Curse of dimensionality

Christophe Giraud

Université Paris-Sud

M2 DS

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 de façon opportunistes. Ex: eBirds collecte des millions d'observations d'oiseaux en Amérique du Nord, les hôpitaux collectent des données médicales sur leurs patients, etc.

Blessing?

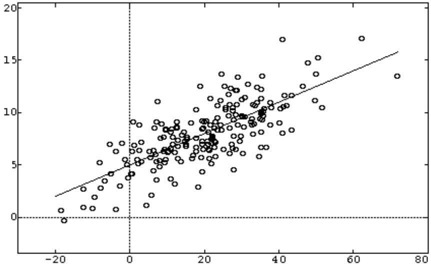
☺ we can sense thousands of variables on each "individual" : potentially we will be able to scan every variables that may influence the phenomenon under study.

☹ 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.

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 ☹
- and the garanties are less tight ☹

Typical quantities involved

For $f : \mathbb{R}^d \to \mathbb{R}$ and $X_1, \ldots, X_n$ i.i.d.

Empirical processes

$$R(f) = \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - \mathbb{E}[f(X_i)])$$

Suprema of Empirical Processes

$$R(\mathcal{F}) = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - \mathbb{E}[f(X_i)])$$

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

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

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

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

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

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

with $\xi_F \sim \mathcal{E}xp(1)$.

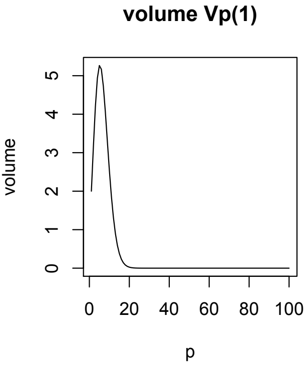
Very useful to assess the random fluctuations in supervised classification.

High-dimensional spaces

The Strange Geometry of High-Dimensional Spaces (I)

High-dimensional balls have a vanishing volume!

$$V_p(r) = \text{volume of a ball of radius } r \text{ in dimension } p \underset{p\to\infty}{\sim} \left(\frac{2\pi e r^2}{p}\right)^{p/2} (p\pi)^{-1/2}.$$



The Strange Geometry of High-Dimensional Spaces (II)

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

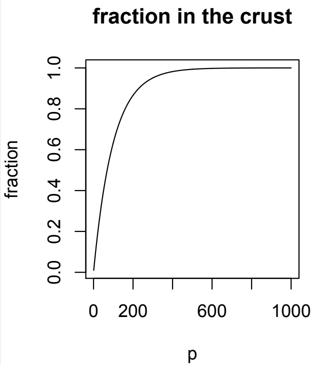
Ball: $B_p(0, r)$

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!



Take Home Message

Forget your low-dimensional intuitions in high-dimensions!

Curse of dimensionality

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

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

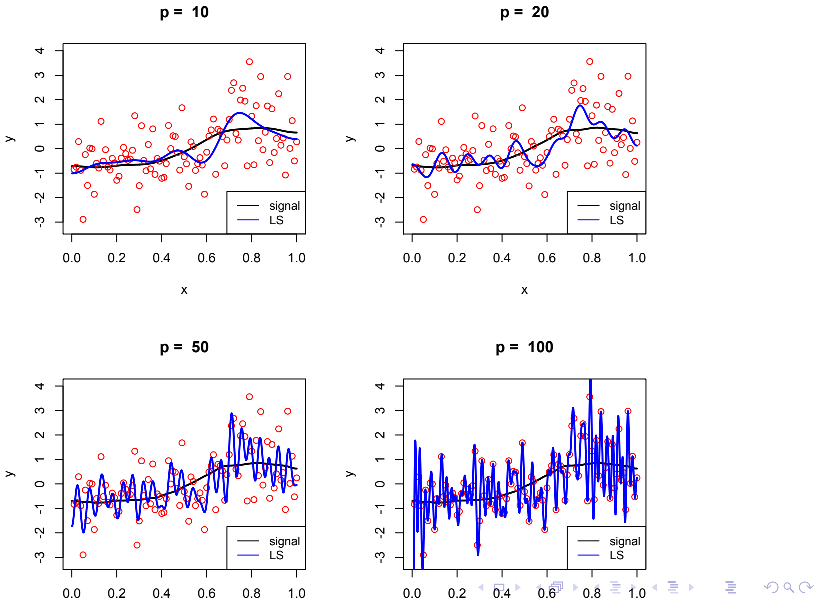
Illustration :

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

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

Canadian high school graduate earnings.

Curse 2 : locality is lost

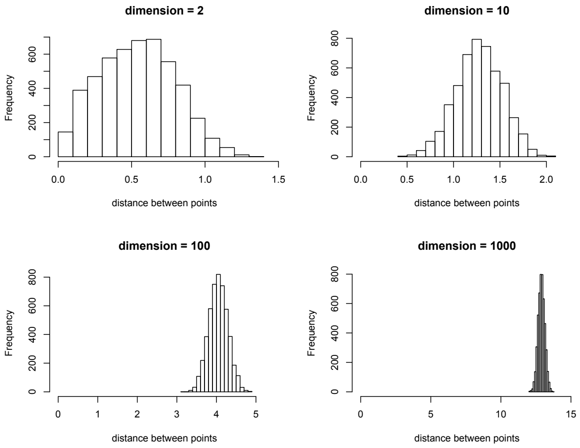


Figure: Histograms of the pairwise-distances between $n = 100$ points sampled uniformly in the hypercube $[0, 1]^p$, for $p = 2, 10, 100$ and 1000.

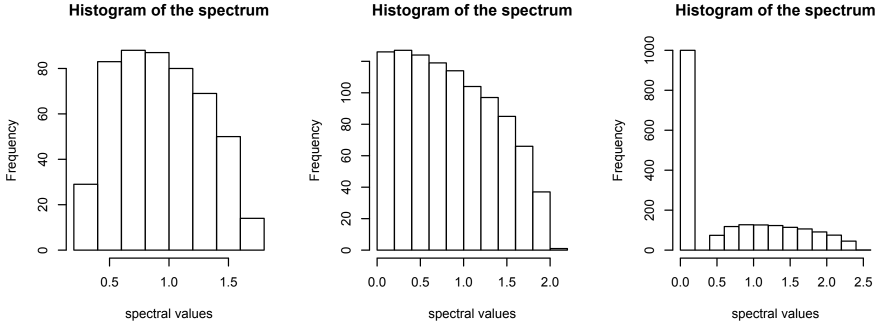
Curse 2 : locality is lost

Number n of points $x_1, \dots, x_n$ 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!

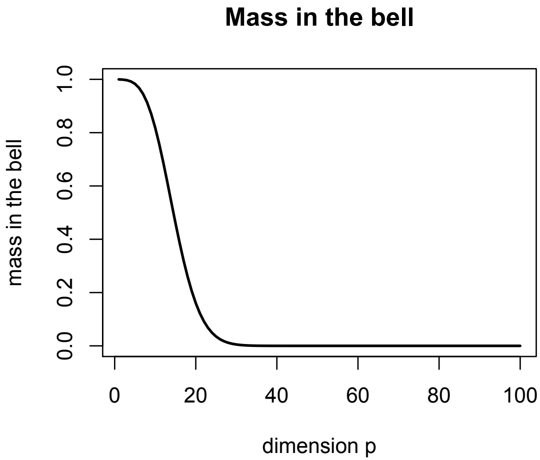


Figure: Mass of the standard Gaussian distribution $g_p(x)\,dx$ in the "bell" $B_{p,0.001} = \{x \in \mathbb{R}^p : g_p(x) \geq 0.001 g_p(0)\}$ for increasing dimensions p.

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.

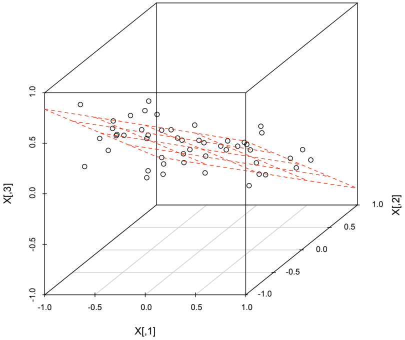
La voie du succès

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

Dimension reduction

Type of dimension reduction

- "unsupervised" (PCA)
- "estimation-oriented"



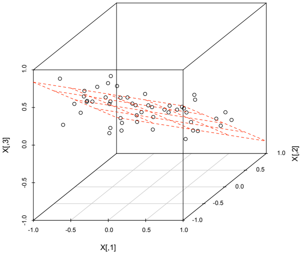
Unsupervised dimension reduction

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 $p = 3$.

Remark: since V is a vector space, always start by centering the data $X^{(i)} \leftarrow X^{(i)} - \frac{1}{n} \sum_{j=1}^{n} X^{(j)}$

PCA = truncated SVD

We set

$$\mathbf{X} = \begin{bmatrix} (X^{(1)})^T \\ \vdots \\ (X^{(n)})^T \end{bmatrix}$$

PCA outcome

Let $\mathbf{X} = \sum_k \sigma_k u_k v_k^T$ be a SVD of $\mathbf{X}$.
Then the matrix of projected data is

$$\begin{bmatrix} (\mathrm{Proj}_{V_d} X^{(1)})^T \\ \vdots \\ (\mathrm{Proj}_{V_d} X^{(n)})^T \end{bmatrix} = \sum_{k=1}^{d} \sigma_k u_k v_k^T$$

PCA and covariance matrix

Principal Vectors

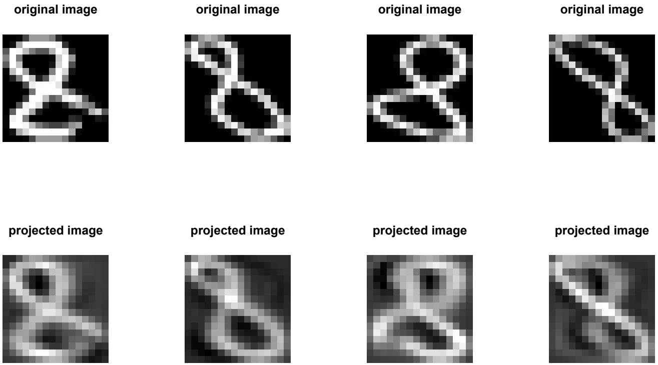
The space V_d is spanned by the k eigenvectors associated to the k largest eigenvalues of the empirical covariance matrix

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X^{(i)}(X^{(i)})^T.$$

Hence

$$\hat{v}_1 \in \operatorname*{argmax}_{\|v\|=1} v^T \widehat{\Sigma} v.$$

PCA in action



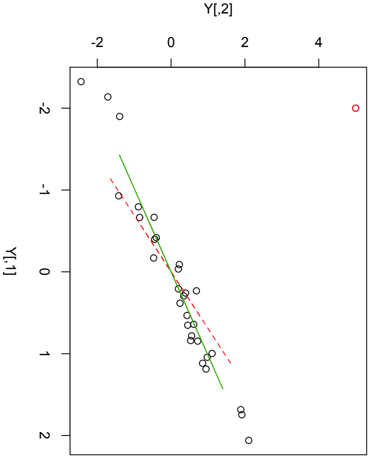
MNIST : 1100 scans of each digit. Each scan is a 16×16 image which is encoded by a vector in $\mathbb{R}^{256}$. The original images are displayed in the first row, their projection onto 10 first principal axes in the second row.

Yet some weakness

Weakness

1. not-robust to outliers
2. fail when $p \approx n$ or $p \gg n$

PCA is not robust to outliers

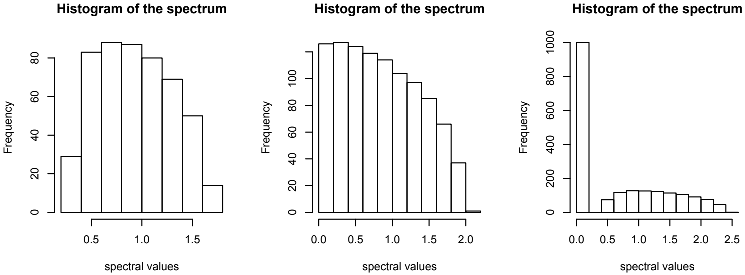


A single error in measurements can strongly impact the PCA.

Outliers nature

- heavy-tailed distribution
- error in data (e.g. gross measurement error)

PCA is not robust to high-dimensions



The empirical covariance matrix is not reliable when is p/n is not small.

References

Curse of dimensionality

- Introduction to high-dimensional statistics, chapter 1.

Robust PCA

- E. Candès, X. Li, Y. Ma, J. Wright. *Robust principal component analysis?* (2011) J. ACM

Sparse PCA

- T. Wang, Q. Berthet, R. Samworth. *Statistical and computational trade-offs in estimation of sparse principal components.* Ann. Stat. (2016)
- V. Vu, J. Lei. *Minimax sparse principal subspace estimation in high-dimensions*. Ann. Stat. (2013)