High-Dimensional Probability and Statistics

MATH/STAT/ECE 888: Topics in Mathematical Data Science

Sebastien Roch (Math+Stat)
UW-Madison

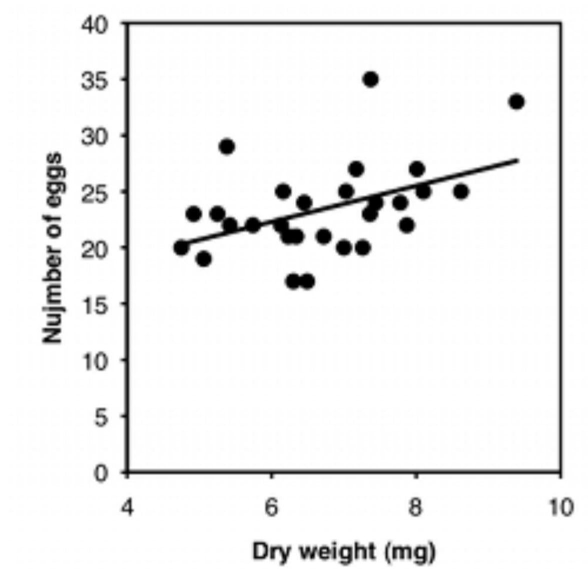
Fall 2021

Lecture 1 (09/08/21)

Course website

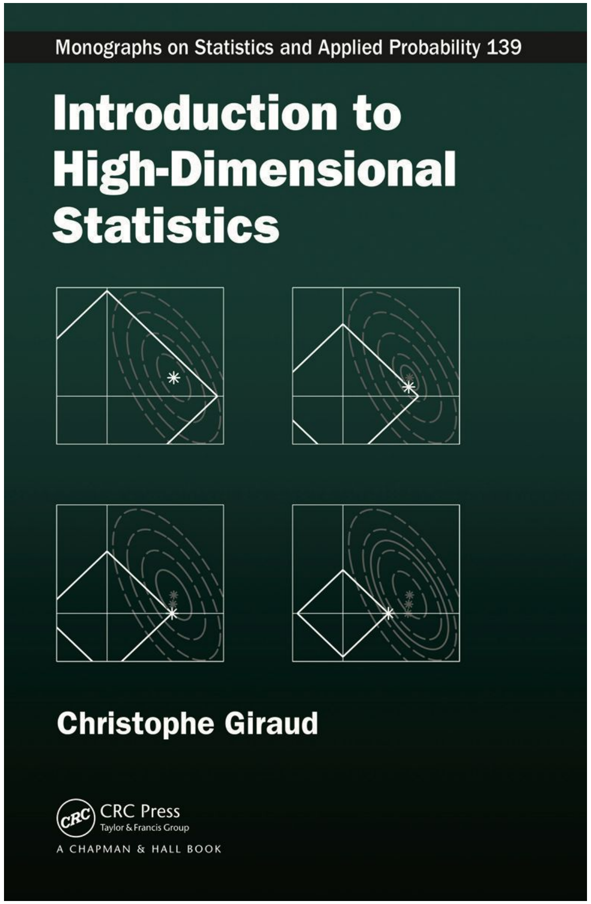
Classical statistics (we'll review some):

- small number p of parameters
- large number n of observations
- investigate performance of estimators as n → ∞ (CLT...)



Graph of number of eggs vs. dry weight in the amphipod *Platorchestia platensis*.
McDonald (1989)

Today’s slides based on Chap 1 (read it!) of Giraud



https://www.imo.universite-paris-saclay.fr/~giraud/Orsay/slides/slidesC1.pdf

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 <u>in general</u> 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 $\mathrm{cov}(X) = \sigma^2 I_p$. 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 : 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 $X^{(i)} \sim \mathcal{U}([0,1]^p)$

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

Curse 2 : locality is lost

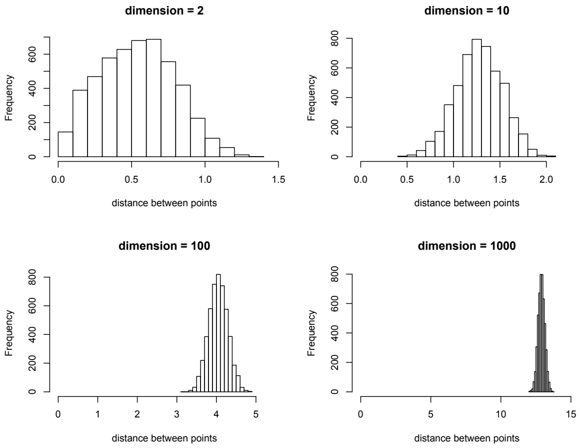


Figure: Histograms of the pairwise-distances between $n = 100$ points sampled uniformly in the hypercube $[0,1]^p$, for $p = 2, 10, 100$ and 1000.

Why?

Square distances.

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

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

Standard deviation of the square distances

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

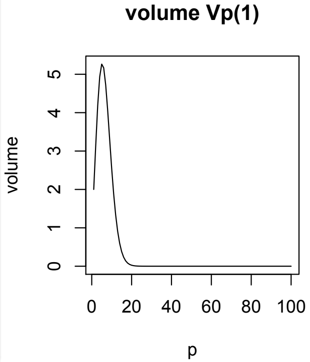
Curse 3 : lost in high-dimensional spaces

High-dimensional 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

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



Curse 3 : lost in high-dimensional space

Which sample size to avoid the lost of locality?

Number n of points $x_1, \dots, x_n$ 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 4: Thin tails concentrate the mass!

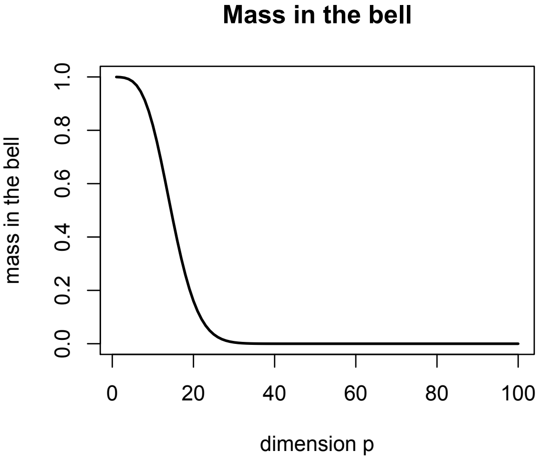


Figure: Mass of the standard Gaussian distribution $g_p(x)\,dx$ in the "bell" $B_{p,0.001} = \{x \in \mathbb{R}^p : g_p(x) \geq 0.001 g_p(0)\}$ for increasing dimensions p.

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!

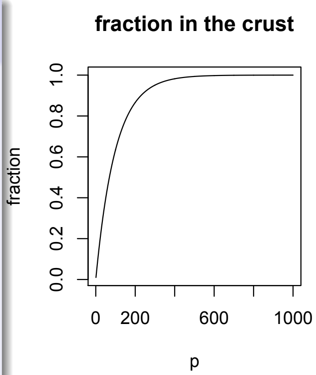
Ball: $B_p(0, r)$

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

The fraction of the volume in the crust

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

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

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

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

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

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

Curse 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

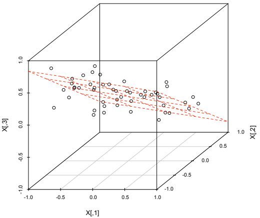
Hopeless?

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"

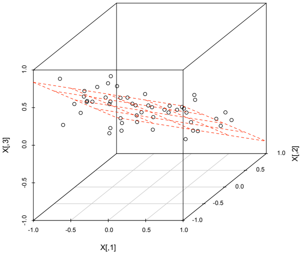


Principal Component Analysis

For any data points $X^{(1)}, \dots, X^{(n)} \in \mathbb{R}^p$ 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.

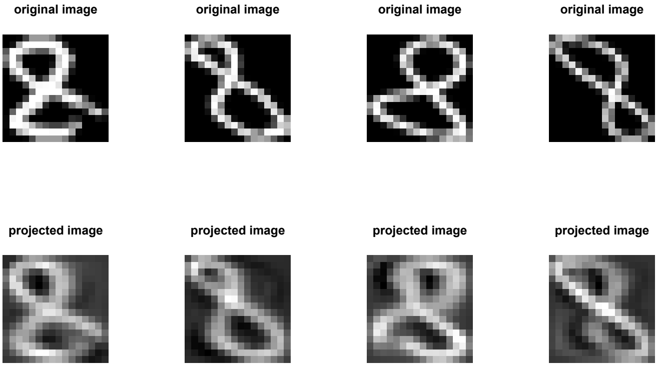


V_2 in dimension $p = 3$.

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

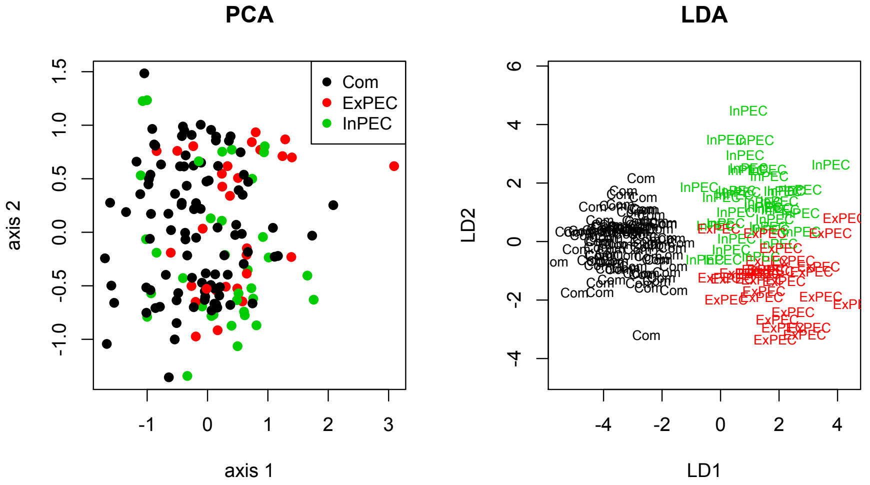


Figure: 55 chemical measurements of 162 strains of *E. coli*.
Left : the data is projected on the plane given by a PCA.
Right : the data is projected on the plane given by a LDA.

Summary

Statistical difficulty

- high-dimensional data
- 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...)

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 ☹