Model-based Statistical Learning

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

The curse of dimensionality

The curse of dimensionality:

- this term was first used by R. Bellman in the introduction of his book "Dynamic programming" in 1957:

  All [problems due to high dimension] *may be subsumed under the heading "the curse of dimensionality". Since this is a curse,* [...], *there is no need to feel discouraged about the possibility of obtaining significant results despite it.*

- he used this term to talk about the difficulties to find an optimum in a high-dimensional space using an exhaustive search,
- in order to promotate dynamic approaches in programming.

The curse of dimensionality

$t_{ik} = p(z=k \mid x=x_i)$

In the mixture model context:

- the building of the data partition mainly depends on:

$$\underline{H_k(x)} = -2\log(\underline{\pi_k f(x,\theta_k)}),$$

- model Full-GMM:

$$\underbrace{H_k(x) = (x-\mu_k)^t\Sigma_k^{-1}(x-\mu_k) + \log(\det\Sigma_k) - 2\log(\pi_k) + \gamma.}$$

$p(p+1)/2$

Consequently:

- it is necessary to invert Σ_k which have a number of parameters proportional to $\underline{p^2}$,
- if n is small compared to p^2, the estimates of Σ_k are ill-conditionned or singular and it will be difficult or impossible to invert Σ_k.

$\Rightarrow$ if n is too small compared to p, it won't be possible to run an EM algorithm with a full-GMM!

The curse of dimensionality

From the estimation point of view:

- let us consider the normalized trace $\tau(\Sigma) = \mathrm{tr}(\Sigma^{-1})/p$ of the inverse covariance matrix Σ^{-1} of a multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$,
- the estimation of τ from a sample of n observations $\{x_1, \ldots, x_n\}$ conduces to:

$$\widehat{\tau(\Sigma)} = \tau(\hat{\Sigma}) = \frac{1}{p}\mathrm{tr}(\hat{\Sigma}^{-1}),$$

$$E[\tau(\hat{\Sigma})] = \left(1 - \frac{p}{n-1}\right)^{-1} \tau(\Sigma).$$

- consequently, if the ratio $p/n \to 0$ when $n \to +\infty$, then $E[\tau(\hat{\Sigma})] \to \tau(\Sigma)$,
- however, if the dimension p is comparable with n, then $E[\tau(\hat{\Sigma})] \to c\tau(\Sigma)$ when $n \to +\infty$, where $c = \lim_{n\to+\infty} p/n$.

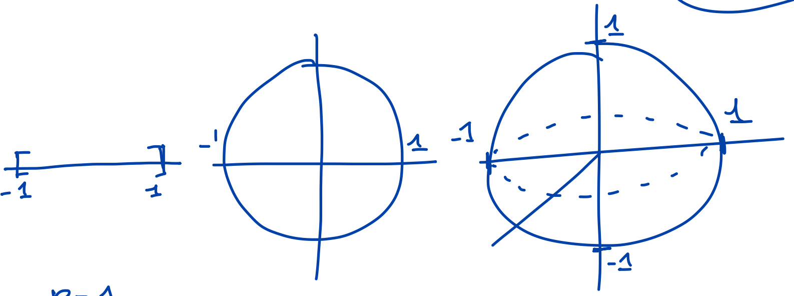
The blessings of dimensionality

As Bellman thought:

- all is not bad in high-dimensional spaces (hopefully!)
- there are interesting things which happen in high-dimensional spaces.

First example: volume of the unit sphere is $V(p) = \frac{\pi^{p/2}}{\Gamma(p/2+1)}$,

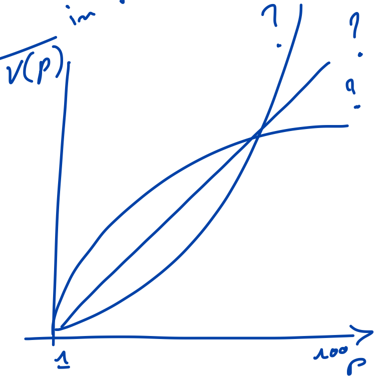
in R: Gamma( )



p = 1

p = 2

p = 3



The blessings of dimensionality

As Bellman thought:

- all is not bad in high-dimensional spaces (hopefully!)
- there are interesting things which happen in high-dimensional spaces.

First example: volume of the unit sphere is $V(p) = \frac{\pi^{p/2}}{\Gamma(p/2+1)}$,

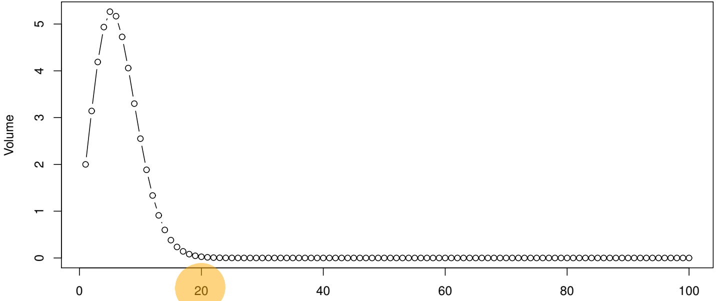
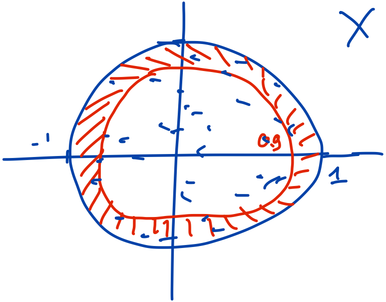


***Fig. Volume of a sphere of radius 1 regarding to the dimension** p.*

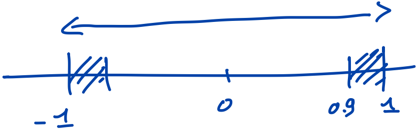
The blessings of dimensionality

Second example: probability that a uniform variable on the unit sphere belongs to the shell between the spheres of radius 0.9 and 1 is

$$P(X \in S_{0.9}(p)) = 1 - 0.9^p \underset{p \to \infty}{\longrightarrow} 1$$



$X \sim U(HS1)$



$P(X \in S_{0.9}) = 0.1$

The blessings of dimensionality

Second example: probability that a uniform variable on the unit sphere belongs to the shell between the spheres of radius 0.9 and 1 is

$$P(X \in S_{0.9}(p)) = 1 - 0.9^p \underset{p\to\infty}{\longrightarrow} 1$$

⚠ The message here is that HD spaces are mostly empty. The data are located in subspaces of lower dimensionalities.

⇒ the "empty space" phenomenon

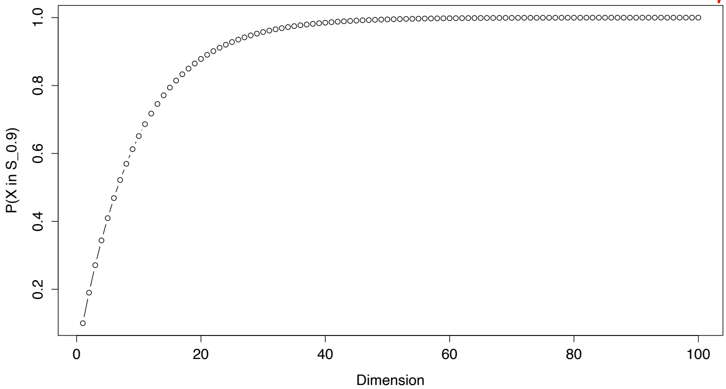


Fig. Probability that X belongs to the shell $S_{0.9}$ regarding to the dimension p.

The blessings of dimensionality

Third example:

- since high-dimensional spaces are almost empty,
- it should be easier to separate groups in high-dimensional space with an adapted classifier,
- a way to observe this is to look at the Bayes classifier behaviour.

The blessings of dimensionality

Third example:

- since high-dimensional spaces are almost empty,
- it should be easier to separate groups in high-dimensional space with an adapted classifier,
- a way to observe this is to look at the Bayes classifier behaviour.

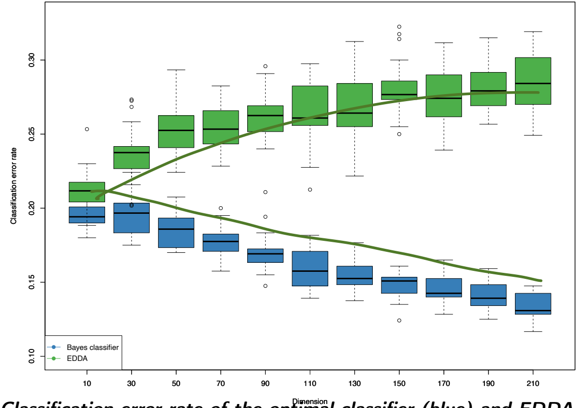


Fig. Classification error rate of the optimal classifier (blue) and EDDA (green) versus the data dimension on simulated data.

Classical ways to avoid the curse of dimensionality

Dimension reduction:

- the problem comes from that p is too large,
- therefore, reduce the data dimension to $d \ll p$,
- such that the curse of dimensionality vanishes!

Regularization:

- the problem comes from that parameter estimates are unstable,
- therefore, regularize these estimates,
- such that the parameter are correctly estimated!

$$\tilde{\Sigma}^{-1} = (\hat{\Sigma} + \varepsilon I)^{-1}$$

ε: $\underline{0.01}$

Parsimonious models:

- the problem comes from that the number of parameters to estimate is too large,
- therefore, make restrictive assumptions on the model,
- such that the number of parameters to estimate becomes more "decent"!

$$\Sigma = \sigma^2 I$$

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.

16

16 $\boxed{3}$

PCA

best discriminative subspace of dim 2.

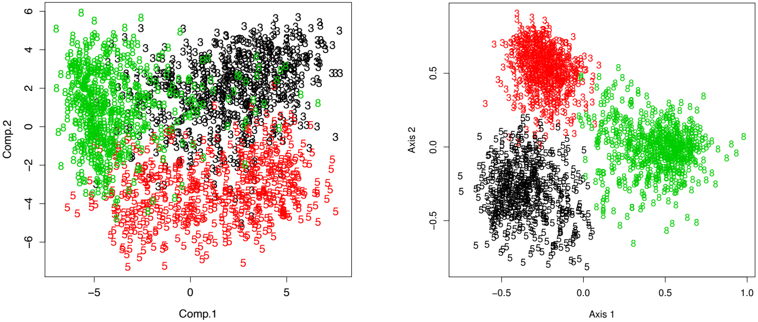
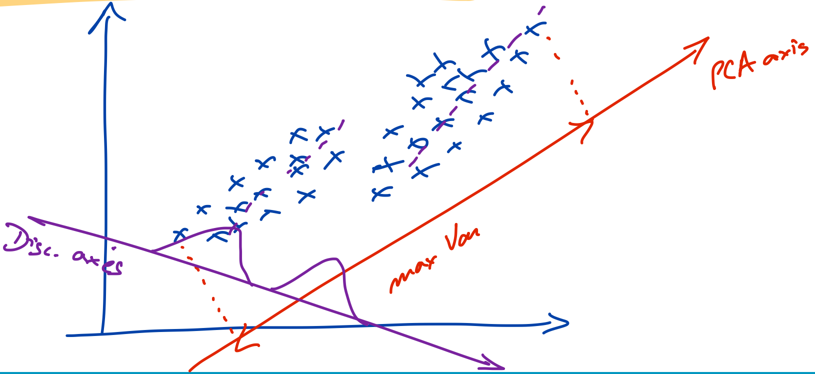


Figure: Projection of the 256-dimensional USPS data with PCA (left, unsupervised) and FDA (right, supervised).

PCA: principal component analysis

The goal of PCA is to find d new axes build on the p original axes by linear combinations such that the variance of the projected data is maximum.



PCA: the principle

The goal of PCA : $\underset{\substack{U \\ p \times d}}{\arg\max} \; \mathrm{Var}(\underline{XU})$ ← projection of X in the d-dimensional space

Solution : the d PCA axes $u_1, \dots, u_d$ are the eigenvectors of $\bar{X}^t \bar{X} = S$ associated with the largest eigenvalues $\lambda_1, \dots \lambda_d$ of S.

In practice : 1) compute $S = \bar{X}^t \bar{X}$ where $\bar{X}$ are the centered data

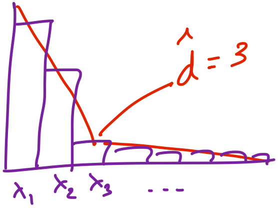
2) do the eigendecomposition of S.

3) keep the eigenvectors of S associated with the largest λ_j

PCA: the principle

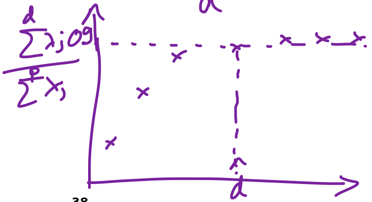
How to choose d?:

1) the "elbow" method: plot the eigenvalue scree and select d in the place of the elbow in the scree



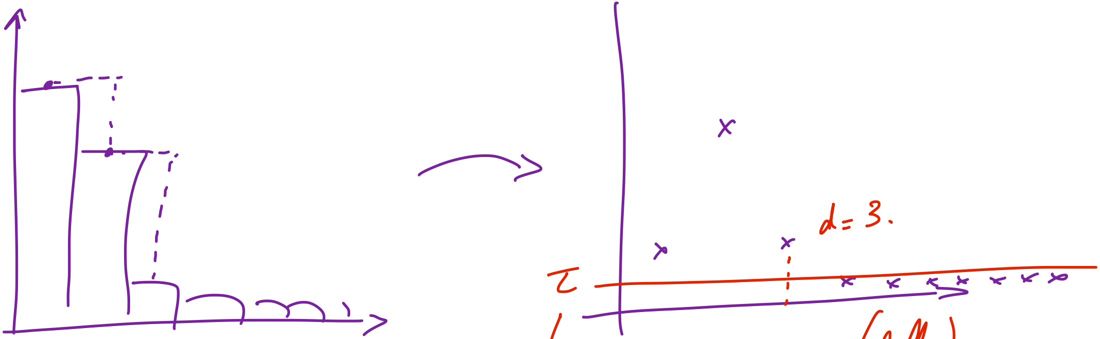
2) the "90% rule":

$$\hat{d} = \underset{d}{\operatorname{argmin}} \sum_{j=1}^{d} \lambda_j \Big/ \sum_{j=1}^{p} \lambda_j \geq 90\%.$$



3) the scree-test of Cattell

Compute the differences between consecutive eigenvalues and select d when all differences after d are smaller than a threshold



↳ $\tau = 0.1 \max(\text{diff})$

__A general remark:__ in case of hesitation, it is better to keep too much dimensions than too few!

PCA: projection

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

will contain the projected data points on the d-dimensional PCA subspace.

principal scores

the loading matrix.

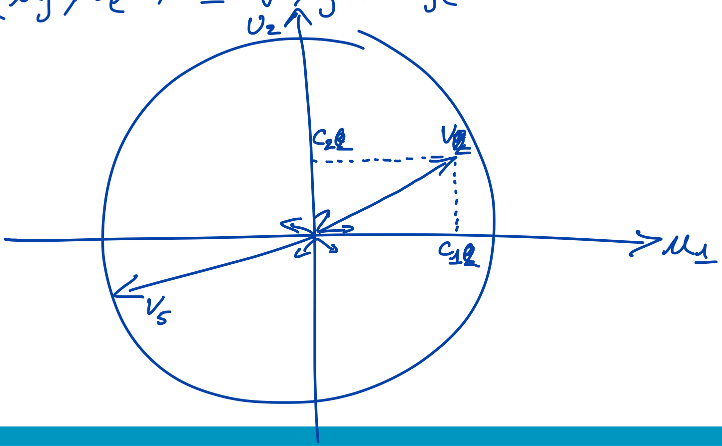
PCA: how many axes?

PCA: how many axes?

PCA: correlation circle

It is easy to relate the PCA axes with the original variables by calculating the correlation between them:

$$c_{j\ell} = \mathrm{Cor}(u_j, v_\ell) = \sqrt{\lambda_j} \cdot u_{j\ell}$$



PCA: analysis of the marathon data

PPCA: a probabilistic version of PCA

Probabilistic PCA (Tipping & Bishop, 1996) extends PCA in a statistical framework. PPCA justifies PCA under the Gaussian assumption.

PPCA assumes a linear link between X the observed variable and Y a latent low-dimensional variable:

$$X = U^t Y + \varepsilon$$

where $X \in \mathbb{R}^p$

$Y \in \mathbb{R}^d$ with $d \ll p$

$\varepsilon \in \mathbb{R}^p$ is a noise term.

U is a $p \times d$ matrix

PPCA: a probabilistic version of PCA

PPCA adds some statistical assumptions:

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

$$Y \sim \mathcal{N}(\mu, I_d) \quad \text{with } \mu \in \mathbb{R}^d$$

Interestingly, the marginal distribution of X is:

$$X \sim \mathcal{N}(U\mu, U^t U + \sigma^2 I_p)$$

Let's introduce $\theta = \{\mu, \sigma^2, U\}$, the set of model parameters

Thanks to this statistical model, we have now a new way to perform PCA: estimate the model parameters of PPCA from the data.

$$\text{performing PCA} \iff \text{estimate the models of PPCA.}$$

Maximum Likelihood can be used for that, through the EM algorithm.

Theorem: the estimation of U_{PPCA} through the EM algo is the eigenvectors of $S = \bar{X}^t \bar{X}$ associated with the largest eigenvalues of S.

PPCA: why is it interesting?

i) PPCA offers a theoretical justification of PCA under the Gaussian assumption.

ii) it also helps to understand the limitation of PCA

iii) we can rely on model selection (AIC, BIC, ...) to select d

iv) new statistical models can be ellaborated from PPCA.

PPCA in practice...

$\hat{U} = \text{eigen}(\text{cov}(X))$

1) Let's do PCA as before! $\hat{\mu} = \text{mean}(XU)$

$\hat{\sigma}^2 = \text{mean}\left(\text{diag}\left(\text{cov}\left(X - X\hat{U}\hat{U}^t\right)\right)\right)$

2) Use BIC to select d:

$$\text{BIC}(d) = \log L(X; \hat{\theta}_{PPCA}) - \frac{\gamma(d)}{2} \log(n)$$

where $\gamma(d) = \underset{\mu}{d} + \underset{\sigma^2}{1} + \underset{\substack{U \\ p \times d}}{p\frac{(d+1)}{2}}$

Outline

The limits of GMM in high dimensional spaces

⊕ Empty HD spaces ⇒ good for clustering

⊖ Model-based approaches are usually over-parametrized which is problem when learning!

⇒ The solution: adapt the statistical techniques to HD spaces by taking into account the properties of those spaces.

Parsimonious models for GMM

$\Sigma_k = \sigma^2 I_d$

$\Sigma_k = \sigma_k I_d$

$\Sigma_k = \Sigma$

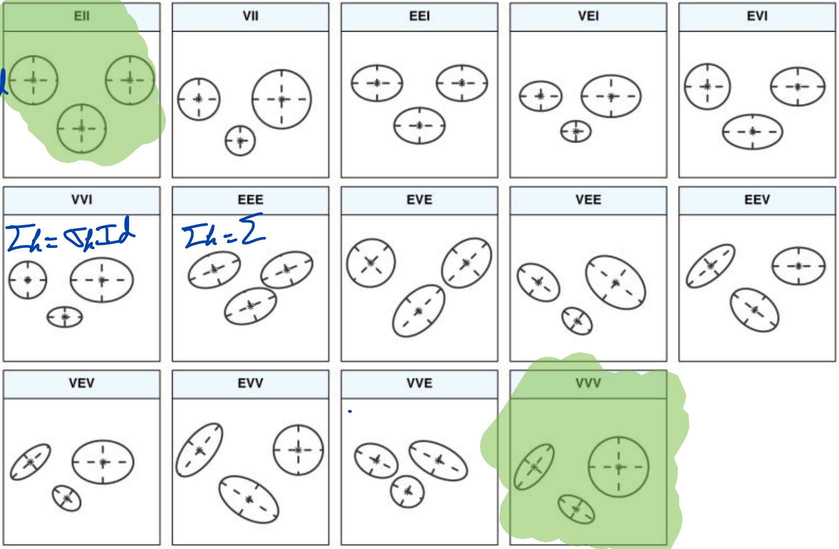
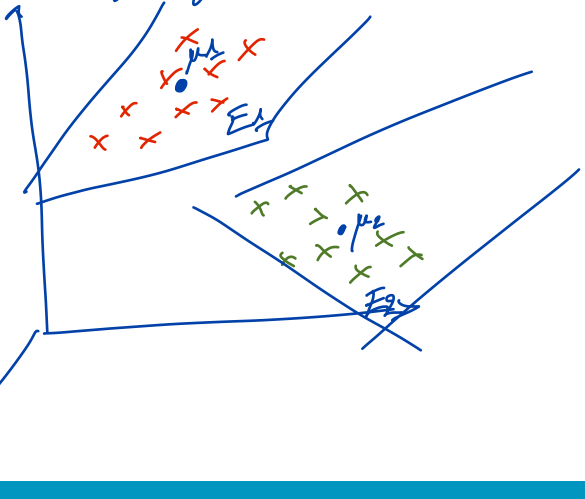


Figure: The parsimonious models of Mclust.

The solutions : Subspace clustering

The idea of subspace clustering is to assume that each group leaves in a specific low-dimensional subspace.



The mixture of PPCA

Tipping & Bishop proposed in 1996 the mixture of PPCA

$$Z \sim \mathcal{M}(1;\pi)$$

$$X_{|Z=k} = Y U_k + \varepsilon_k$$ where U_k is a $p \times d$ matrix

$$\varepsilon_k \sim \mathcal{N}(0, \sigma_k^2 I_p)$$

$$Y|Z=k \sim \mathcal{N}_k(\mu_k, I_d)$$

$$\Rightarrow X \sim \sum_{k=1} \pi_k \, \phi(\mu_k U_k, U_k^t U_k + \sigma_k^2 I_p)$$