stochastic neighbor embedding

t-SNE is a model-based version of MDS

(i) SNE : $\forall j = 1 \dots n, j \neq i$ $\quad \underline{x_j \sim N(x_i, \sigma_i^2)}$

$p_{j|i}$ = probability that i chooses j as neighborhood

$$\underset{\text{Bayes}}{\propto} \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)}$$

Min KL div

$\vdots$

$$q_{j|i} = \frac{\exp\left(-\|y_i - y_j\|^2 / 2\sigma_i^2\right)}{\sum \_\_\_\_\_\_\_\_}$$

t-SNE: from SNE to T-SNE

$$\min_{y} \sum_{i \neq j} p_{j|i} \log \left(\frac{p_{j|i}}{q_{j|i}} \right).$$

⊖ very difficult to optimize

⊖ a very asymmetric modeling.

t-SNE: the principle

(ii) t-SNE :

- symmetry : $p_{ij} = p_{ji} = \frac{p_{j|i} + p_{i|j}}{2n}$

- distribution in the reduced space

$$\forall j \neq i, \quad y_j \sim T(y_i, 1)$$

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq j} \text{———}}$$