The curse of dimensionality

Julie Delon

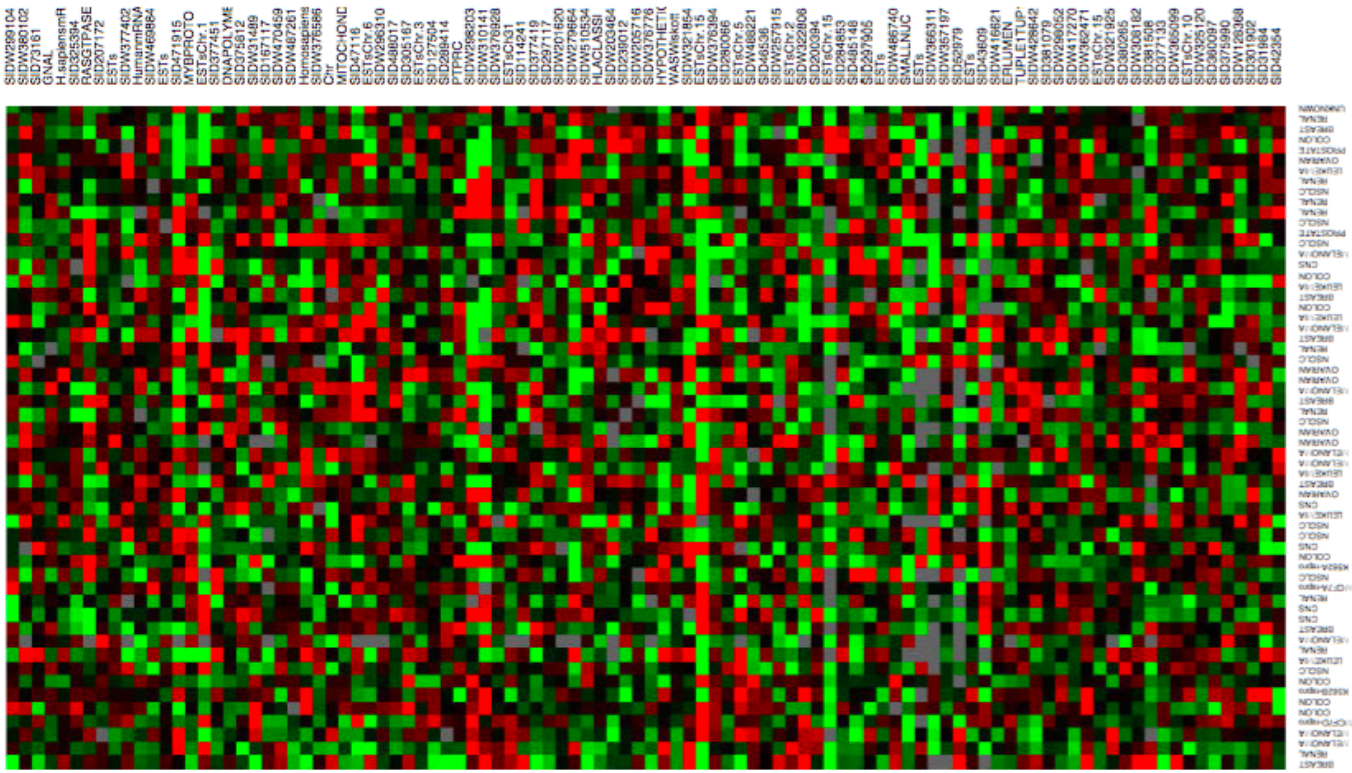
Laboratoire MAP5, UMR CNRS 8145
Université Paris Descartes

up5.fr/delon

Introduction

Modern data are often high dimensional.

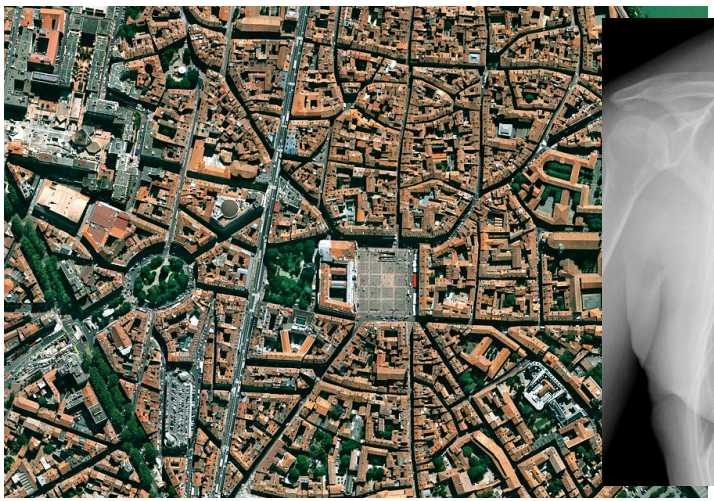
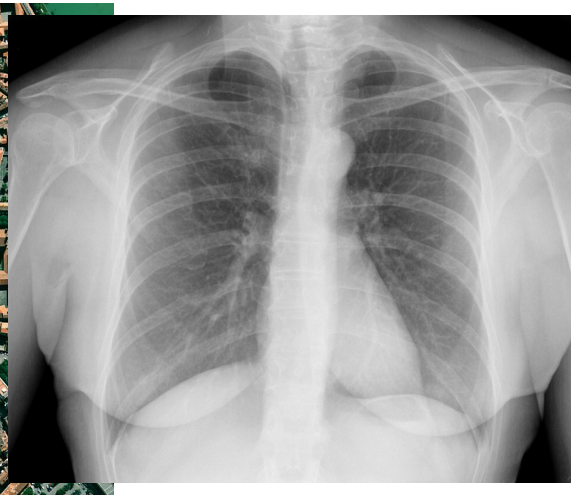
- computational biology: DNA, few observations and huge number of variables ;



Introduction

- images or videos: an image from a digital camera has millions of pixels, 1h of video contains more than 130000 images



Introduction

- data coming from consumer preferences: *Netflix* for instance owns a huge (but sparse) database of ratings given by millions of users on thousands of movies or TV shows.

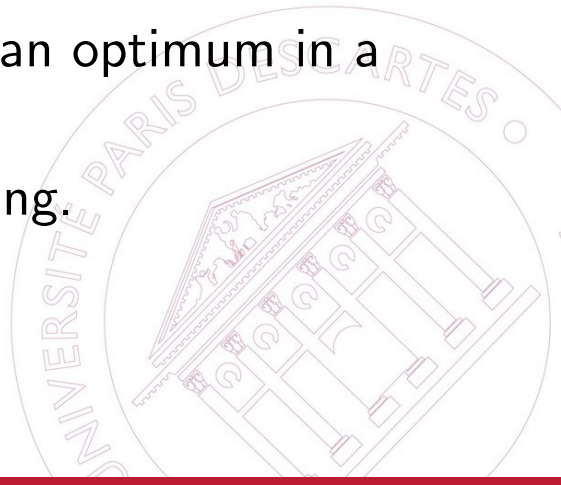
The curse of dimensionality

The curse of dimensionality:

- this term was first used by R. Bellman in the introduction of his book "Dynamic programming" in 1957:

  All [problems due to high dimension] *may be subsumed under the heading "the curse of dimensionality". Since this is a curse,* [...], *there is no need to feel discouraged about the possibility of obtaining significant results despite it.*

- he used this term to talk about the difficulties to find an optimum in a high-dimensional space using an exhaustive search,
- in order to promote dynamic approaches in programming.



Outline

In high dimensional spaces, nobody can hear you scream

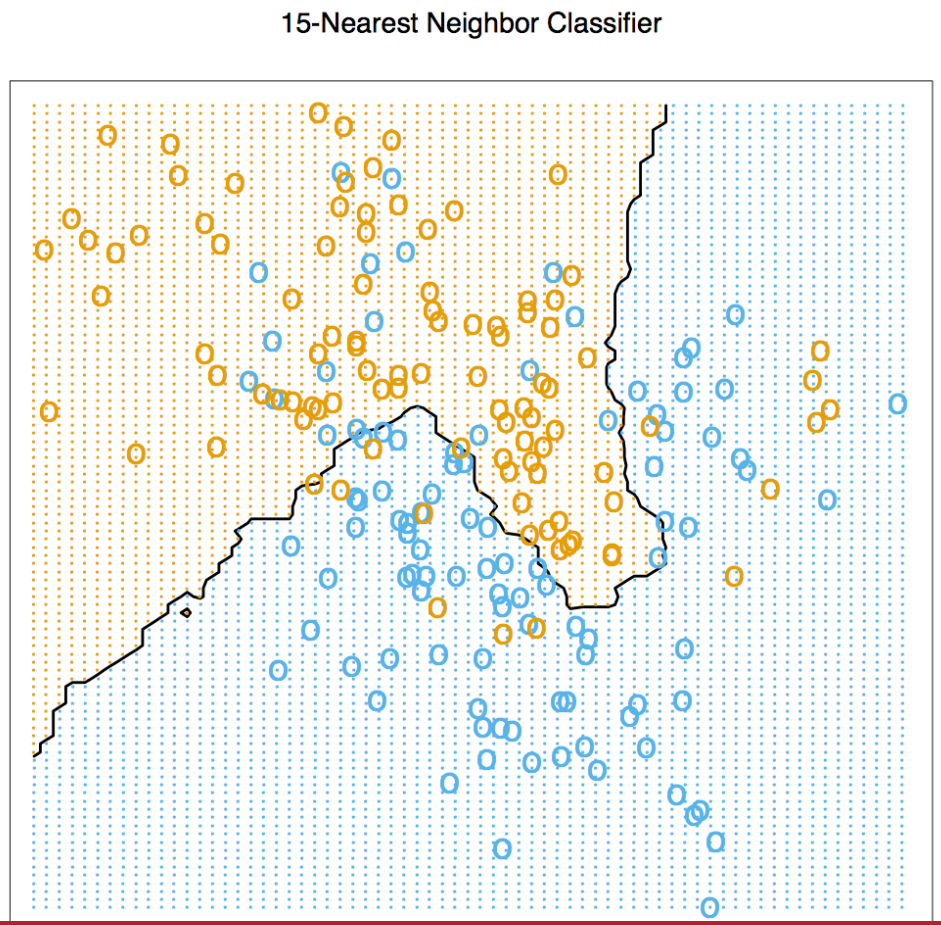
Concentration phenomena

Surprising asymptotic properties for covariance matrices

Nearest neighbors and neighborhoods in estimation

Supervised classification or regression often rely on local averages:

- Classification : you know the classes of n points from your learning database, you can classify a new point x by computing the most represented class in the neighborhood of x.

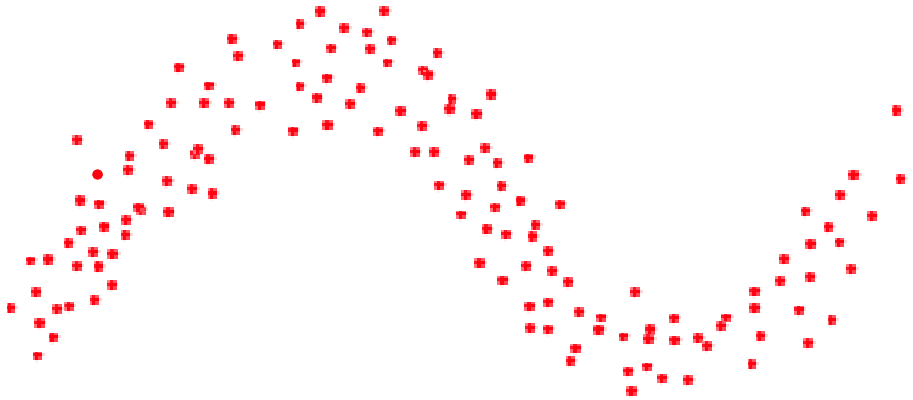


Nearest neighbors and neighborhoods in estimation

- Regression : you observe n i.i.d observations (x^i, y^i) from the model

$$y^i = f(x^i) + \epsilon_i,$$

and you want to estimate f. If you assume f is smooth, a simple solution consists in estimating $f(x)$ as the average of all y_i corresponding to the k nearest neighbors x_i of x.

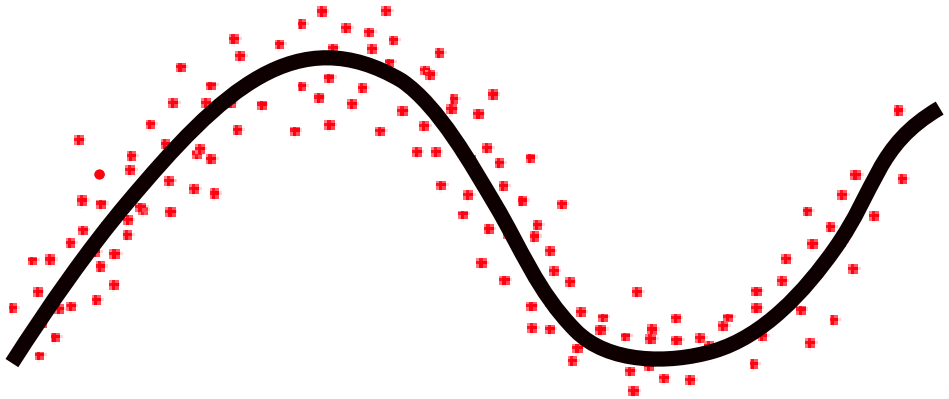


Nearest neighbors and neighborhoods in estimation

- Regression : you observe n i.i.d observations (x^i, y^i) from the model

$$y^i = f(x^i) + \epsilon_i,$$

and you want to estimate f. If you assume f is smooth, a simple solution consists in estimating $f(x)$ as the average of all y_i corresponding to the k nearest neighbors x_i of x.

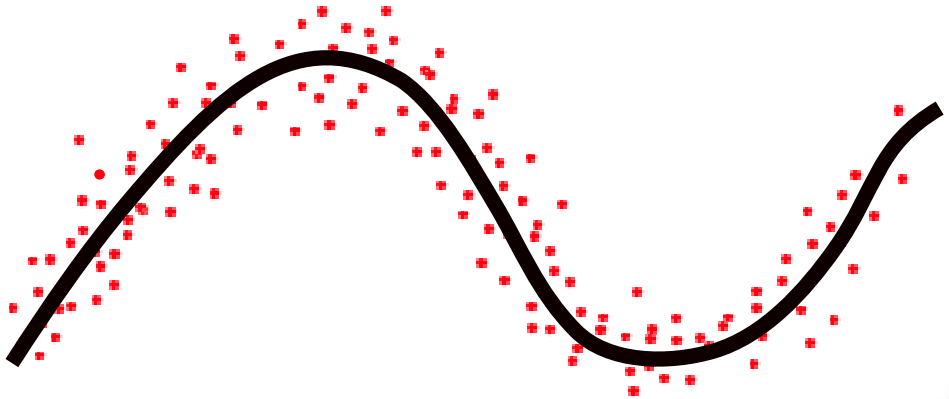


Nearest neighbors and neighborhoods in estimation

- Regression : you observe n i.i.d observations (x^i, y^i) from the model

$$y^i = f(x^i) + \epsilon_i,$$

and you want to estimate f. If you assume f is smooth, a simple solution consists in estimating $f(x)$ as the average of all y_i corresponding to the k nearest neighbors x_i of x.



Makes sense in small dimension. Unfortunately, not so much when the dimension p increases...

High dimensional spaces are empty

Assume your data lives in $[0,1]^p$. To capture a neighborhood which represents a fraction s of the hypercube volume, you need the edge length to be $s^{1/p}$

High dimensional spaces are empty

Assume your data lives in $[0,1]^p$. To capture a neighborhood which represents a fraction s of the hypercube volume, you need the edge length to be $s^{1/p}$

- $s = 0.1$, $p = 10$, $s^{1/p} = 0.63$

High dimensional spaces are empty

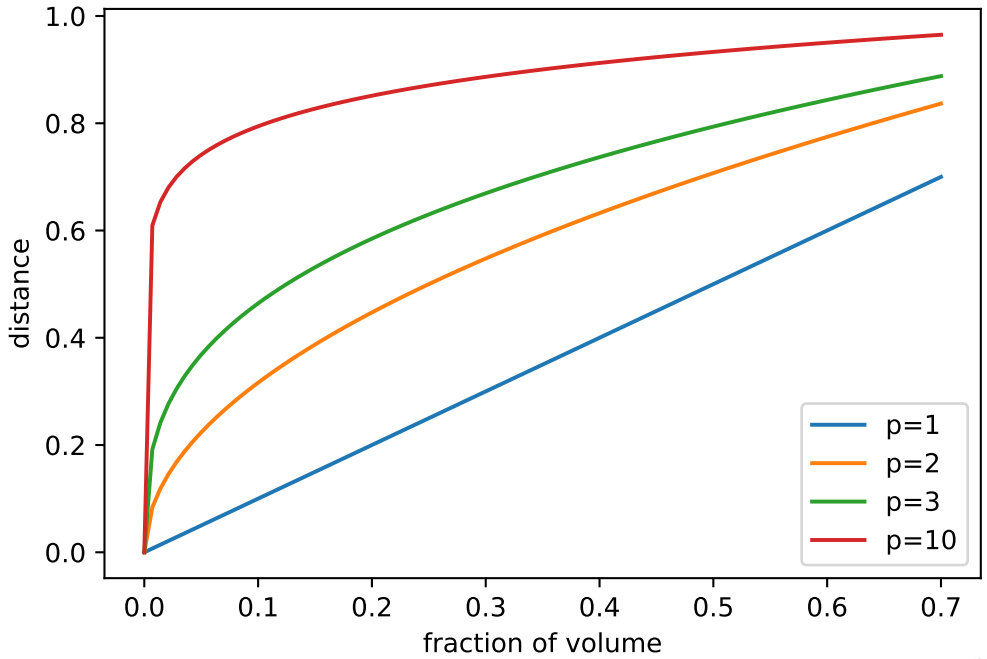
Assume your data lives in $[0,1]^p$. To capture a neighborhood which represents a fraction s of the hypercube volume, you need the edge length to be $s^{1/p}$

- $s = 0.1$, $p = 10$, $s^{1/p} = 0.63$
- $s = 0.01$, $p = 10$, $s^{1/p} = 0.8$

High dimensional spaces are empty

Assume your data lives in $[0, 1]^p$. To capture a neighborhood which represents a fraction s of the hypercube volume, you need the edge length to be $s^{1/p}$

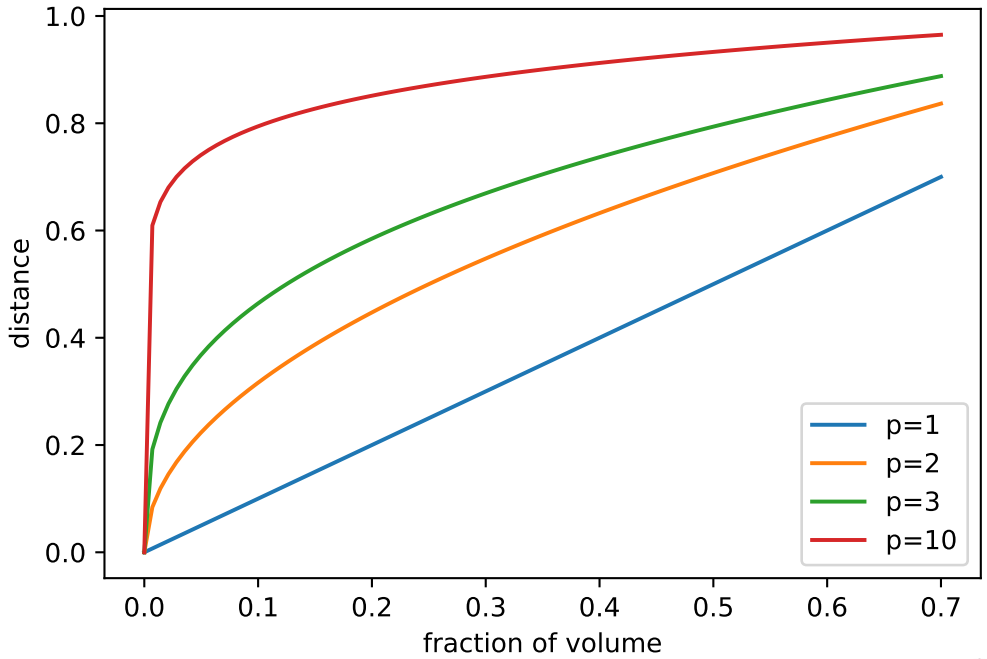
- $s = 0.1$, $p = 10$, $s^{1/p} = 0.63$
- $s = 0.01$, $p = 10$, $s^{1/p} = 0.8$



High dimensional spaces are empty

Assume your data lives in $[0,1]^p$. To capture a neighborhood which represents a fraction s of the hypercube volume, you need the edge length to be $s^{1/p}$

- $s = 0.1,\ p = 10,\ s^{1/p} = 0.63$
- $s = 0.01,\ p = 10,\ s^{1/p} = 0.8$



Neighborhoods are no longer local

High dimensional spaces are empty

The volume of an hypercube with an edge length of $r = 0.1$ is $0.1^p \rightarrow$ when p grows, it quickly becomes so small that the probability to capture points from your database becomes very very small...

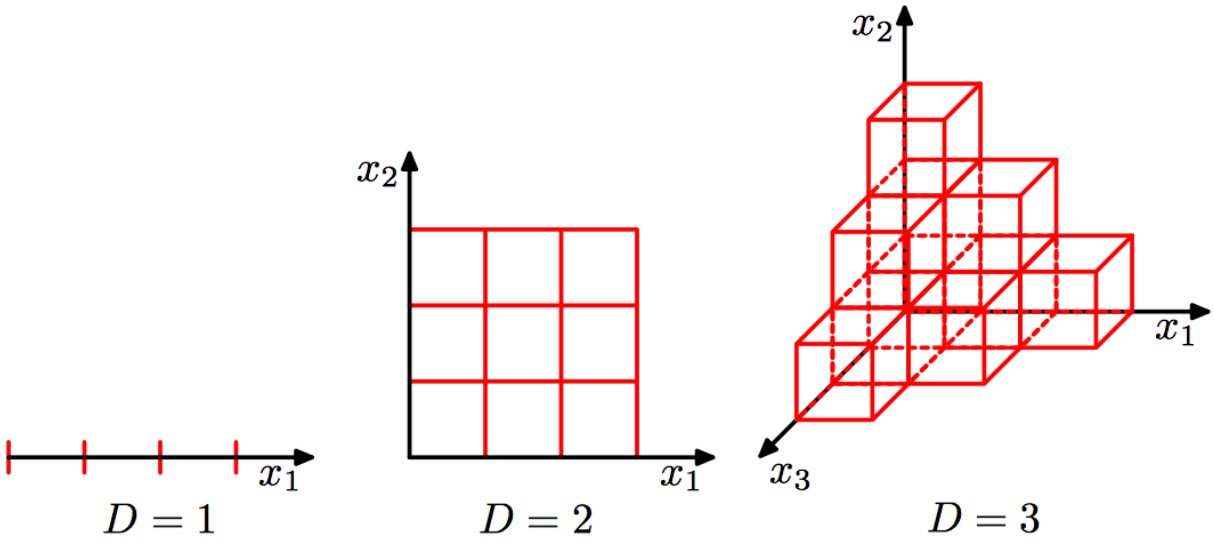
Points in high dimensional spaces are isolated

High dimensional spaces are empty

The volume of an hypercube with an edge length of $r = 0.1$ is $0.1^p \rightarrow$ when p grows, it quickly becomes so small that the probability to capture points from your database becomes very very small...

Points in high dimensional spaces are isolated

To overcome this limitation, you need a number of sample which grows exponentially with p...



Nearest neighbors

X, Y two independent variables, with uniform distribution on $[0,1]^p$. The mean square distance $\|X-Y\|^2$ satisfies

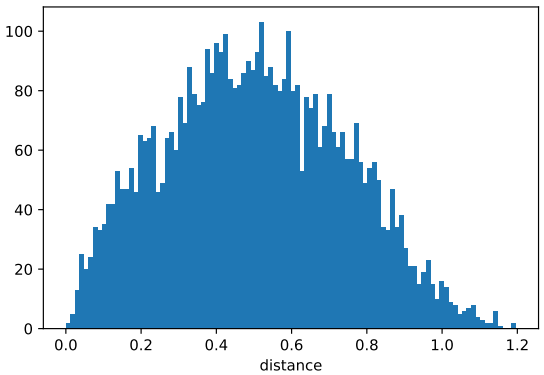
$$\mathbb{E}[\|X-Y\|^2] = p/6 \text{ and } Std[\|X-Y\|^2] \simeq 0.2\sqrt{p}.$$

Nearest neighbors

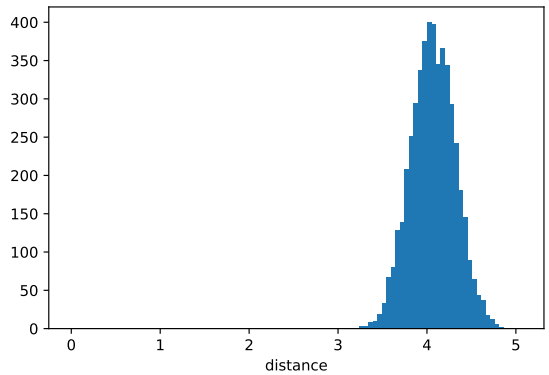
X, Y two independent variables, with uniform distribution on $[0,1]^p$. The mean square distance $\|X-Y\|^2$ satisfies

$$\mathbb{E}[\|X-Y\|^2] = p/6 \text{ and } Std[\|X-Y\|^2] \simeq 0.2\sqrt{p}.$$

$p = 2$



$p = 100$



$p = 1000$

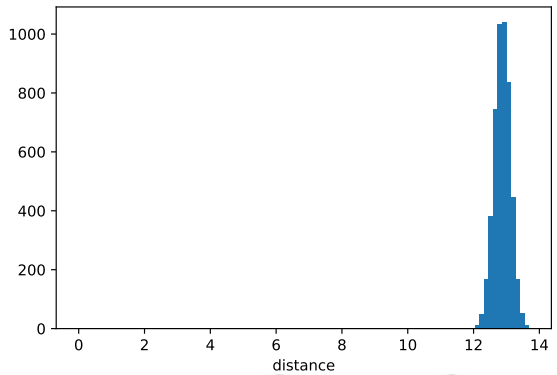


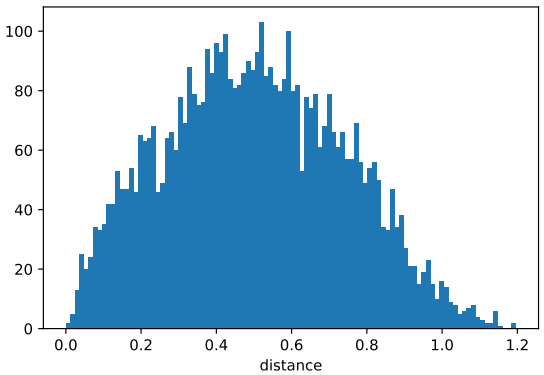
Figure: Histograms of pairwise-distances between $n = 100$ points sampled uniformly in the hypercube $[0,1]^p$

Nearest neighbors

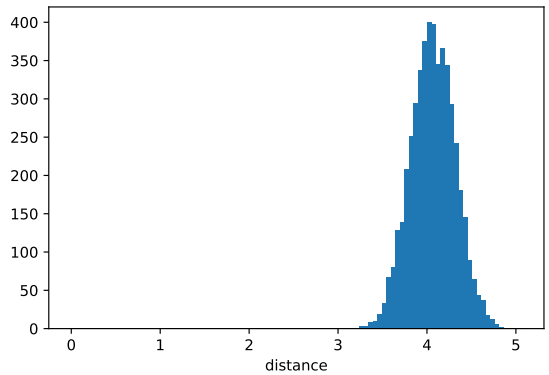
X, Y two independent variables, with uniform distribution on $[0,1]^p$. The mean square distance $\|X-Y\|^2$ satisfies

$$\mathbb{E}[\|X-Y\|^2] = p/6 \text{ and } Std[\|X-Y\|^2] \simeq 0.2\sqrt{p}.$$

$p = 2$



$p = 100$



$p = 1000$

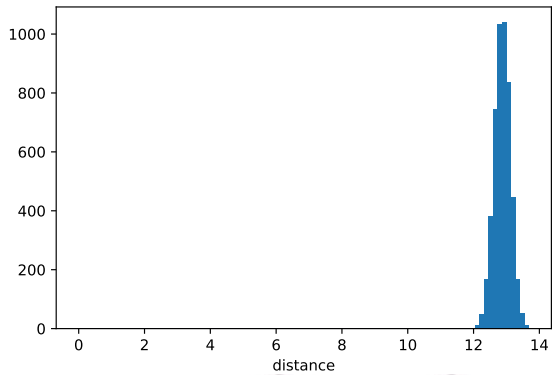


Figure: Histograms of pairwise-distances between $n = 100$ points sampled uniformly in the hypercube $[0,1]^p$

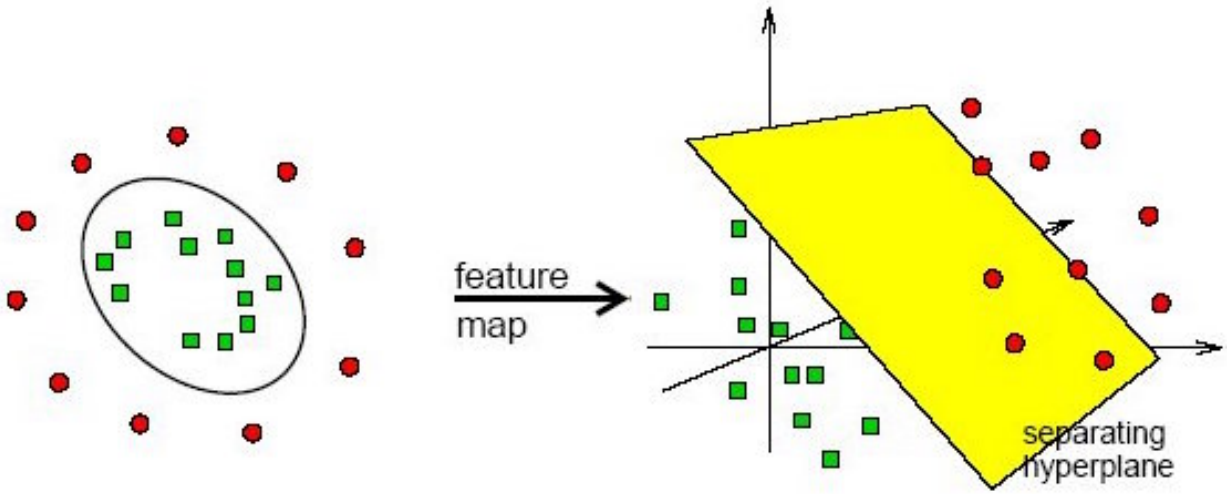
The notion of nearest neighbors vanishes.

Classification in high dimension

- since high-dimensional spaces are almost empty,
- it should be easier to separate groups in high-dimensional space with an adapted classifier,

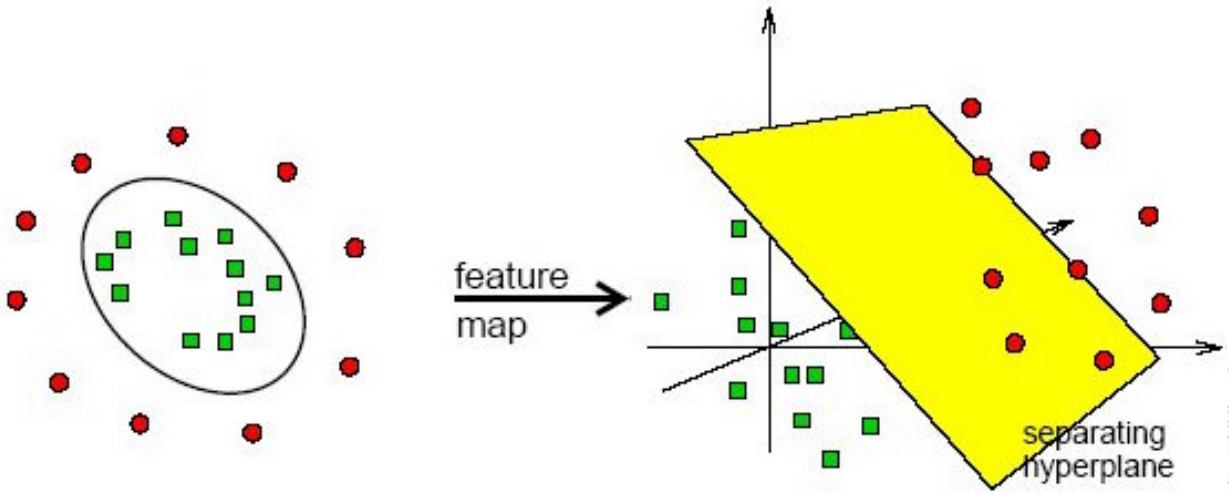
Classification in high dimension

- since high-dimensional spaces are almost empty,
- it should be easier to separate groups in high-dimensional space with an adapted classifier,
- the larger p is, the higher the likelihood that we can separate the classes perfectly with a hyperplane



Classification in high dimension

- since high-dimensional spaces are almost empty,
- it should be easier to separate groups in high-dimensional space with an adapted classifier,
- the larger p is, the higher the likelihood that we can separate the classes perfectly with a hyperplane



Overfitting

Outline

In high dimensional spaces, nobody can hear you scream

Concentration phenomena

Surprising asymptotic properties for covariance matrices

Volume of the ball

Volume of the ball of radius r is $V_p(r) = r^p \frac{\pi^{p/2}}{\Gamma(p/2+1)}$,

Volume of the ball

Volume of the ball of radius r is $V_p(r) = r^p \frac{\pi^{p/2}}{\Gamma(p/2+1)}$,

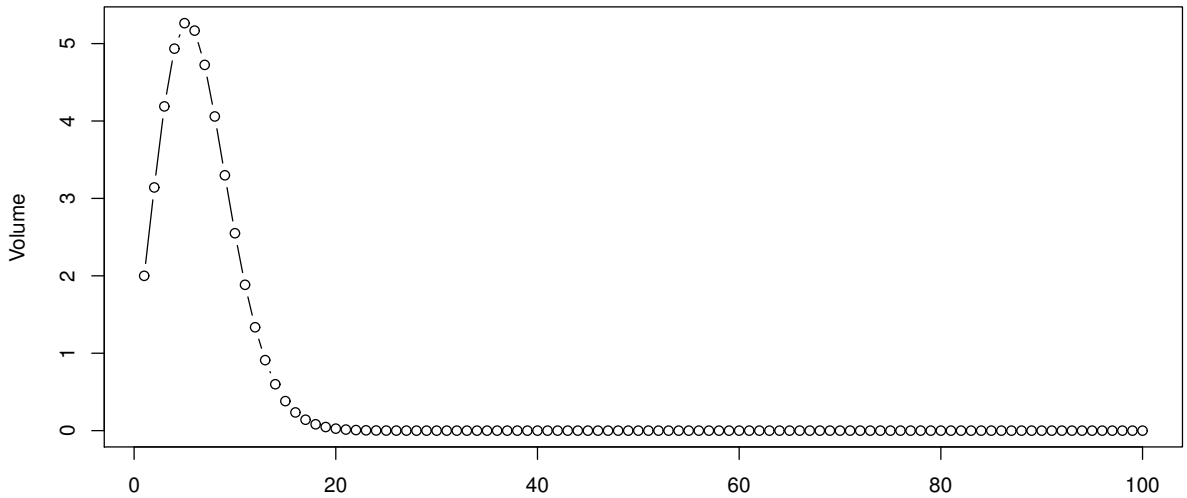


Fig. Volume of a ball of radius 1 regarding to the dimension p.

Volume of the ball

Volume of the ball of radius r is $V_p(r) = r^p \frac{\pi^{p/2}}{\Gamma(p/2+1)}$,

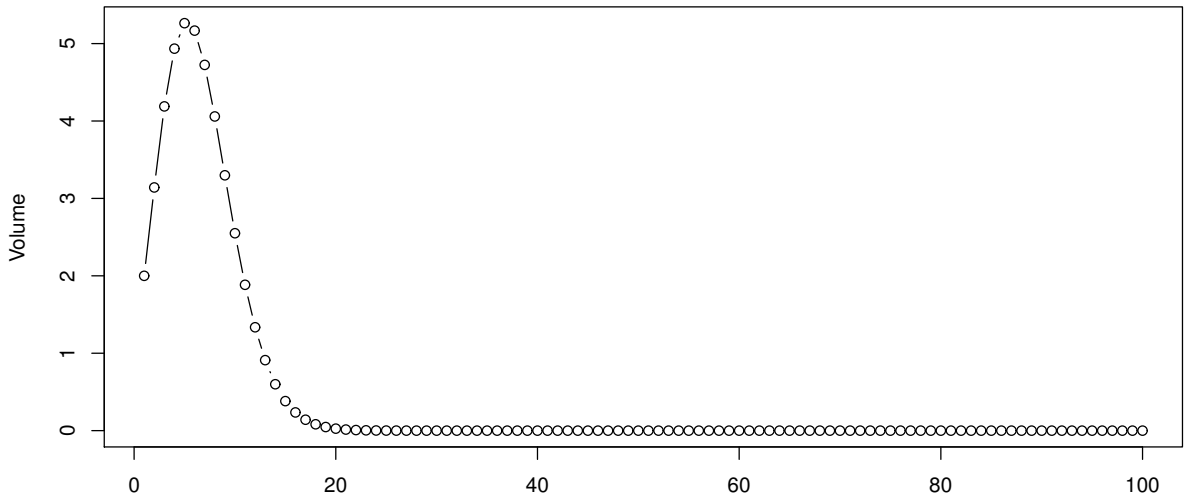


Fig. Volume of a ball of radius 1 regarding to the dimension p.

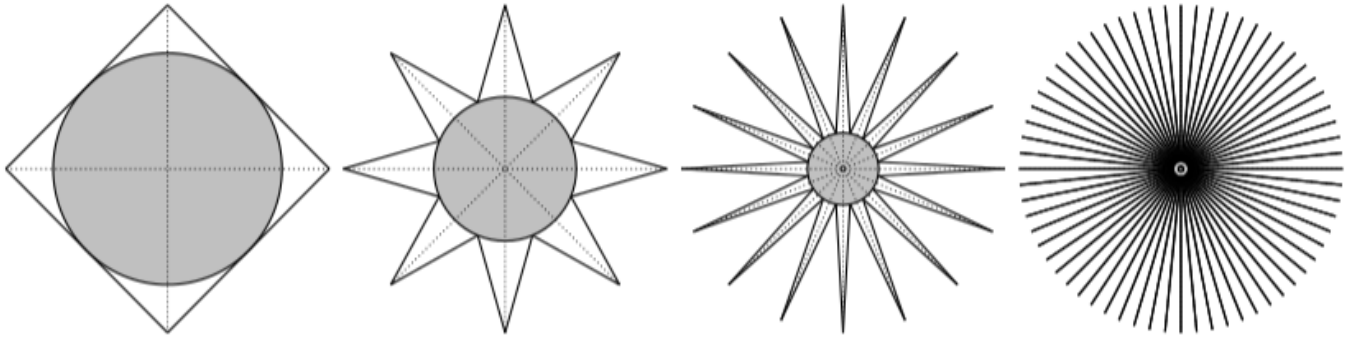
Consequence: if you want to cover $[0,1]^p$ with a union of n unit balls, you need

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

For $p = 100$, $n = 42\ 10^{39}$.

Corners of the hypercube

Assume you draw n samples with uniform law in the hypercube, most sample points will be in corners of the hypercube :



Volume of the shell

Probability that a uniform variable on the unit sphere belongs to the shell between the spheres of radius 0.9 and 1 is

$$P(X \in S_{0.9}(p)) = 1 - 0.9^p \underset{p \to \infty}{\longrightarrow} 1$$

Volume of the shell

Probability that a uniform variable on the unit sphere belongs to the shell between the spheres of radius 0.9 and 1 is

$$P(X \in S_{0.9}(p)) = 1 - 0.9^p \underset{p \to \infty}{\longrightarrow} 1$$

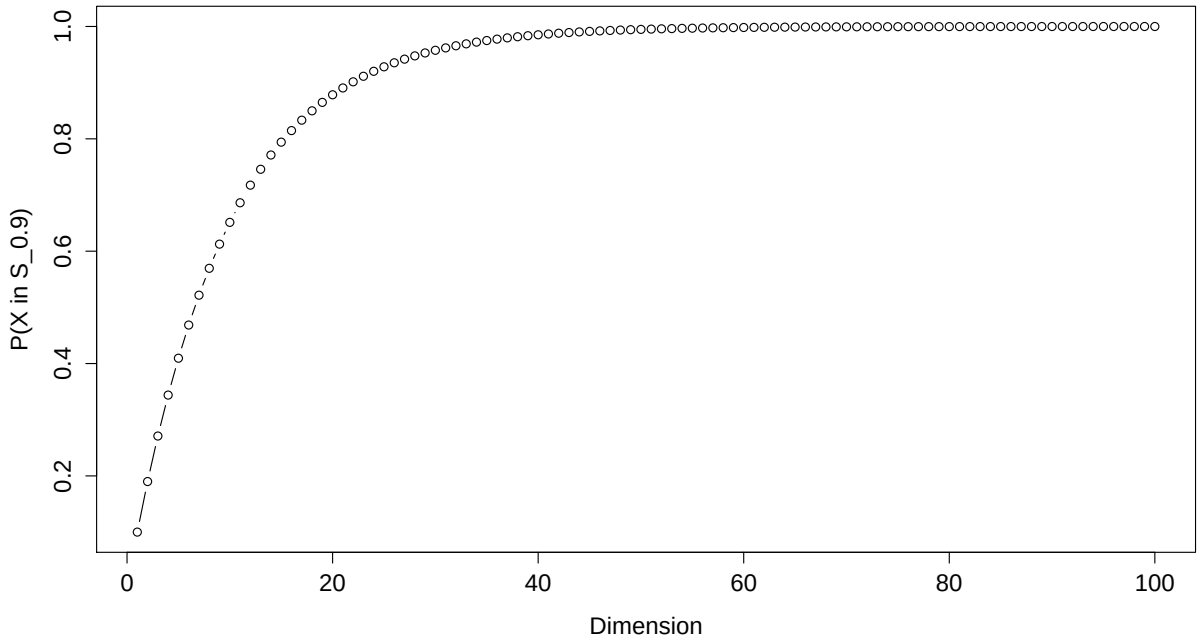


Fig. Probability that X belongs to the shell $S_{0.9}$ regarding to the dimension p.

Samples are close to an edge of the sample

$X_1, \dots X_n$ i.i.d. in dimension p, with uniform distribution on the unit ball. Median distance from the origin to the closest data point is given by

$$med(p, n) = \left(1 - \frac{1}{2^{\frac{1}{n}}}\right)^{\frac{1}{p}}.$$

For $n = 500$ and $p = 10$, $med = 0.52$, which means that most data points are closer to the edge of the ball than to the center.

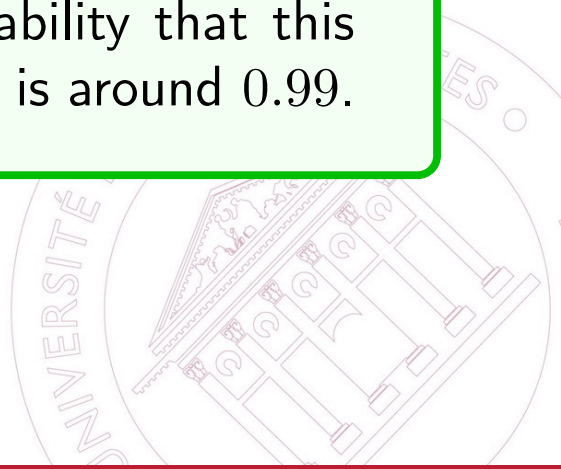
Concentration phenomena and estimation

Consequence: samples are closer to the boundary of the sample space than to other samples, which makes prediction much more difficult. Indeed, near the edges of the training sample, one must extrapolate from neighboring sample points rather than interpolate between them.

Concentration phenomena and estimation

Consequence: samples are closer to the boundary of the sample space than to other samples, which makes prediction much more difficult. Indeed, near the edges of the training sample, one must extrapolate from neighboring sample points rather than interpolate between them.

Exemple. Assume n data sampled independently with a uniform law on $[-1,1]^p$. You want to estimate $e^{-\|x\|^2/8}$ in 0 from your data. You choose as an estimator the observed value in x_i, the nearest neighbor of 0. For $n = 1000$ samples and $p = 10$, the probability that this nearest neighbor is at a distance larger than $\frac{1}{2}$ from 0 is around 0.99.



Where is located the mass of the Gaussian distribution ?

Mass of the standard Gaussian distribution in the ring between radius r and $r + dr$

$$\mathbb{P}[r \leq \|X\| \leq r+dr] \simeq \frac{e^{-r^2/2}}{(2\pi)^{p/2}}(V_p(r+dr)-V_p(r)) \simeq \frac{e^{-r^2/2}}{(2\pi)^{p/2}} r^{p-1}\ p dr\ V_p(1).$$

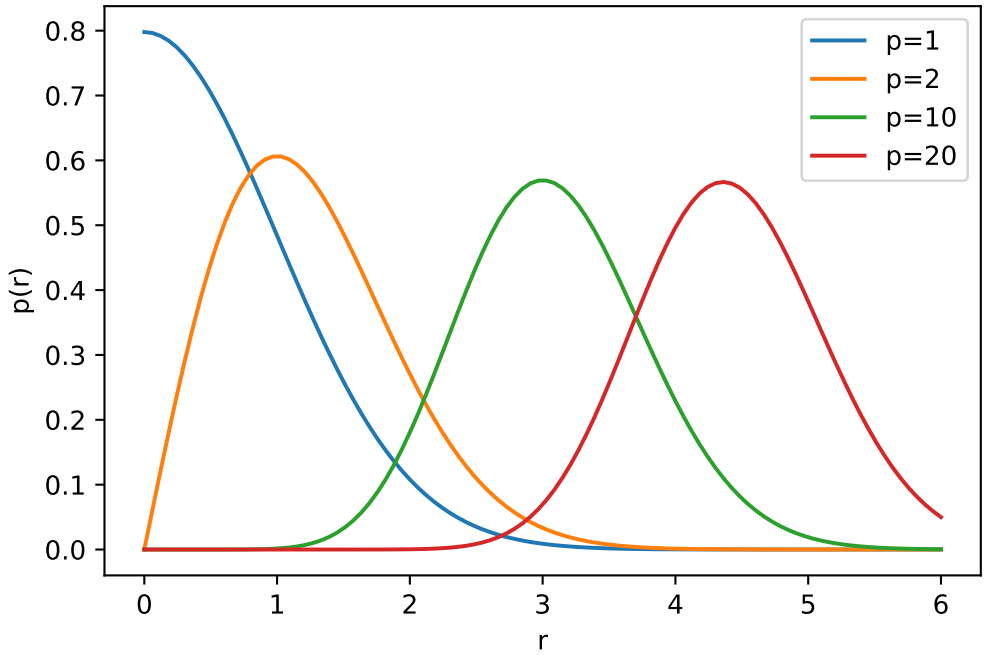
$\longrightarrow$ maximum for $r = \sqrt{p-1}$

Where is located the mass of the Gaussian distribution ?

Mass of the standard Gaussian distribution in the ring between radius r and $r + dr$

$$\mathbb{P}[r \leq \|X\| \leq r+dr] \simeq \frac{e^{-r^2/2}}{(2\pi)^{p/2}}(V_p(r+dr)-V_p(r)) \simeq \frac{e^{-r^2/2}}{(2\pi)^{p/2}} r^{p-1}\, p dr\, V_p(1).$$

$\longrightarrow$ maximum for $r = \sqrt{p-1}$

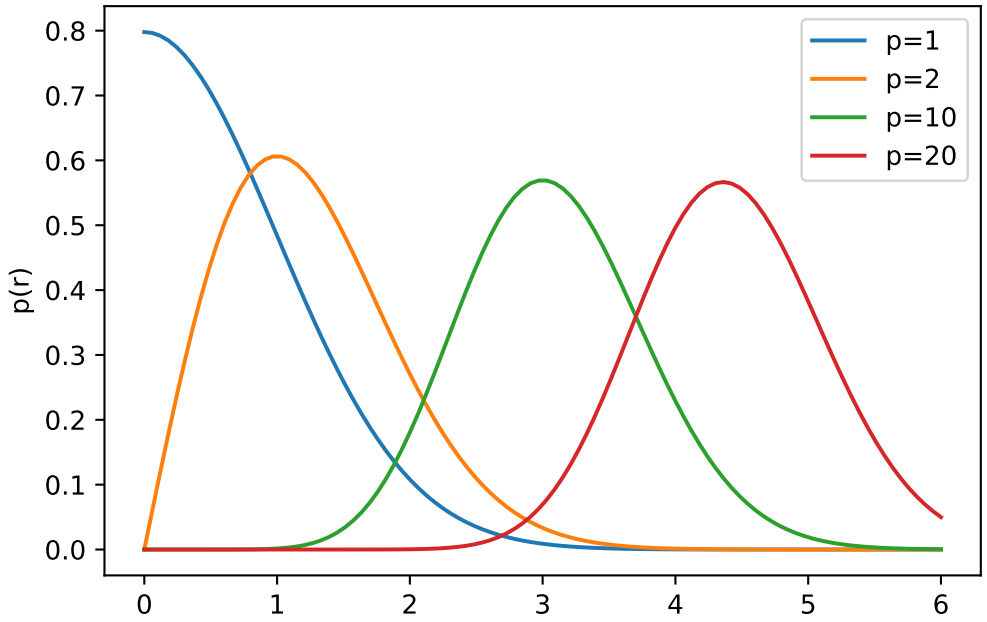


Where is located the mass of the Gaussian distribution ?

Mass of the standard Gaussian distribution in the ring between radius r and $r + dr$

$$\mathbb{P}[r \leq \|X\| \leq r+dr] \simeq \frac{e^{-r^2/2}}{(2\pi)^{p/2}}(V_p(r+dr)-V_p(r)) \simeq \frac{e^{-r^2/2}}{(2\pi)^{p/2}} r^{p-1}\, p dr\, V_p(1).$$

$\longrightarrow$ maximum for $r = \sqrt{p-1}$



Most of the mass of a Gaussian distribution is located in areas where the density is extremely small compared to its maximum value.

Outline

In high dimensional spaces, nobody can hear you scream

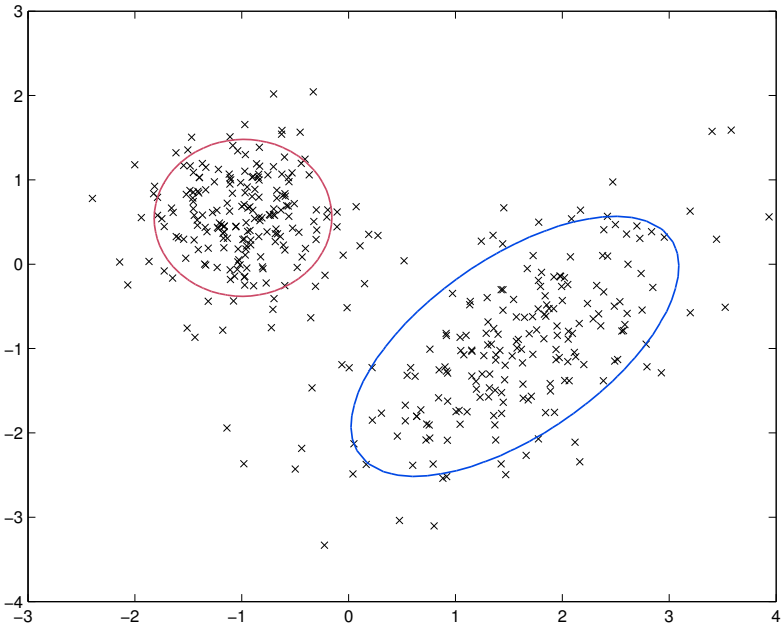
Concentration phenomena

Surprising asymptotic properties for covariance matrices

Covariance matrices

Sample Covariance Matrices appear everywhere in statistics

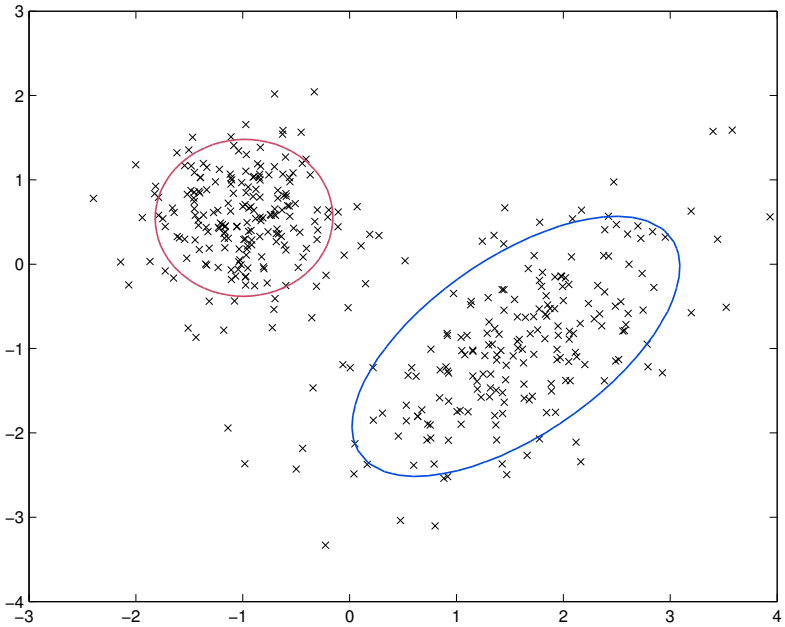
- classification with gaussian mixture models
- principal component analysis (PCA)
- in linear regression with least squares, etc...



Covariance matrices

Sample Covariance Matrices appear everywhere in statistics

- classification with gaussian mixture models
- principal component analysis (PCA)
- in linear regression with least squares, etc...



Problems:

- often necessary to invert Σ
- if n is not large enough, the estimates of Σ are ill-conditionned or singular
- sometimes necessary to estimate the eigenvalues of Σ

Covariance matrices

Context $x_1 \ldots, x_n \in \mathbb{R}^p$ i.i.d. samples from a gaussian multivariate distribution $\mathcal{N}(0, \Sigma_p)$.

Covariance matrices

Context $x_1 \ldots, x_n \in \mathbb{R}^p$ i.i.d. samples from a gaussian multivariate distribution $\mathcal{N}(0, \Sigma_p)$. The maximum likelihood estimator for Σ_p is the sample covariance matrix

$$\widehat{\Sigma}_p = \frac{1}{n} \sum_{k=1}^{n} x_k x_k^T .$$

Covariance matrices

Context $x_1 \ldots, x_n \in \mathbb{R}^p$ i.i.d. samples from a gaussian multivariate distribution $\mathcal{N}(0, \Sigma_p)$. The maximum likelihood estimator for Σ_p is the sample covariance matrix

$$\widehat{\Sigma}_p = \frac{1}{n} \sum_{k=1}^{n} x_k x_k^T .$$

If p is fixed and $n \to \infty$, then (strong law of larger numbers) for any matrix norm

$$\|\widehat{\Sigma}_p - \Sigma_p\| \xrightarrow{a.s.} 0$$

Covariance matrices

Context $x_1 \ldots, x_n \in \mathbb{R}^p$ i.i.d. samples from a gaussian multivariate distribution $\mathcal{N}(0, \Sigma_p)$. The maximum likelihood estimator for Σ_p is the sample covariance matrix

$$\widehat{\Sigma}_p = \frac{1}{n} \sum_{k=1}^{n} x_k x_k^T .$$

If p is fixed and $n \to \infty$, then (strong law of larger numbers) for any matrix norm

$$\|\widehat{\Sigma}_p - \Sigma_p\| \xrightarrow{a.s.} 0$$

Random matrices

- If $n, p \to \infty$ with $p/n \to c > 0$, then

$$\|\widehat{\Sigma}_p - I_p\|_2 \not\to 0 \quad (\|\|_2 \text{ denotes the spectral norm}).$$

- Even false for $p/n = 1/100$.

Covariance matrices - Random matrix regime

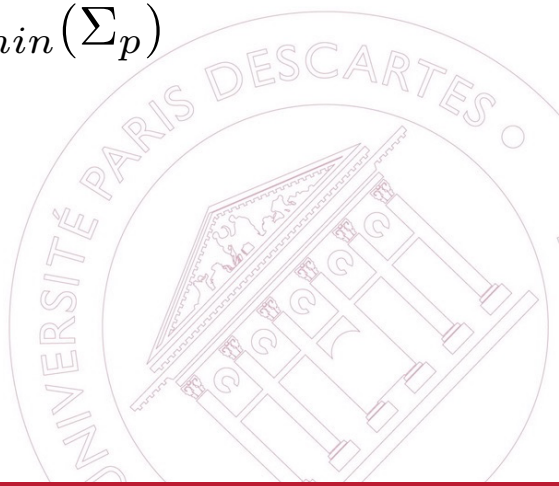
Context $x_1 \dots, x_n \in \mathbb{R}^p$ i.i.d. samples from a gaussian multivariate distribution $\mathcal{N}(0, I_p)$. Note $X = (x_1, \dots, x_p)$.

- $p/n = c > 1$
- Convergence in $\|\|_\infty$

$$\max_{i,j} |\widehat{\Sigma}_{i,j} - \delta_{i,j}| \xrightarrow{a.s.} 0$$

- However, we lose the convergence in spectral norm since

$$\text{rank}(X) \leq p \;\Rightarrow\; \lambda_{min}(\widehat{\Sigma_p}) = 0 < 1 = \lambda_{min}(\Sigma_p)$$



Covariance matrices - Random matrix regime

Context $x_1 \ldots, x_n \in \mathbb{R}^p$ i.i.d. samples from a gaussian multivariate distribution $\mathcal{N}(0, I_p)$. Note $X = (x_1, \ldots, x_p)$.

- $p/n = c > 1$
- Convergence in $\|\|_\infty$

$$\max_{i,j} |\widehat{\Sigma}_{i,j} - \delta_{i,j}| \xrightarrow{a.s.} 0$$

- However, we lose the convergence in spectral norm since

$$\text{rank}(X) \leq p \;\Rightarrow\; \lambda_{min}(\widehat{\Sigma_p}) = 0 < 1 = \lambda_{min}(\Sigma_p)$$

No contradiction with the fact that all norms are equivalent in finite dimension.

Covariance matrices - Random matrix regime

More precisely, the random matrices theory tells us that when $p, n \to \infty$ with $p/n \to c > 0$, then **[Marčenko-Pastur Theorem, 1967]**

$$\frac{1}{p} \sum_{k=1}^{p} \delta_{\lambda_k(\widehat{\Sigma_p})} \xrightarrow{a.s.} \mu \quad \text{weakly,}$$

with μ the Marčenko-Pastur law of parameter c, which satisfies

- $\mu(\{0\}) = max(0, 1 - c^{-1})$
- on $(0, \infty)$, μ has a continuous density supported on $[(1 - \sqrt{c})^2, (1 + \sqrt{c})^2]$.

The Marčenko–Pastur law

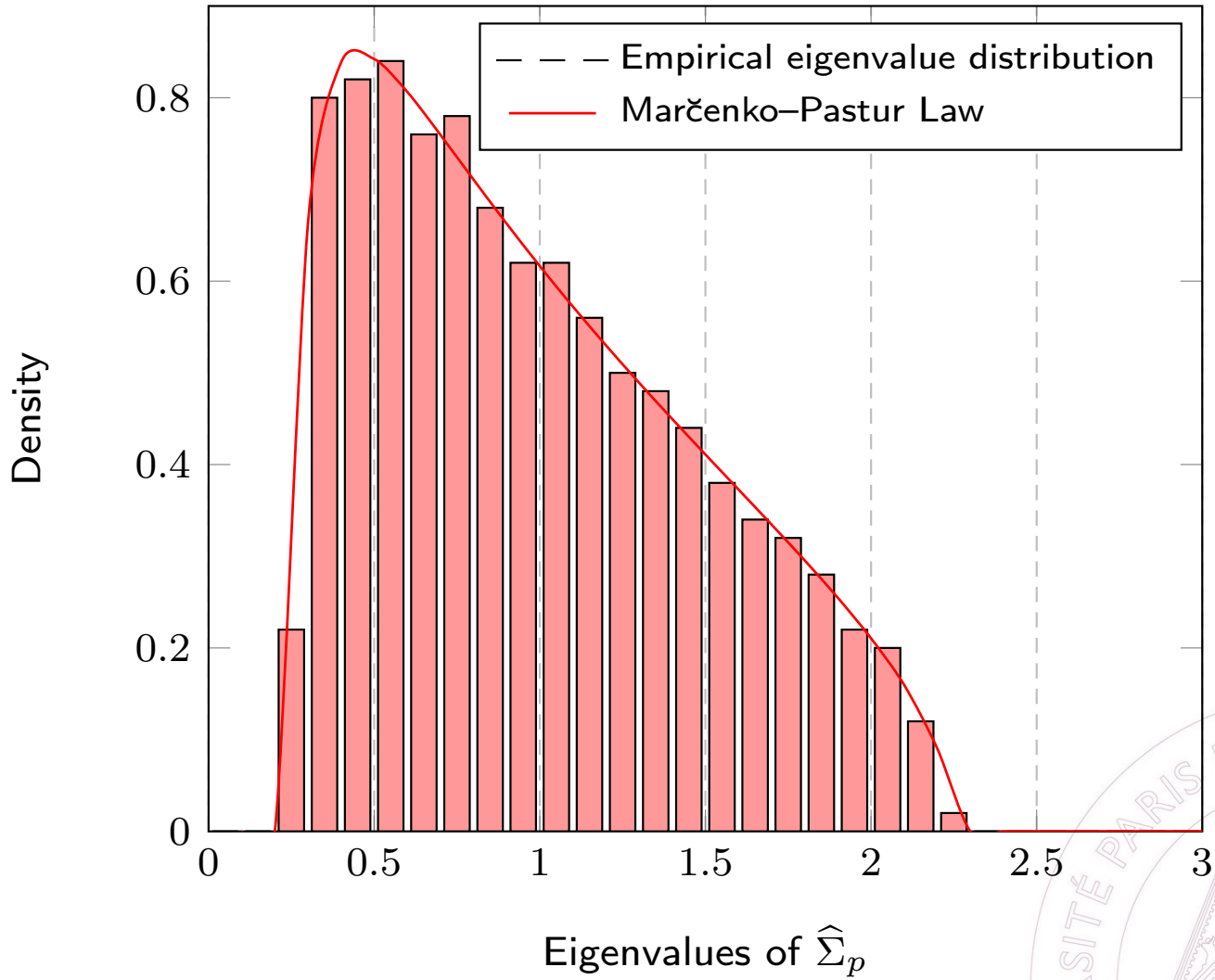


Figure: Histogram of the eigenvalues of $\widehat{\Sigma}_p$ for $p = 500$, $n = 2000$, $\Sigma_p = I_p$.

The Marčenko–Pastur law

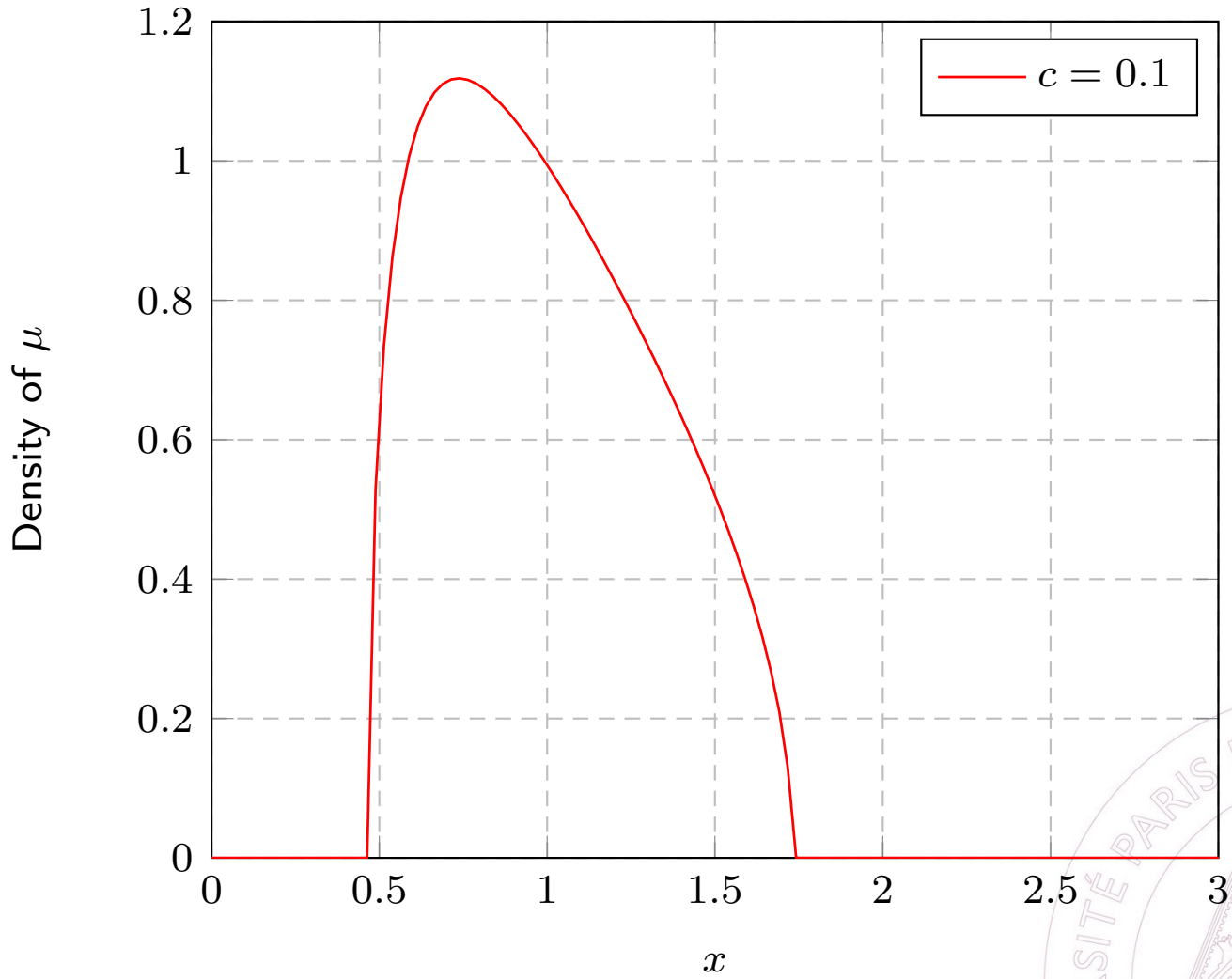


Figure: Marčenko-Pastur law for different limit ratios $c = \lim_{p\to\infty} p/n$.

The Marčenko–Pastur law

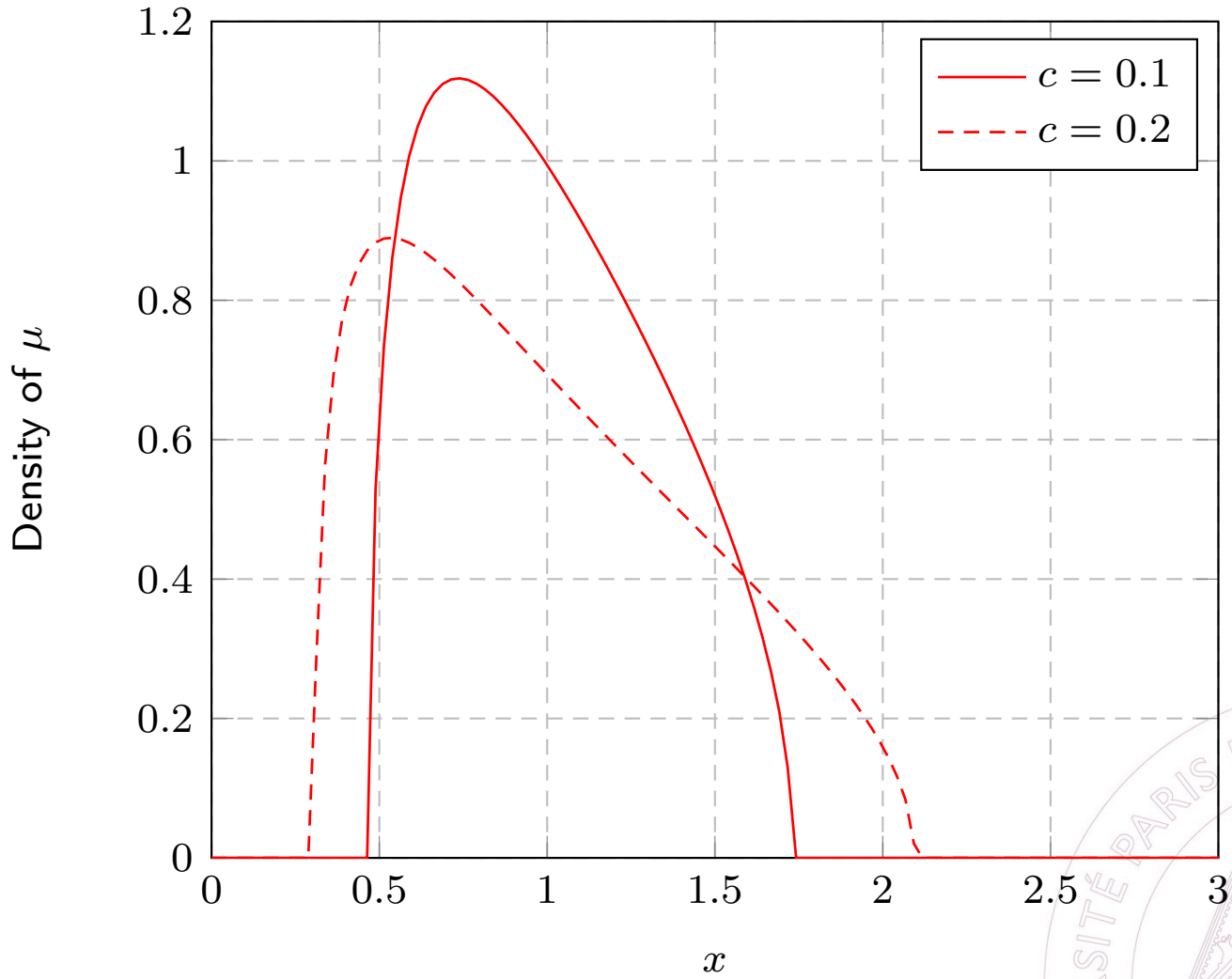


Figure: Marčenko-Pastur law for different limit ratios $c = \lim_{p\to\infty} p/n$.

The Marčenko–Pastur law

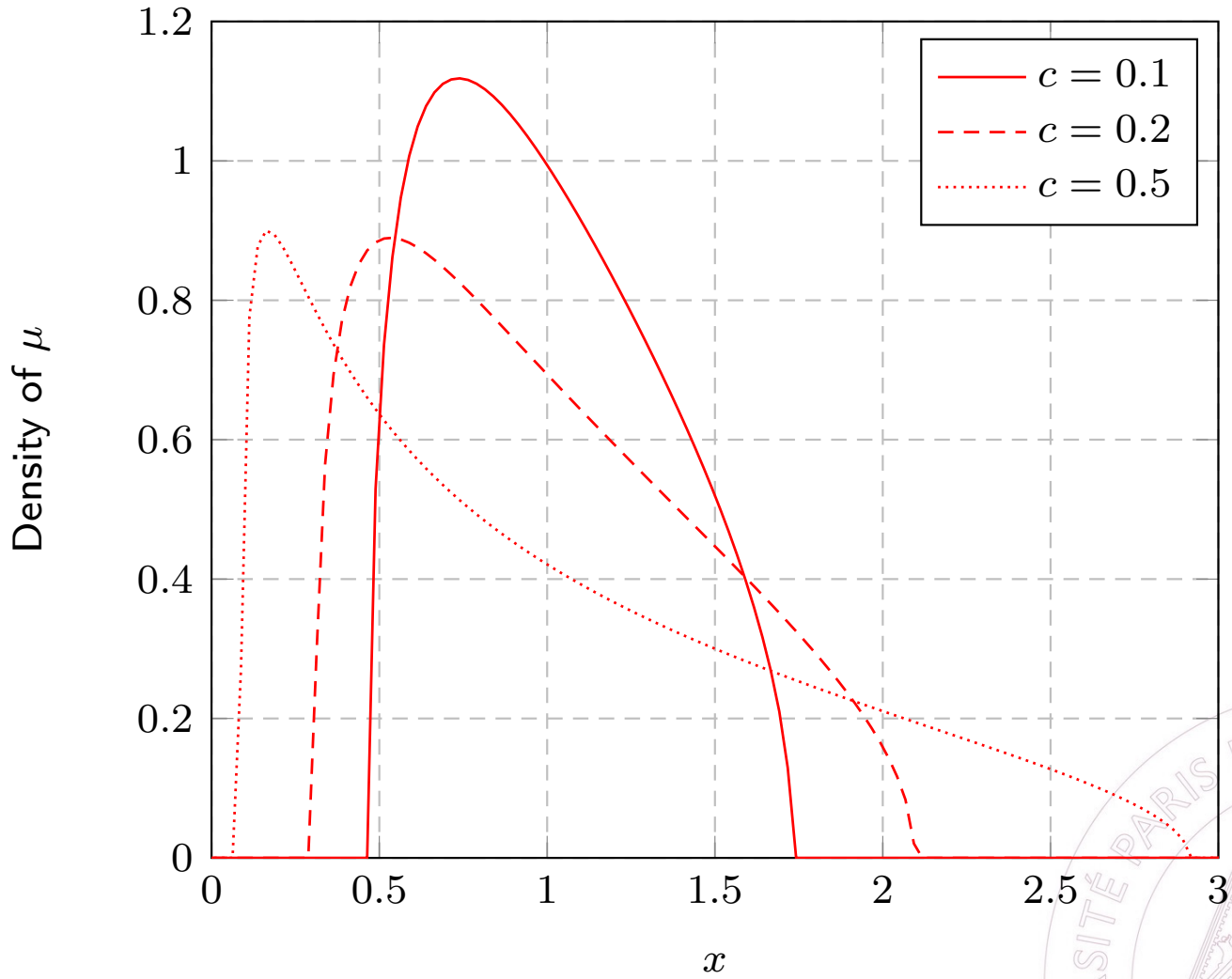


Figure: Marčenko-Pastur law for different limit ratios $c = \lim_{p\to\infty} p/n$.

Classical ways to avoid the curse of dimensionality

Dimension reduction:

- the problem comes from that p is too large,
- therefore, reduce the data dimension to $d \ll p$,
- such that the curse of dimensionality vanishes!

Regularization:

- the problem comes from that parameter estimates are unstable,
- therefore, regularize these estimates,
- such that the parameter are correctly estimated!

Parsimonious models:

- the problem comes from that the number of parameters to estimate is too large,
- therefore, make restrictive assumptions on the model,
- such that the number of parameters to estimate becomes more "decent"!