High-dimensional statistics and probability

Christophe Giraud[1], Matthieu Lerasle[2] and Zacharie Naulet[3,1]

(1) Université Paris-Saclay
(2) ENSAE
(3) INRAE - Jouy en Josas

M2 Maths Aléa & MathSV & Math-IA

Informations on the course

Objective

1. To understand the main features of high-dimensional observations;
2. To learn the mains concepts and methods to handle the curse of dimensionality;
3. To get prepared for a PhD in statistics or machine learning
4. [MSV] Some complement and biological illustrations by Z. Naulet.

$\longrightarrow$ conceptual and mathematical course

$\longrightarrow$ blackboard course (except today)

Agenda (1/2)

Structure

The course has two parts

- **Part 1** [MDA+MSV+MIA]: 6 weeks with C. Giraud: central concepts in high-dimensional statistics
- **Part 2** [MDA]: 6 weeks with M. Lerasle: essential probabilistic tools for stats and ML
- **Part 2** [MSV]: 3 weeks with Z. Naulet: false discoveries, supervised classification and illustrations

Agenda (2/2)

[MDA+MSV]+MIA 26/09 – 07/11

1. Curse of dimensionality + principle of model selection
2. Model selection theory
3. Information theoretic lower bounds
4. Convexification: principle and theory
5. Iterative algorithms
6. Low rank regression

MDA (Matthieu)

6 weeks on central probabilistic tools for ML and statistics

MSV (Zacharie)

3 weeks on false discoveries, supervised classification, algorithmic aspects, and illustrations. November 14, 21, and 28 from 15h to 18h, room 1A11.

Organisation

Organisation for the first part

- Lectures: the lectures will take place every Thursday (26/09 – 07/11) at 15h-19h room 0A1. A recorded version of the lectures (2020) is available on the Youtube channel
  https://www.youtube.com/channel/UCDo2g5DETs2s-GKu9-jT_BQ
- Lecture notes: lectures notes are available on the website of the course https://www.imo.universite-paris-saclay.fr/~giraud/Orsay/HDPS.html as well as handwritten notes for each lecture
- Exercises: the list of assigned exercises is given on the website
- December 19: exam on the first part of the course
  - 7 pt: on 1 or 2 exercises from the assigned list
  - 13pt: research oriented problem

Learn by doing

- you follow actively the lectures:
  - you try to understand all the explanations;
  - if a point is not clear, please ask questions. You can also look back at the explanations on the lecture notes and the Youtube channel.
- you work out the lecture notes: take a pen and a sheet of paper, and redo all the computations. You have understood something, only when you are able to
  - explain it to someone else;
  - answer the question "why have we done this, instead of anything else?"
- you work out the assigned exercises.
- you interact with the others: discussing with the others is very efficient for making progress (both when explaining something, and when receiving an explanation).

Documents

Documents

- Lecture notes: pdf & printed versions, handwritten notes
- Website of the course
  https://www.imo.universite-paris-saclay.fr/~giraud/Orsay/HDPS.html
- Youtube channel
  https://www.youtube.com/channel/UCDo2g5DETs2s-GKu9-jT_BQ
- A wiki website for sharing solutions to the exercises
  http://high-dimensional-statistics.wikidot.com

Evaluation

[MDA+MSV+MIA] Exam December 19

- 1 or 2 (part of) exercises of the list (7/20)
  - list = those on the website
    https://www.imo.universite-paris-saclay.fr/~giraud/Orsay/HDPS.html
- a research oriented problem (13/20)
- you can take with you the printed lecture notes

[MDA] second evaluation in January

Project related to the material presented by Matthieu Lerasle

Any questions so far?

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

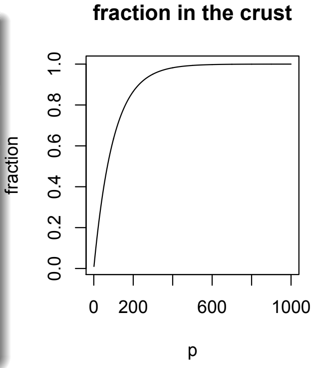
The volume of a high-dimensional ball is concentrated in its crust!

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

The fraction of the volume in the crust

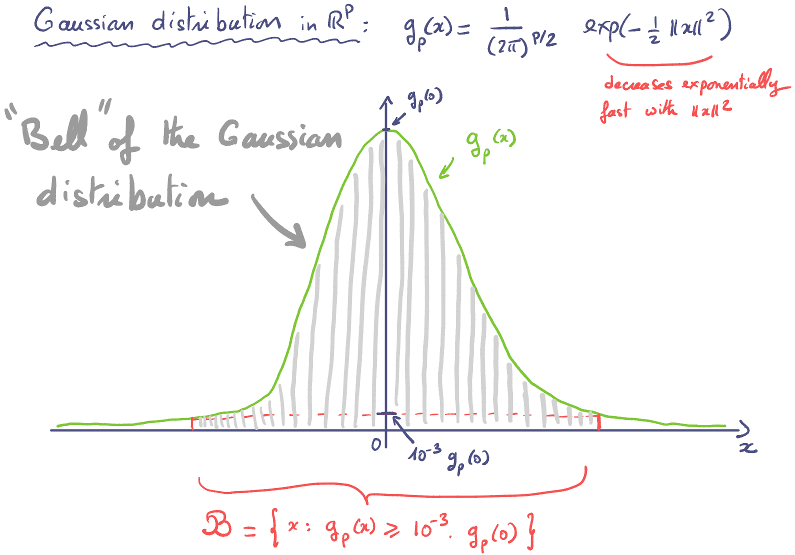
$$\frac{\text{volume}(C_p(r))}{\text{volume}(B_p(0, r))} = 1 - 0.99^p$$

goes exponentially fast to 1!



Forget your low-dimensional intuitions!

Thin tails can concentrate the mass!



Thin tails can concentrate the mass!

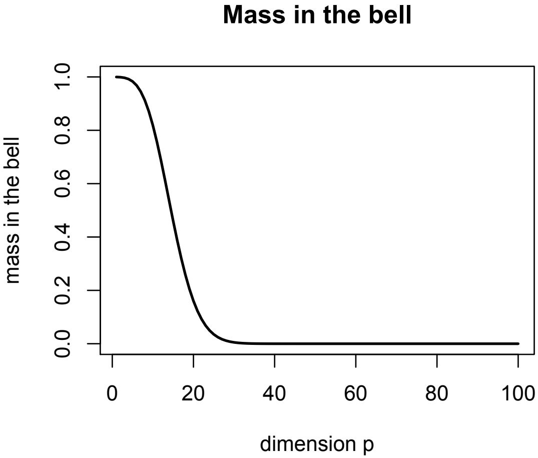


Figure: Mass of the standard Gaussian distribution $g_p(x)$ in the "bell" $\mathcal{B} = \{x \in \mathbb{R}^p : g_p(x) \geq 0.001 g_p(0)\}$ for increasing dimension p.

Thin tails can concentrate the mass!

Where is the Gaussian mass located?

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

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

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 $x \geq 0$

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

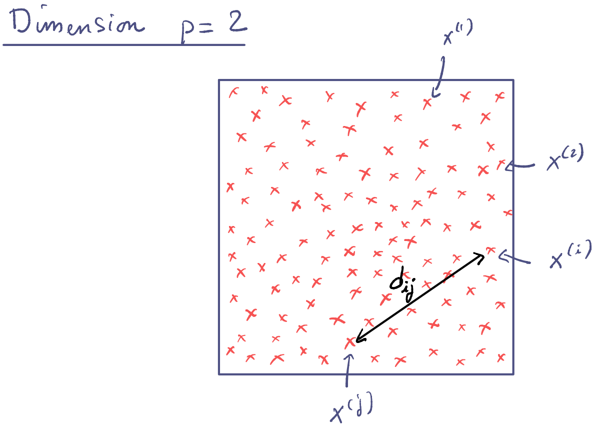
Proof: Chernoff bound (Exercise 1.6.6).

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

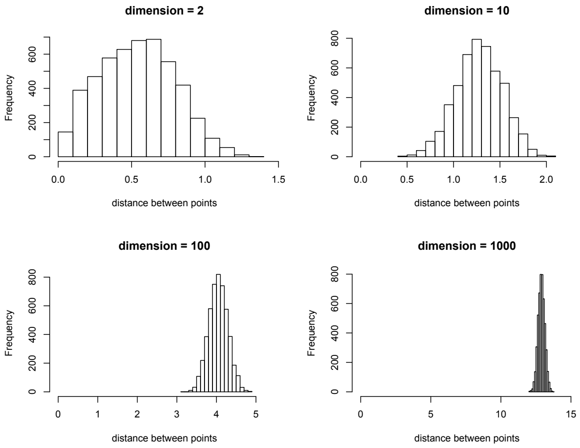


Figure: Histograms of the pairwise-distances between $n = 100$ points sampled uniformly in the hypercube $[0,1]^p$, for $p = 2, 10, 100$ and 1000.

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

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

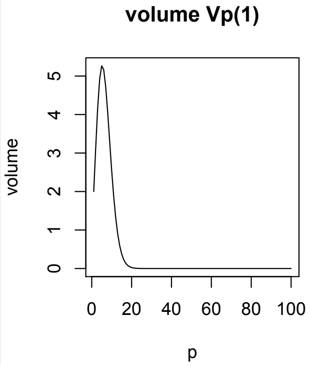
Lost in high-dimensional spaces

High-dimensional unit balls have a vanishing volume!

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

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

Unreliable empirical covariance matrix

Empirical covariance in High-Dimension

- Let $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(0, \underset{p\times p}{\Sigma})$ with $\Sigma = I_p$.
- $Sp(\Sigma) = (1, \dots, 1)$ (p times)
- Empirical covariance

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

We have $\mathrm{rank}(\hat{\Sigma}) = n$ so

$$n\,\mathbb{E}[|\hat{\Sigma}|_{op}] \geq \mathbb{E}[\mathrm{Tr}(\hat{\Sigma})] = \mathrm{Tr}(\mathbb{E}[\hat{\Sigma}]) = \mathrm{Tr}\Big(\frac{1}{n}\sum_{i=1}^{n}\underbrace{\mathbb{E}[X_i X_i^T]}_{=\Sigma}\Big) = \mathrm{Tr}(\Sigma) = p$$

So $\mathbb{E}[|\hat{\Sigma}|_{op}] \geq \frac{p}{n} \gg 1 = |\Sigma|_{op}$ if $p \gg n$

- Furthermore, we can prove (later) that

$$\mathbb{E}[|\hat{\Sigma}|_{op}] \leq \left(1 + \sqrt{\frac{p}{n}}\right)^2 \underset{p\gg n}{=} \frac{p}{n}(1+o(1))$$

So

$$Sp(\hat{\Sigma}) \underset{p\gg n}{\approx} \underbrace{\left(\frac{p}{n}(1+o(1)), \dots, \frac{p}{n}(1+o(1))\right)}_{n \text{ times}}$$

⟶ very different from $Sp(\Sigma)$,

so we cannot rely on $\hat{\Sigma}$ when $p \gg n$.

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 : $X^{(1)}, \dots, X^{(n)} \in \mathbb{R}^p$ 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

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

☹ 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 $x_1, \dots, x_n$ required for having at least one observation at distance $\delta = 1$ with probability $1/2$:

$$n \geq \frac{1}{2V_p(1)} \overset{p\to\infty}{\sim} \left(\frac{p}{2\pi e}\right)^{p/2} \sqrt{\frac{p\pi}{4}}$$

p	20	30	50	100	200
n	39	45630	$6 \,.\, 10^{12}$	$42 \,.\, 10^{39}$	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 $W_1, \ldots, W_p$ i.i.d. with $\mathcal{N}(0, \sigma^2)$ distribution, we have

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

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 ☹

Some other curses

- Curse 6 : an accumulation of rare events may not be rare (false discoveries, etc)
- Curse 7 : 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...
- etc

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.

Back to low dimensional statistics! Projecting the data on the low-dimensional structures, we are back to classical low-dimensional statistics ☺

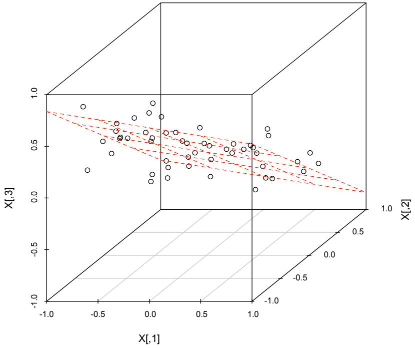
Low-dimensional structures in high-dimensional data

But the low-dimensional structures are unknown!

☹

Dimension reduction :

- "unsupervised" (PCA)
- "supervised"

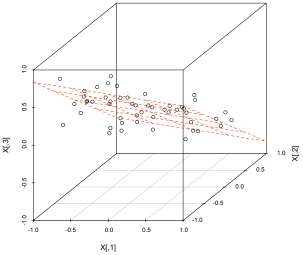


Principal Component Analysis

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

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

where Proj_V is the orthogonal projection matrix onto V.

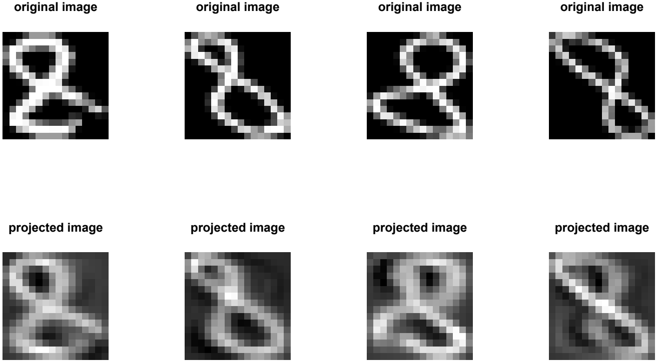


V_2 in dimension $p = 3$.

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

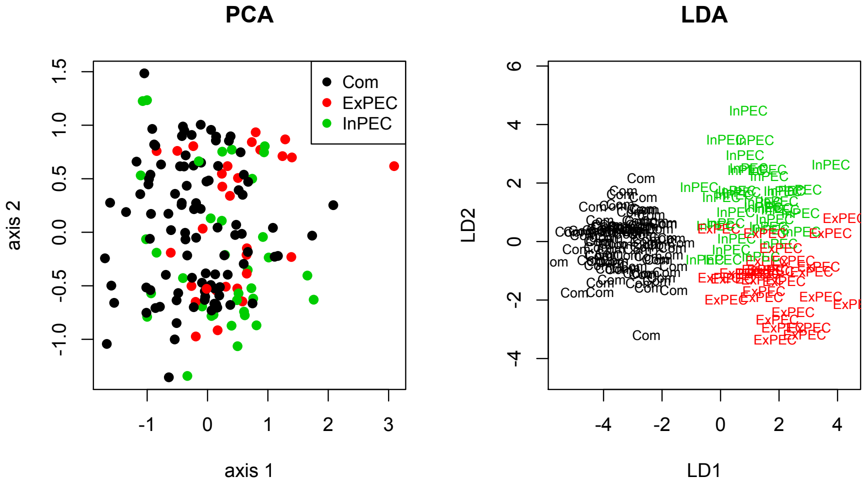


Figure: 55 chemical measurements of 162 strains of *E. coli*.
Left : the data is projected on the plane given by a PCA.
Right : the data is projected on the plane given by a LDA.

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.

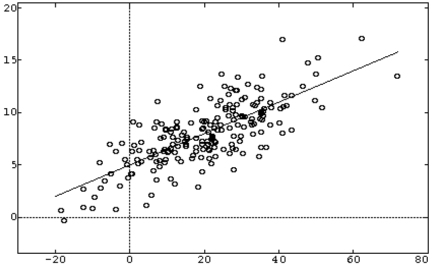
Paradigm shift

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: $n \approx p$ or $n \ll p$

$\Longrightarrow$ 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)