Statistiques en grande dimension

Christophe Giraud[1,2]

(1) Université Paris-Sud
(2) Ecole Polytechnique

M2 DS & ISG

High-dimensional data

Données en grande dimension

- **Données biotech:** mesure des milliers de quantités par "individu".
- **Images :** images médicales, astrophysique, video surveillance, etc. Chaque image est constituées de milliers ou millions de pixels ou voxels.
- **Marketing:** les sites web et les programmes de fidélité collectent de grandes quantités d'information sur les préférences et comportements des clients. Ex: systèmes de recommandation...
- **Business:** exploitation des données internes et externes de l'entreprise devient primordial
- **Crowdsourcing data :** données récoltées online par des volontaires. Ex: eBirds collecte des millions d'observations d'oiseaux en Amérique du Nord

Blessing?

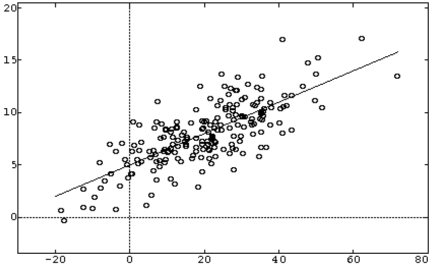
☺ we can sense thousands of variables on each "individual" : potentially we will be able to scan every variables that may influence the phenomenon under study.

☹ 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.

Renversement de point de vue

Cadre statistique classique:

- petit nombre p de paramètres
- grand nombre n d'expériences
- on étudie le comportement asymptotique des estimateurs lorsque $n \to \infty$ (résultats type théorème central limite)



Renversement de point de vue

Cadre statistique classique:

- petit nombre p de paramètres
- grand nombre n d'expériences
- on étudie le comportement asymptotique des estimateurs lorsque $n \to \infty$ (résultats type théorème central limite)

Données actuelles:

- inflation du nombre p de paramètres
- taille d'échantillon réduite: $n \approx p$ ou $n \ll p$

$\Longrightarrow$ penser différemment les statistiques!
(penser $n \to \infty$ ne convient plus)

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 ☺
- 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 ☹

The tools of non-asymptotic statistics (1/3)

Typical tool of asymptotic analysis: CLT

For $f : \mathbb{R}^d \to \mathbb{R}$ and $X_1, \ldots, X_n$ i.i.d. such that $\mathrm{var}(f(X_1)) < +\infty$, when $n \to +\infty$

$$\sqrt{\frac{n}{\mathrm{var}(f(X_1))}} \left(\frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}\left[f(X_1)\right] \right) \xrightarrow{\mathrm{d}} Z, \quad \text{with } Z \sim \mathcal{N}(0,1).$$

Ex: If f is L-Lipschitz, and $\mathrm{var}(X_i) = \sigma^2$, we have

$$\mathrm{var}(f(X_1)) = \frac{1}{2}\mathbb{E}\left[(f(X_1) - f(X_2))^2\right] \leq \frac{L^2}{2}\mathbb{E}\left[(X_1 - X_2)^2\right] = L^2\sigma^2,$$

so,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} f(X_i) \geq \mathbb{E}\left[f(X_1)\right] + \frac{L\sigma}{\sqrt{n}} x \right) \leq \mathbb{P}(Z \geq x) \leq e^{-x^2/2}$$

The tools of non-asymptotic statistics (2/3)

Concentration inequalities provide some non asymptotic versions of such results.

Gaussian concentration inequality

If $X_1, \ldots, X_n$ are i.i.d. with $\mathcal{N}(0, \sigma^2)$ Gaussian distribution and $F : \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz then

$$F(X_1, \ldots, X_n) \leq \mathbb{E}\left[F(X_1, \ldots, X_n)\right] + L\sigma\sqrt{2\xi}, \quad \text{where } \xi \sim \mathcal{E}xp(1)$$

Ex: If $f : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz, the Gaussian concentration inequality ensures for any $x > 0$ and $n \geq 1$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) \geq \mathbb{E}\left[f(X_1)\right] + \frac{L\sigma}{\sqrt{n}}\, x\right) \leq e^{-x^2/2}.$$

Proof:

The Cauchy–Schwartz inequality gives

$$\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \frac{1}{n}\sum_{i=1}^{n} f(Y_i)\right| \leq \frac{L}{n}\sum_{i=1}^{n}|X_i - Y_i| \leq \frac{L}{\sqrt{n}}\sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2}\ ,$$

so $F(X_1,\ldots,X_n) = n^{-1}\sum_{i=1}^{n} f(X_i)$ is $(n^{-1/2}L)$-Lipschitz.

Hence

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}\left[f(X_1)\right] \geq \frac{L\sigma}{\sqrt{n}}x\right) \leq \mathbb{P}\left(\sqrt{2\xi} \geq x\right) = e^{-x^2/2}.$$

The tools of non-asymptotic statistics (3/3)

McDiarmid concentration inequality

Let $F : \mathcal{X}^n \to \mathbb{R}$ be a measurable function, such that

$$\left|F(x_1, \ldots, x_i', \ldots, x_n) - F(x_1, \ldots, x_i, \ldots, x_n)\right| \le \delta_i, \quad \text{for all} \quad x_1, \ldots, x_n, x_i' \in \mathcal{X},$$

for all $i = 1, \ldots, n$. Then, for any independent random variables $X_1, \ldots, X_n \in \mathcal{X}$, we have

$$F(X_1, \ldots, X_n) \le \mathbb{E}\left[F(X_1, \ldots, X_n)\right] + \sqrt{\frac{\delta_1^2 + \ldots + \delta_n^2}{2}\, \xi}\ .$$

Very useful to assess the random fluctuations in supervised classification.

Fléau de la dimension

Curse 1 : fluctuations cumulate

Exemple : linear regression $Y = \mathbf{X}\beta^* + \varepsilon$ with $\mathbf{cov}(\varepsilon) = \sigma^2 I_n$. The Least-Square estimator $\widehat{\beta} \in \mathrm{argmin}_{\beta \in \mathbb{R}^p} \|Y - \mathbf{X}\beta\|^2$ has a risk

$$\mathbb{E}\left[\|\widehat{\beta} - \beta^*\|^2\right] = \mathrm{Tr}\left((\mathbf{X}^T\mathbf{X})^{-1}\right)\sigma^2.$$

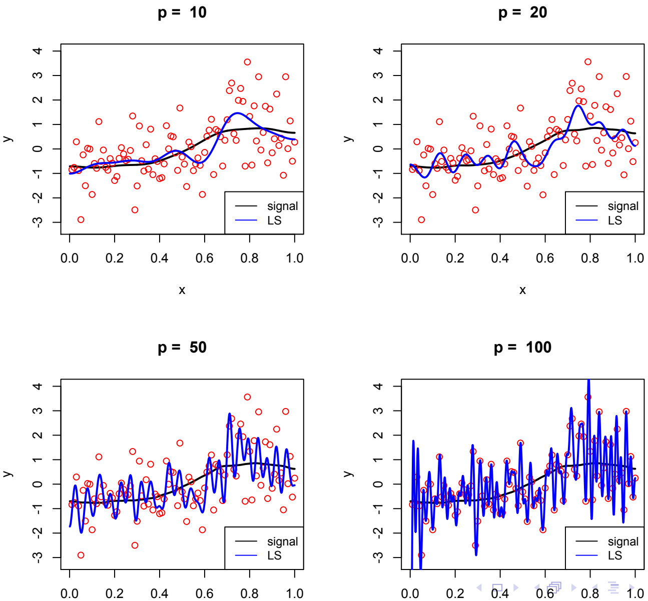
Illustration :

$$Y_i = \sum_{j=1}^{p} \beta_j^* \cos(\pi j i/n) + \varepsilon_i = f_{\beta^*}(i/n) + \varepsilon_i, \quad \text{for } i = 1, \ldots, n,$$

with

- $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d with $\mathcal{N}(0, 1)$ distribution
- β_j^* independent with $\mathcal{N}(0, j^{-4})$ distribution

Curse 1 : fluctuations cumulate



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.

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

Curse 2 : locality is lost

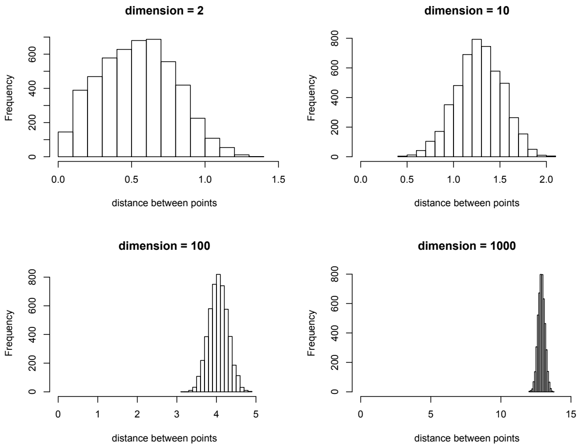


Figure: Histograms of the pairwise-distances between $n = 100$ points sampled uniformly in the hypercube $[0, 1]^p$, for $p = 2, 10, 100$ and 1000.

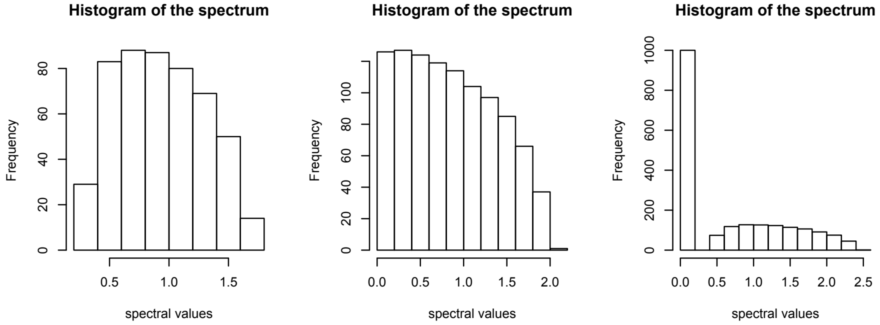
Curse 2 : locality is lost

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

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

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: empirical covariance fails



Histogram of the spectral values of the empirical covariance matrix $\widehat{\Sigma}$ of $\Sigma = Id$, with $n = 1000$ and $p = n/2$ (left), $p = n$ (center), $p = 2n$ (right).

Curse 4: Thin tails concentrate the mass!

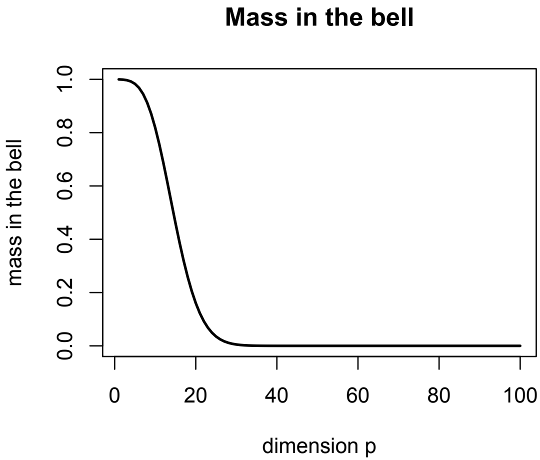


Figure: Mass of the standard Gaussian distribution $g_p(x)\,dx$ in the "bell" $B_{p,0.001} = \{x \in \mathbb{R}^p : g_p(x) \geq 0.001 g_p(0)\}$ for increasing dimensions p.

Some other curses

- Curse 5 : an accumulation of rare events may not be rare (false discoveries, etc)
- Curse 6 : algorithmic complexity must remain low

Low-dimensional structures in high-dimensional data

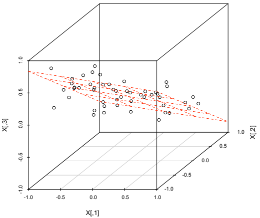
Hopeless?

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)
- "estimation-oriented"

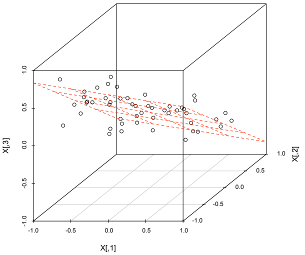


Principal Component Analysis

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 $\mathbb{R}^p$

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

where Proj_V is the orthogonal projection matrix onto V.

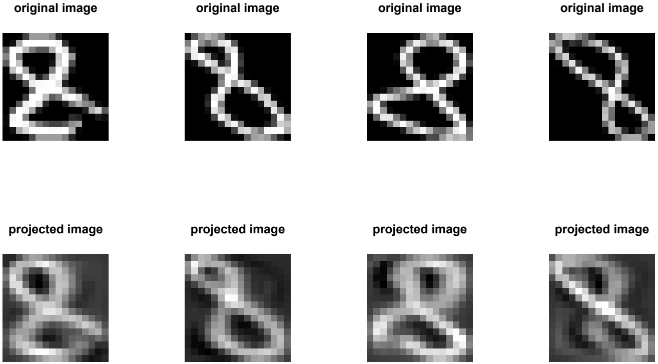


V_2 in dimension $p = 3$.

To do

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.

"Estimation-oriented" dimension reduction

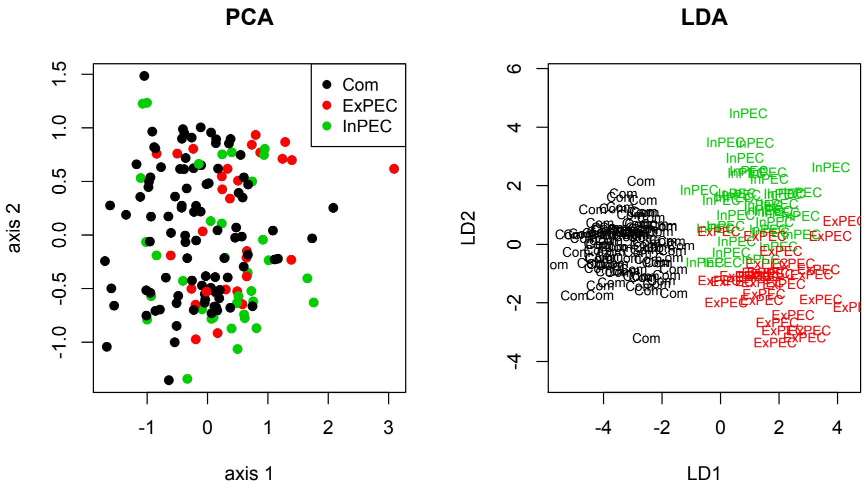


Figure: 55 chemical measurements of 162 strains of *E. coli*.
Left : the data is projected on the plane given by a PCA.
Right : the data is projected on the plane given by a LDA.

Résumé

Difficulté statistique

- données de très grande dimension
- peu de répétitions

Pour nous aider

Données issues d'un vaste système dynamique (plus ou moins stochastique)

- existence de structures de faible dimension "effective"
- existence de modèles théoriques

La voie du succès

Trouver, à partir des données, ces structures "cachées" pour pouvoir les exploiter.

La voie du succès

Trouver, à partir des données, les structures cachées pour pouvoir les exploiter.

Exemples de structures

Regression Model

Regression model

$$Y_i = f(x^{(i)}) + \varepsilon_i, \quad i = 1, \ldots, n \quad \text{with}$$

- $f : \mathbb{R}^p \to \mathbb{R}$
- $\mathbb{E}[\varepsilon_i] = 0$

Vectorial representation

The observations can be summarized in a vector form

$$Y = f^* + \varepsilon \in \mathbb{R}^n$$

with $f_i^* = f(x^{(i)})$, $i = 1, \ldots, n$.

Low-dimensional x

Example 1: sparse piecewise constant regression

It corresponds to the case where $f : \mathbb{R} \to \mathbb{R}$ is piecewise constant with a small number of jumps.

This situation appears for example for CGH profiling.

Low-dimensional x

Example 2: sparse basis/frame expansion

We estimate $f : \mathbb{R} \to \mathbb{R}$ by expanding it on a basis or frame $\{\varphi_j\}_{j\in\mathcal{J}}$

$$f(x) = \sum_{j\in\mathcal{J}} c_j \varphi_j(x),$$

with a small number of nonzero c_j. Typical examples of basis are Fourier basis, splines, wavelets, etc.

The most simple example is the piecewise linear decomposition where $\varphi_j(x) = (x - z_j)_+$ where $z_1 < z_2 < \ldots$ and $(x)_+ = \max(x, 0)$.

High-dimensional x

Example 3: sparse linear regression

It corresponds to the case where f is linear: $f(x) = \langle \beta, x \rangle$ where $\beta \in \mathbb{R}^p$ has only a few nonzero coordinates.

This model can be too rough to model the data.

Example: relationship between some phenotypes and some protein abundances.

- only a small number of proteins influence a given phenotype,
- but the relationship between these proteins and the phenotype is unlikely to be linear.

High-dimensional x

Example 4: sparse additive model

In the sparse additive model, we expect that $f(x) = \sum_k f_k(x_k)$ with most of the f_k equal to 0.

If we expand each function f_k on a frame or basis $\{\varphi_j\}_{j \in \mathcal{J}_k}$ we obtain the decomposition

$$f(x) = \sum_{k=1}^{p} \sum_{j \in \mathcal{J}_k} c_{j,k} \varphi_j(x_k),$$

where most of the vectors $\{c_{j,k}\}_{j \in J_k}$ are zero.

Such a model can be hard to fit from a small sample since it requires to estimate a relatively large number of non-zero $c_{j,k}$.

High-dimensional x

In some cases the basis expansion of f_k can be sparse itself.

Example 5: sparse additive piecewise linear regression

The sparse additive piecewise linear model, is a sparse additive model $f(x) = \sum_k f_k(x_k)$ with sparse piecewise linear functions f_k. We then have two levels of sparsity :

1. most of the f_k are equal to 0,
2. the nonzero f_k have a sparse expansion in the following representation

$$f_k(x_k) = \sum_{j \in \mathcal{J}_k} c_{j,k} \, (x_k - z_{j,k})_+$$

In other words, the matrix $c = [c_{j,k}]$ of the sparse additive model has only a few nonzero columns, and this nonzero columns are sparse.

Reduction to a structured linear model

Reduction to a structured linear model

In all the above examples, we have a linear representation

$$f_i^* = \langle \alpha, \psi_i \rangle \quad \text{for } i = 1, \ldots, n,$$

with a structured α.

Examples (continued)

Representation $f_i^* = \langle \alpha, \psi_i \rangle$

- Sparse piecewise constant regression: $\psi_i = e_i$ with $\{e_1, \ldots, e_n\}$ the canonical basis of $\mathbb{R}^n$ and $\alpha = f^*$ is piecewise constant.
- Sparse basis expansion: $\psi_i = [\varphi_j(x^{(i)})]_{j \in \mathcal{J}}$ and $\alpha = c$.
- Sparse linear regression: $\psi_i = x^{(i)}$ and $\alpha = \beta$.
- Sparse additive models: $\psi_i = [\varphi_j([x_k^{(i)}])]_{\substack{k=1,\ldots,p \\ j \in \mathcal{J}_k}}$ and $\alpha = [c_{j,k}]_{\substack{k=1,\ldots,p \\ j \in \mathcal{J}_k}}$.