High-dimensional statistics

Christophe Giraud

Université Paris Saclay

Geometry and Statistics in Data Sciences, Cargèse 2022

Informations on the lectures

Plan

5 lectures of 1h30

1. Curse of dimensionality
2. Structure learning
3. Convexification
4. Iterative algorithms
5. Implicit regularisation, benign overfitting and overparametrisation

A mixture of very standard materials and some more recent results.
(I will adapt depending on your background)

Emphasize on ideas, concepts and intuitions, rather than on mathematical technics.

Documents

Documents

- Website of the lectures
  https://www.imo.universite-paris-saclay.fr/~giraud/Orsay/Cargese2022.html
  (where you can download the slides)
- Book available online (pdf)
  https://www.imo.universite-paris-saclay.fr/~giraud/Orsay/Bookv3.pdf
- A wiki website for sharing solutions to the exercises
  http://high-dimensional-statistics.wikidot.com
- Youtube channel: related lectures on High-Dimensional Statistics
  https://www.youtube.com/channel/UCDo2g5DETs2s-GKu9-jT_BQ

High-dimensional data

High-dimension data

- biotech data (sense thousands of features)
- images (millions of pixels / voxels)
- web data
- crowdsourcing data
- etc

Blessing?

☺ we can sense thousands of variables on each "individual" : potentially we will be able to scan every variables that may influence the phenomenon under study.

☹ the curse of dimensionality : separating the signal from the noise is in general almost impossible in high-dimensional data and computations can rapidly exceed the available resources.

Probability in high-dimension

Chapter 1

A ball is essentially a sphere

Volume of a ball $B_p(0, r)$ **of radius** r**:** $V_p(r) = r^p V_p(1)$

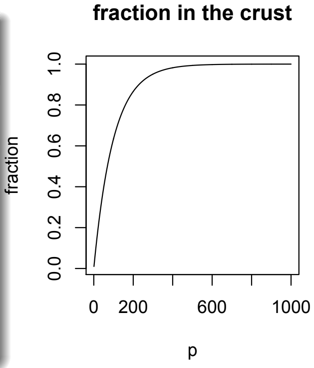
The volume of a high-dimensional ball is concentrated in its crust!

Crust: $C_p(r) = B_p(0, r) \setminus B_p(0, 0.99r)$

The fraction of the volume in the crust

$$\frac{\text{volume}(C_p(r))}{\text{volume}(B_p(0, r))} = 1 - 0.99^p$$

goes exponentially fast to 1!

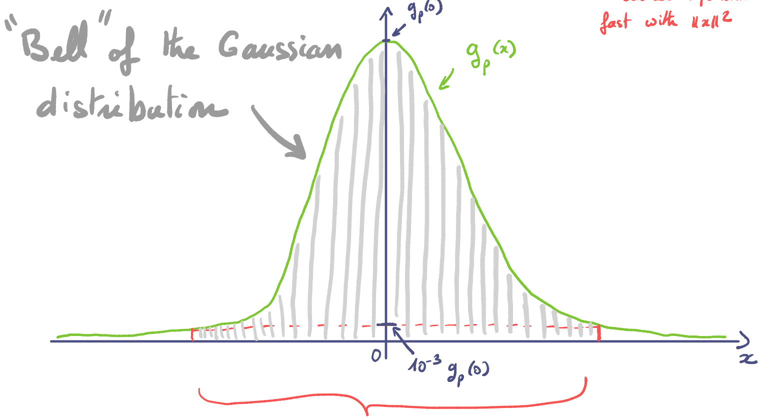


Forget your low-dimensional intuitions!

Thin tails can concentrate the mass!

Gaussian distribution in $\mathbb{R}^p$: $g_p(x) = \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}\|x\|^2\right)$

decreases exponentially fast with $\|x\|^2$



$\mathcal{B} = \{x : g_p(x) \geq 10^{-3} \cdot g_p(0)\}$

Thin tails can concentrate the mass!

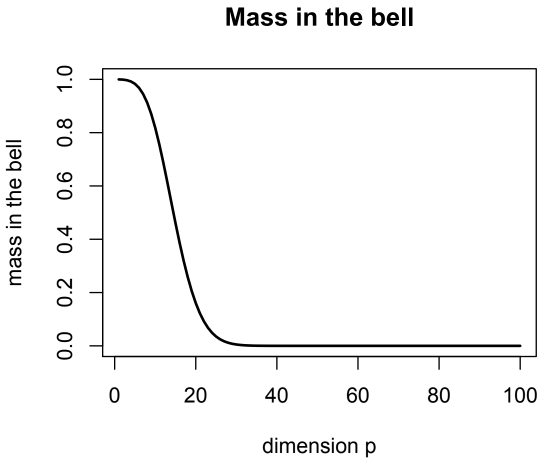


Figure: Mass of the standard Gaussian distribution $g_p(x)$ in the "bell" $\mathcal{B} = \{x \in \mathbb{R}^p : g_p(x) \geq 0.001 g_p(0)\}$ for increasing dimension p.

Thin tails can concentrate the mass!

Where is the Gaussian mass located?

For $X \sim \mathcal{N}(0, I_p)$ and $\varepsilon > 0$ small

$$\frac{1}{\varepsilon}\mathbb{P}[R \leq \|X\| \leq R+\varepsilon] = \frac{1}{\varepsilon}\int_{R\leq\|x\|\leq R+\varepsilon} e^{-\|x\|^2/2}\ \frac{dx}{(2\pi)^{p/2}}$$
$$= \frac{1}{\varepsilon}\int_{R}^{R+\varepsilon} e^{-r^2/2}\ r^{p-1}\frac{pV_p(1)\,dr}{(2\pi)^{p/2}}$$
$$\approx \frac{p}{2^{p/2}\Gamma(1+p/2)}\ R^{p-1} \times e^{-R^2/2}.$$

This mass is concentrated around $R^* = \sqrt{p-1}$!

Remark: the density ratio $\frac{g_p(R^*)}{g_p(0)}$ is smaller than $2e^{-p/2}$.

Thin tails can concentrate the mass!

Concentration of the square Norm

Let $X \sim \mathcal{N}(0, I_p)$. We have for all $x \geq 0$

$$\mathbb{P}\left[p - 2\sqrt{px} \leq \|X\|^2 \leq p + 2\sqrt{2px} + 2x\right] \geq 1 - 2e^{-x}.$$

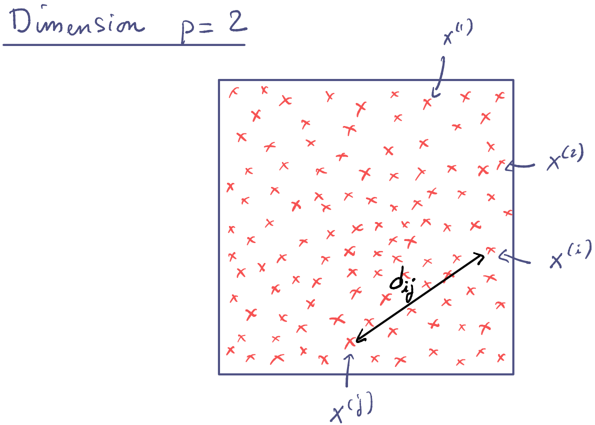
Proof: Chernoff bound (Exercise 1.6.6).

Gaussian $\approx$ Uniform on the sphere $S(0, \sqrt{p})$

As a first approximation, the Gaussian $\mathcal{N}(0, I_p)$ distribution can be thought as a uniform distribution on the sphere of radius $\approx \sqrt{p}$!

Lost in high-dimensional spaces

We sample $n = 100$ data points $X^{(1)}, \dots, X^{(n)} \overset{i.i.d.}{\sim} \mathcal{U}([0,1]^p)$ i.i.d. uniformly in the hypercube $[0,1]^p$.



let us look at the distribution of the pairwise distances $d_{ij} = \|X^{(i)} - X^{(j)}\|$ between the points.

Lost in high-dimensional spaces

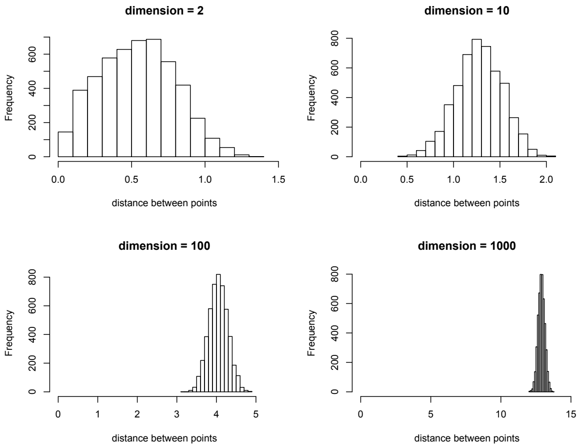


Figure: Histograms of the pairwise-distances between $n = 100$ points sampled uniformly in the hypercube $[0, 1]^p$, for $p = 2, 10, 100$ and 1000.

Lost in high-dimensional spaces

Square distances.

$$\mathbb{E}\left[\|X^{(i)} - X^{(j)}\|^2\right] = \sum_{k=1}^{p} \mathbb{E}\left[\left(X_k^{(i)} - X_k^{(j)}\right)^2\right] = p\,\mathbb{E}\left[(U - U')^2\right] = p/6,$$

with U, U' two independent random variables with $\mathcal{U}[0,1]$ distribution.

Standard deviation of the square distances

$$\begin{aligned}\mathrm{sdev}\left[\|X^{(i)} - X^{(j)}\|^2\right] &= \sqrt{\sum_{k=1}^{p} \mathrm{var}\left[\left(X_k^{(i)} - X_k^{(j)}\right)^2\right]} \\ &= \sqrt{p\,\mathrm{var}\left[(U' - U)^2\right]} \approx 0.2\sqrt{p}\,.\end{aligned}$$

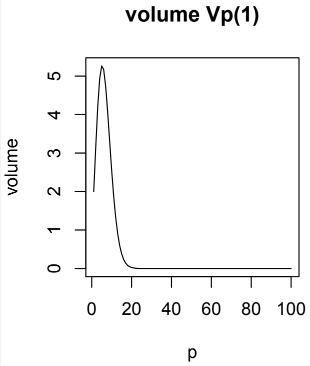
Lost in high-dimensional spaces

High-dimensional unit balls have a vanishing volume!

$$\begin{aligned} V_p(r) &= \text{volume of a ball of radius } r \\ &\quad \text{in dimension } p \\ &= r^p V_p(1) \end{aligned}$$

with

$$V_p(1) \overset{p\to\infty}{\sim} \left(\frac{2\pi e}{p}\right)^{p/2} (p\pi)^{-1/2}.$$



Vanishing volume for $r \leq \sqrt{\frac{p}{2\pi e}}$!

Take home message (so far)

In high-dimensional spaces,
be careful
not to be mislead by
your low dimensional intuitions.

The curse of dimensionality

Chapter 1

Curse 1 : fluctuations cumulate

Example : $X^{(1)}, \ldots, X^{(n)} \in \mathbb{R}^p$ i.i.d. with $\text{cov}(X) = \sigma^2 I_p$. We want to estimate $\mathbb{E}[X]$ with the sample mean

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X^{(i)}.$$

Then

$$\begin{aligned}
\mathbb{E}\left[\|\bar{X}_n - \mathbb{E}[X]\|^2\right] &= \sum_{j=1}^{p} \mathbb{E}\left[\left([\bar{X}_n]_j - \mathbb{E}[X_j]\right)^2\right] \\
&= \sum_{j=1}^{p} \mathrm{var}\left([\bar{X}_n]_j\right) = \frac{p}{n}\sigma^2.
\end{aligned}$$

☹ It can be huge when $p \gg n$...

Curse 2 : local averaging is ineffective (in general)

Observations $(Y_i, X^{(i)}) \in \mathbb{R} \times [0,1]^p$ for $i = 1, \ldots, n$.

Model: $Y_i = f(X^{(i)}) + \varepsilon_i$ with f smooth.

assume that $(Y_i, X^{(i)})_{i=1,\ldots,n}$ i.i.d. and that $X^{(i)} \sim \mathcal{U}([0,1]^p)$

Local averaging: $\widehat{f}(x) = \text{average of}\left\{ Y_i : X^{(i)} \text{ close to } x \right\}$

Problem: for $x \in [0,1]^p$, we have

$$\begin{aligned} \mathbb{P}\left[\exists i = 1, \ldots, n : \|x - X_i\| \leq \delta\right] &\leq n\,\mathbb{P}\left[\|x - X_1\| \leq \delta\right] \leq n\, V_p(\delta) \\ &\approx n \left(\frac{2\pi e}{p}\right)^{p/2} \frac{\delta^p}{\sqrt{\pi p}}. \end{aligned}$$

which goes more than exponentially fast to 0 when $p \to \infty$.

Curse 2 : local averaging is ineffective

Which sample size to avoid the lost of locality?

Number n of points $x_1, \ldots, x_n$ required for covering $[0,1]^p$ by the balls $B(x_i, 1)$:

$$n \geq \frac{1}{V_p(1)} \overset{p\to\infty}{\sim} \left(\frac{p}{2\pi e}\right)^{p/2} \sqrt{p\pi}$$

p	20	30	50	100	200
n	39	45630	$5.7\,10^{12}$	$42\,10^{39}$	larger than the estimated number of particles in the observable universe

Curse 3: weak signals are lost

Finding active genes: we observe n repetitions for p genes

$$Z_j^{(i)} = \theta_j + \varepsilon_j^{(i)}, \quad j = 1, \ldots, p, \quad i = 1, \ldots, n,$$

with the $\varepsilon_j^{(i)}$ i.i.d. with $\mathcal{N}(0, \sigma^2)$ Gaussian distribution.

Our goal: find which genes have $\theta_j \neq 0$

For a single gene

Set

$$X_j = n^{-1/2}(Z_j^{(1)} + \ldots + Z_j^{(n)}) \sim \mathcal{N}(\sqrt{n}\theta_j, \sigma^2)$$

Since $\mathbb{P}\left[|\mathcal{N}(0, \sigma^2)| \geq 2\sigma\right] \leq 0.05$, we can detect the active gene with X_j when

$$|\theta_j| \geq \frac{2\sigma}{\sqrt{n}}$$

Curse 3: weak signals are lost

Maximum of Gaussian

For $W_1, \dots, W_p$ i.i.d. with $\mathcal{N}(0, \sigma^2)$ distribution, we have

$$\max_{j=1,\dots,p} W_j \approx \sigma\sqrt{2\log(p)}.$$

Consequence: When we consider the p genes together, we need a signal of order

$$|\theta_j| \geq \sigma\sqrt{\frac{2\log(p)}{n}}$$

in order to dominate the noise ☹

Curse 4: an accumulation of rare events may not be rare

Empirical covariance matrix:

- Let $X_i \overset{iid}{\sim} \mathcal{N}(0, I_p)$ and set

$$\hat{\Sigma} = \frac{1}{m}\sum_{i=1}^{m} X_i X_i^T .$$

- For any $u \in \mathbb{R}^p$ with $\|u\| = 1$:

$$u^T \hat{\Sigma} u = \frac{1}{m}\sum_{i=1}^{m}(X_i^T u)^2 = \frac{1}{m}\|\zeta_u\|^2$$

where $\zeta_u = \begin{bmatrix} X_1^T u \\ \vdots \\ X_m^T u \end{bmatrix} \sim \mathcal{N}(0, I_m)$

- Chernoff bound: (exercise 1.6.6)
For all $L > 0$:

$$\mathbb{P}\left[\frac{1}{m}\|\zeta_u\|^2 \geq \left(1 + \sqrt{\frac{2L}{m}}\right)^2\right] \leq e^{-L}$$

In particular, for any $u \in \mathbb{R}^p$ with $\|u\| = 1$

$$\mathbb{P}\left[u^T \hat{\Sigma} u \geq \left(1 + \sqrt{\frac{p}{m}}\right)^2\right] \leq e^{-p/2}$$

Yet $\mathbb{E}[|\hat{\Sigma}|_{op}] = \mathbb{E}[\max_{\|u\|=1} u^T \hat{\Sigma} u]$

$$\simeq \left(1 + \sqrt{\frac{p}{m}}\right)^2$$

$$\simeq \frac{p}{m} \text{ if } p \gg m$$

- Why $\frac{p}{m}$? for $p \geq m$

→ $\text{rank}(\hat{\Sigma}) = m$ a.s.

→ $m\,\mathbb{E}[|\hat{\Sigma}|_{op}] \geq \mathbb{E}[\text{Tr}(\hat{\Sigma})]$

$$= \text{Tr}(\underbrace{\mathbb{E}[\hat{\Sigma}]}_{= I_p}) = p$$

$$\Rightarrow \mathbb{E}[|\hat{\Sigma}|_{op}] \geq \frac{p}{m}$$

(We will come back to this phenomenon in the last lecture)

Algorithmic complexity must remain low

When p is large, an algorithmic complexity larger than $O(p^2)$ is computationally prohibitive.

For very large p, even a complexity $O(p^2)$ can be an issue...

Low-dimensional structures in high-dimensional data

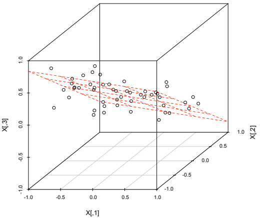
Hopeless?

Low dimensional structures : high-dimensional data are usually concentrated around low-dimensional structures reflecting the (relatively) small complexity of the systems producing the data

- geometrical structures in an image,
- regulation network of a "biological system",
- social structures in marketing data,
- human technologies have limited complexity, etc.

Dimension reduction :

- "unsupervised" (PCA)
- "supervised"

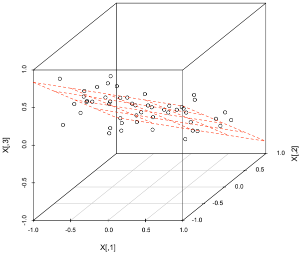


Principal Component Analysis

For any data points $X^{(1)}, \dots, X^{(n)} \in \mathbb{R}^p$ and any dimension $d \le p$, the PCA computes the linear span in $\mathbb{R}^p$

$$V_d \in \underset{\dim(V) \le d}{\operatorname{argmin}} \sum_{i=1}^{n} \|X^{(i)} - \mathrm{Proj}_V X^{(i)}\|^2,$$

where Proj_V is the orthogonal projection matrix onto V.

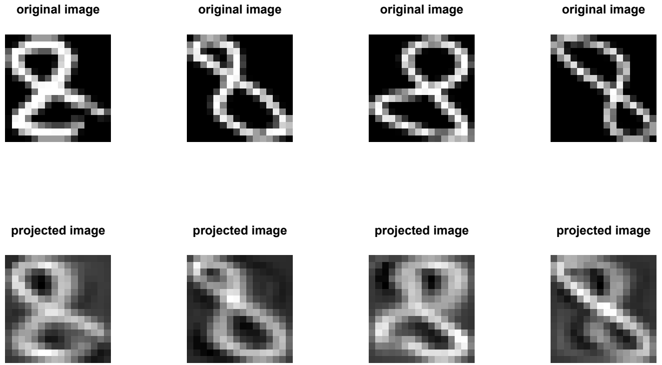


V_2 in dimension $p = 3$.

Recap on PCA

Exercise 1.6.4

PCA in action



MNIST : 1100 scans of each digit. Each scan is a 16×16 image which is encoded by a vector in $\mathbb{R}^{256}$. The original images are displayed in the first row, their projection onto 10 first principal axes in the second row.

"Supervised" dimension reduction

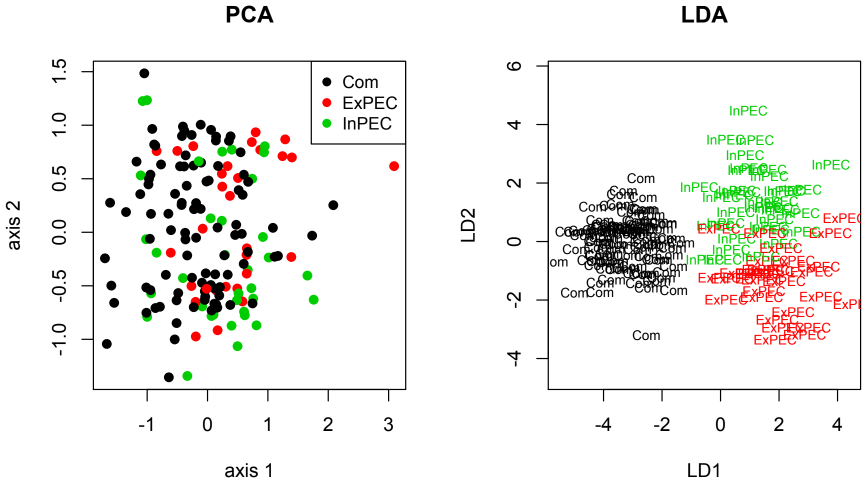


Figure: 55 chemical measurements of 162 strains of *E. coli*.
Left : the data is projected on the plane given by a PCA.
Right : the data is projected on the plane given by a LDA.

Summary

Statistical difficulty

- high-dimensional data
- relatively small sample size

Good feature

Data usually generated by a large stochastic system

- existence of low dimensional structures
- (sometimes: expert models)

The way to success

Finding, from the data, the hidden structure in order to exploit them.

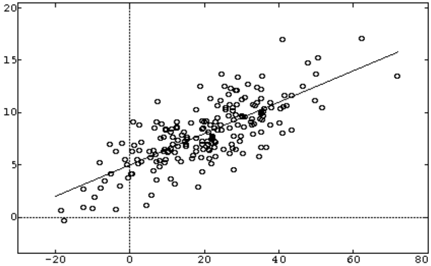
Paradigm shift

Chapter 1

Paradigm shift

Classical statistics:

- small number p of parameters
- large number n of observations
- we investigate the performances of the estimators when $n \to \infty$ (central limit theorem...)



Paradigm shift

Classical statistics:

- small number p of parameters
- large number n of observations
- we investigate the performances of the estimators when $n \to \infty$ (central limit theorem...)

Actual data:

- inflation of the number p of parameters
- small sample size: $n \approx p$ ou $n \ll p$

$\Longrightarrow$ Change our point of view on statistics!
(the $n \to \infty$ asymptotic does not fit anymore)

Statistical settings

- double asymptotic: both $n, p \to \infty$ with $p \sim g(n)$
- non asymptotic: treat n and p as they are

Double asymptotic

- more easy to analyse, sharp results ☺
- but sensitive to the choice of g ☹

ex: if $n = 33$ and $p = 1000$, do we have $g(n) = n^2$ or $g(n) = e^{n/5}$?

Non-asymptotic

- no ambiguity ☺
- but the analysis is more involved ☹
  (based on concentration inequalities)