High-dimensional statistics and probability

Christophe Giraud\textsuperscript{1}, Matthieu Lerasle\textsuperscript{2,3} and Tristan Mary-Huard\textsuperscript{4,5}

\begin{itemize}
  \item \textsuperscript{1} Université Paris-Saclay
  \item \textsuperscript{2} CNRS
  \item \textsuperscript{3} ENSAE
  \item \textsuperscript{4} AgroParistech
  \item \textsuperscript{5} INRA - Le Moulon
\end{itemize}

M2 Maths Aléa & MathSV
Informations on the course
Objective

1. To understand the main features of high-dimensional observations;
2. To learn the main concepts and methods to handle the curse of dimensionality;
3. To get prepared for a PhD in statistics or machine learning;

→ conceptual and mathematical course
Structure

The course has two parts

- **Part 1** [MDA + MSV]: 7 weeks with C. Giraud: central concepts in the simple Gaussian setting
- **Part 2** [MDA]: 7 weeks with M. Lerasle: essential probabilistic tools for stats and ML
- **Part 2** [MSV]: 3 weeks with T. Mary-Huard: supervised classification and illustrations
## Agenda (2/2)

**[MDA+MSV] 23/09 – 10/11**

<table>
<thead>
<tr>
<th>1</th>
<th>Curse of dimensionality + principle of model selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Model selection theory</td>
</tr>
<tr>
<td>3</td>
<td>Information theoretic lower bounds</td>
</tr>
<tr>
<td>4</td>
<td>Convexification: principle and theory</td>
</tr>
<tr>
<td>5</td>
<td>Iterative algorithms</td>
</tr>
<tr>
<td>6</td>
<td>Low rank regression</td>
</tr>
<tr>
<td>7</td>
<td>False discoveries and multiple testing</td>
</tr>
</tbody>
</table>

**MDA (Matthieu)**

7 weeks on central probabilistic tools for ML and statistics

**MSV (Tristan)**

3 weeks on supervised classification, algorithmic aspects, and illustrations. October 4, 11 and 18. Orsay, room 0A5, 9h-12h.
Organisation

Organisation for the first part

- **Lectures:** the lectures will take place every Thursday (23/09 – 28/10) at 3 p.m. room 0A1, and 10/11 at 3 p.m. room 2L8. A recorded version of the lectures (2020) is available on the Youtube channel [https://www.youtube.com/channel/UCDo2g5DETs2s-GKu9-jT_BQ](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](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 16:** exam on the first part of the course
  - 7 pt: on 1 or 2 exercises from the assigned list
  - 13pt: research 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

- 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] Exam December 16

- 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 problem (13/20)

[MDA] second evaluation in January

mainly on the material presented by Matthieu Lerasle
Any questions so far?
High-dimensional data

Chapter 1
High-dimension data

- biotech data (sense thousands of features)
- images (millions of pixels / voxels)
- marketing, business 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 generally almost impossible in high-dimensional data and computations can rapidly exceed the available resources.
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

\[
\mathbb{E} \left[ ||\bar{X}_n - \mathbb{E}[X]||^2 \right] = \sum_{j=1}^{p} \mathbb{E} \left[ (\bar{X}_n)_j - \mathbb{E}[X_j])^2 \right]
\]

\[
= \sum_{j=1}^{p} \text{var} (\bar{X}_n)_j = \frac{p}{n} \sigma^2.
\]

It can be huge when \( p \gg n \)...
Curse 2: locality is lost

**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 \(X^{(i)} \sim U([0, 1]^p)\).

**Local averaging:** \(\hat{f}(x) = \text{average of } \{Y_i : X^{(i)} \text{ close to } x\}\)
Curse 2: locality is lost

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.
Why?

Square distances.

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

with \( U, U' \) two independent random variables with \( U[0, 1] \) distribution.

Standard deviation of the square distances

\[ \text{sdev} \left[ \| X^{(i)} - X^{(j)} \|^2 \right] = \sqrt{\sum_{k=1}^{p} \text{var} \left[ (X^{(i)}_k - X^{(j)}_k)^2 \right]} \]
\[ = \sqrt{p \text{var} \left[ (U' - U)^2 \right]} \approx 0.2 \sqrt{p}. \]
Curse 3: lost in high-dimensional spaces

High-dimensional 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

\[ V_p(1) \xrightarrow{p \to \infty} \left( \frac{2\pi e}{p} \right)^{p/2} (p\pi)^{-1/2}. \]
Curse 3 : lost in high-dimensional space

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 $B(x_i, 1)$:

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

<table>
<thead>
<tr>
<th>$p$</th>
<th>20</th>
<th>30</th>
<th>50</th>
<th>100</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>39</td>
<td>45630</td>
<td>5.7 $10^{12}$</td>
<td>42 $10^{39}$</td>
<td>larger than the estimated number of particles in the observable universe</td>
</tr>
</tbody>
</table>
Curse 4: Thin tails concentrate the mass!

Figure: Mass of the standard Gaussian distribution $g_p(x) \, dx$ in the “bell” $B_{p,0.001} = \{ x \in \mathbb{R}^p : g_p(x) \geq 0.001g_p(0) \}$ for increasing dimensions $p$. 
Why?

Volume of a ball: \( V_p(r) = r^p V_p(1) \)

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

**Ball:** \( B_p(0, r) \)

**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!**
Curse 4: Thin tails concentrate the mass!

Where is the Gaussian mass located?

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

$$\frac{1}{\varepsilon} \mathbb{P}[R \leq \|X\| \leq R + \varepsilon] = \frac{1}{\varepsilon} \int_{R \leq \|x\| \leq R + \varepsilon} e^{-\|x\|^2/2} \frac{dx}{(2\pi)^{p/2}}$$

$$= \frac{1}{\varepsilon} \int_{R}^{R+\varepsilon} e^{-r^2/2} r^{p-1} \frac{pV_p(1)}{(2\pi)^{p/2}} dr$$

$$\approx \frac{p}{2^{p/2}\Gamma(1 + p/2)} R^{p-1} \times e^{-R^2/2}.$$

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

Gaussian = uniform?

The Gaussian $\mathcal{N}(0, I_p)$ distribution looks like a uniform distribution on the sphere of radius $\sqrt{p - 1}$!
Curse 5: 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 $\mathbb{P} [ |\mathcal{N}(0, \sigma^2)| \geq 2\sigma ] \leq 0.05$, we can detect the active gene with $X_j$ when

$$|\theta_j| \geq \frac{2\sigma}{\sqrt{n}}$$
Curse 5: 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 (see later)

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

**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
- 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.

Dimension reduction:

- ”unsupervised” (PCA)
- ”supervised”
Principal Component Analysis

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

$$V_d \in \arg\min_{\dim(V) \leq d} \sum_{i=1}^{n} \|X^{(i)} - \text{Proj}_V X^{(i)}\|^2,$$

where $\text{Proj}_V$ is the orthogonal projection matrix onto $V$.

Recap on PCA

Exercise 1.6.4
PCA in action

original image

projected image

original image

projected image

original image

projected image

original image

projected image

MNIST : 1100 scans of each digit. Each scan is a $16 \times 16$ 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
- small sample size

Good feature

Data 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.
Mathematics of high-dimensional statistics

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...)

![Graph showing a scatter plot with a trend line and data points]

C. Giraud (Paris Saclay)
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 $n \ll p$

$\implies$ 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 😞