On the Curses and Blessings of Dimensionality

Short talk : 19/09/2022

Hugues Van Assel

Setting

$$X \in \mathbb{R}^{n \times p}$$

Today we look at what happens when p is very large.

(Lots of genes etc…)

Space is empty

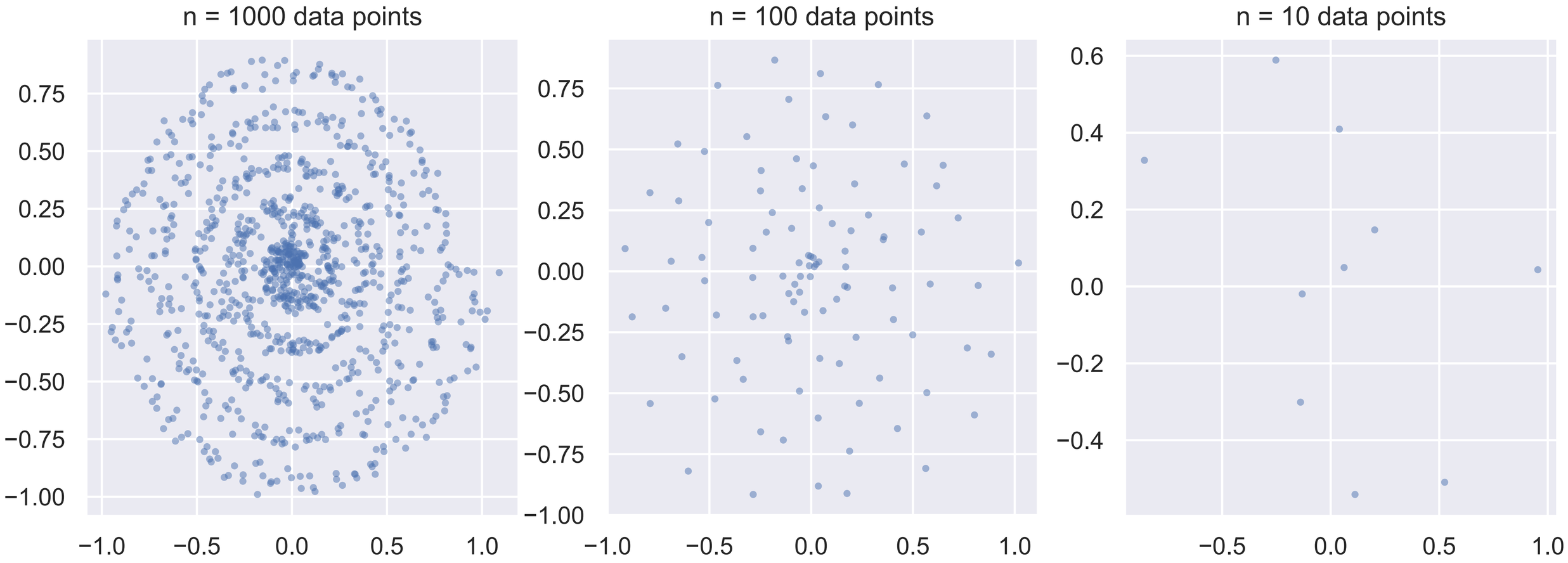
Number n of points $x_1, \ldots, x_n$ required for covering $[0,1]^p$ by the balls $B(x_i, 1)$:

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

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

Space is empty

The density of data in local neighborhoods is too sparse to fit distributions.



All the mass is on the edge

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

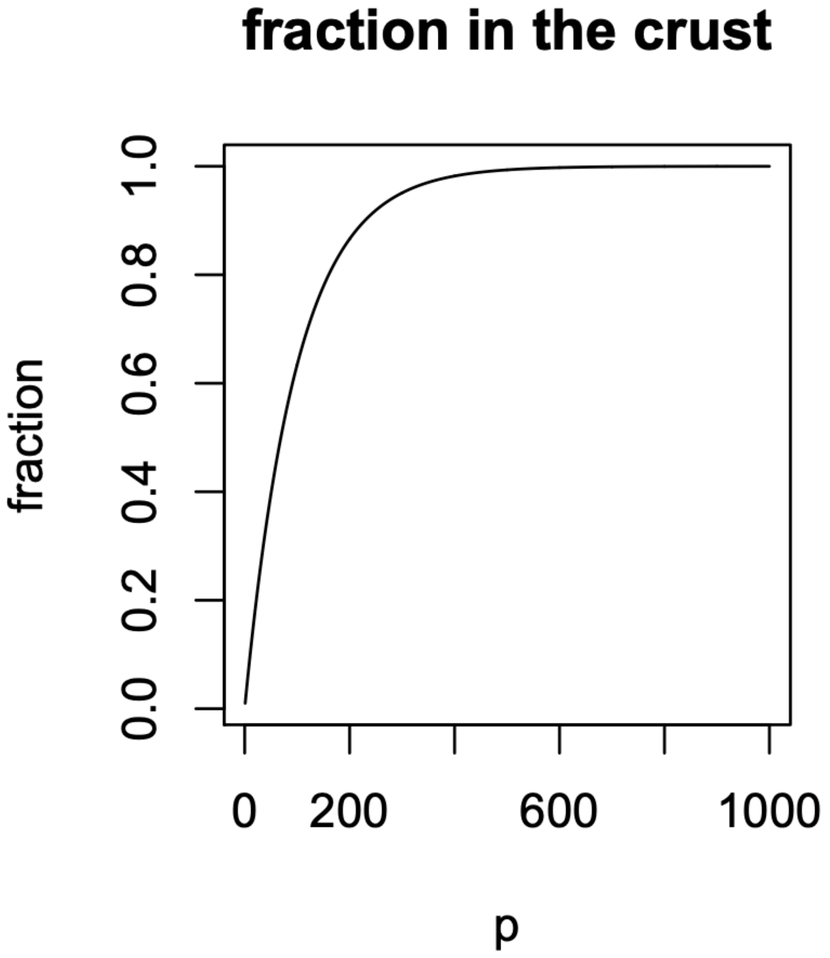
Ball: $B_p(0, r)$

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!



All the mass is on the edge

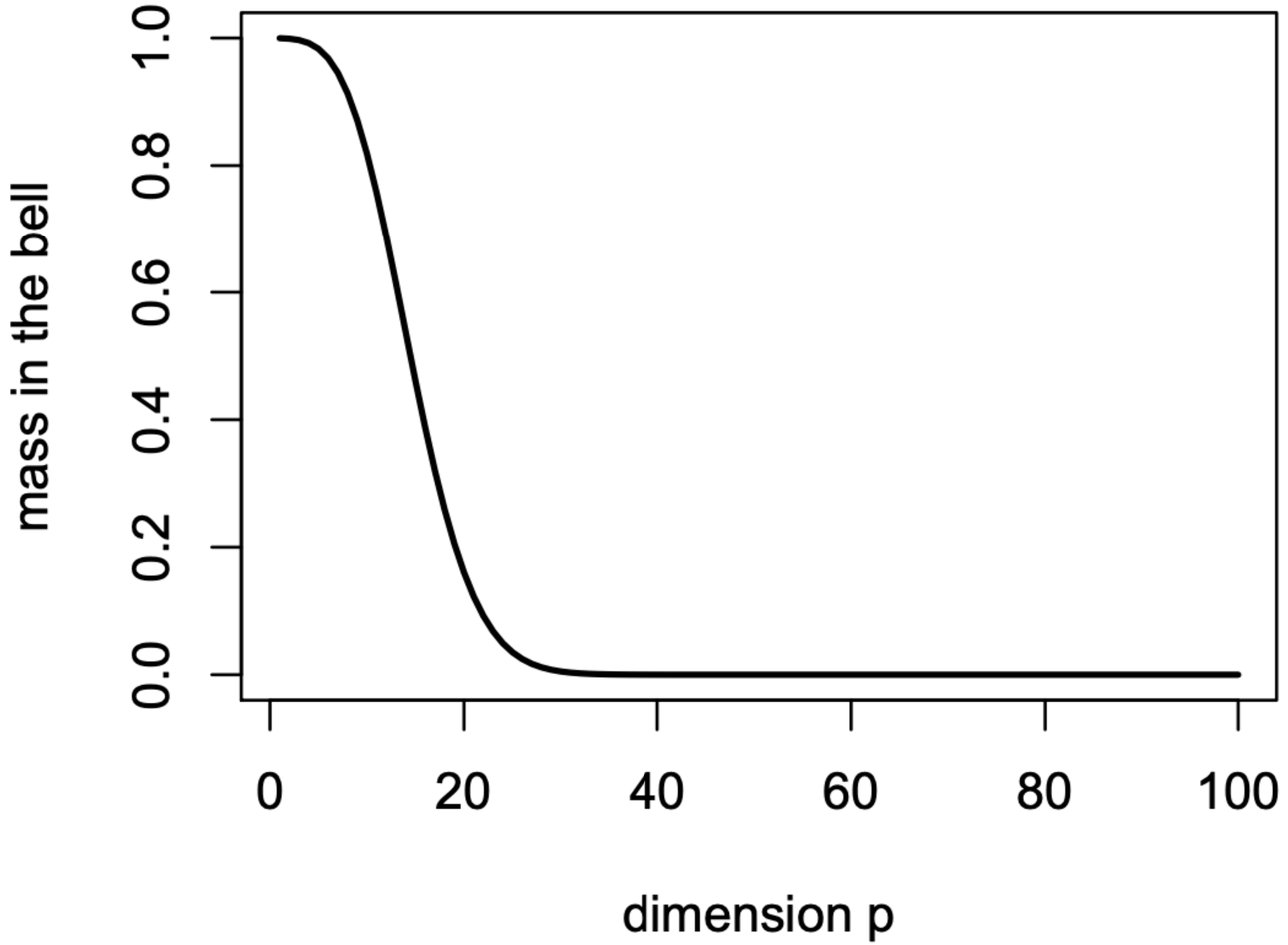


Figure: Mass of the standard Gaussian distribution $g_p(x)\,dx$ in the "bell" $B_{p,0.001} = \{x \in \mathbb{R}^p : g_p(x) \geq 0.001 g_p(0)\}$ for increasing dimensions p.

Distances concentrate

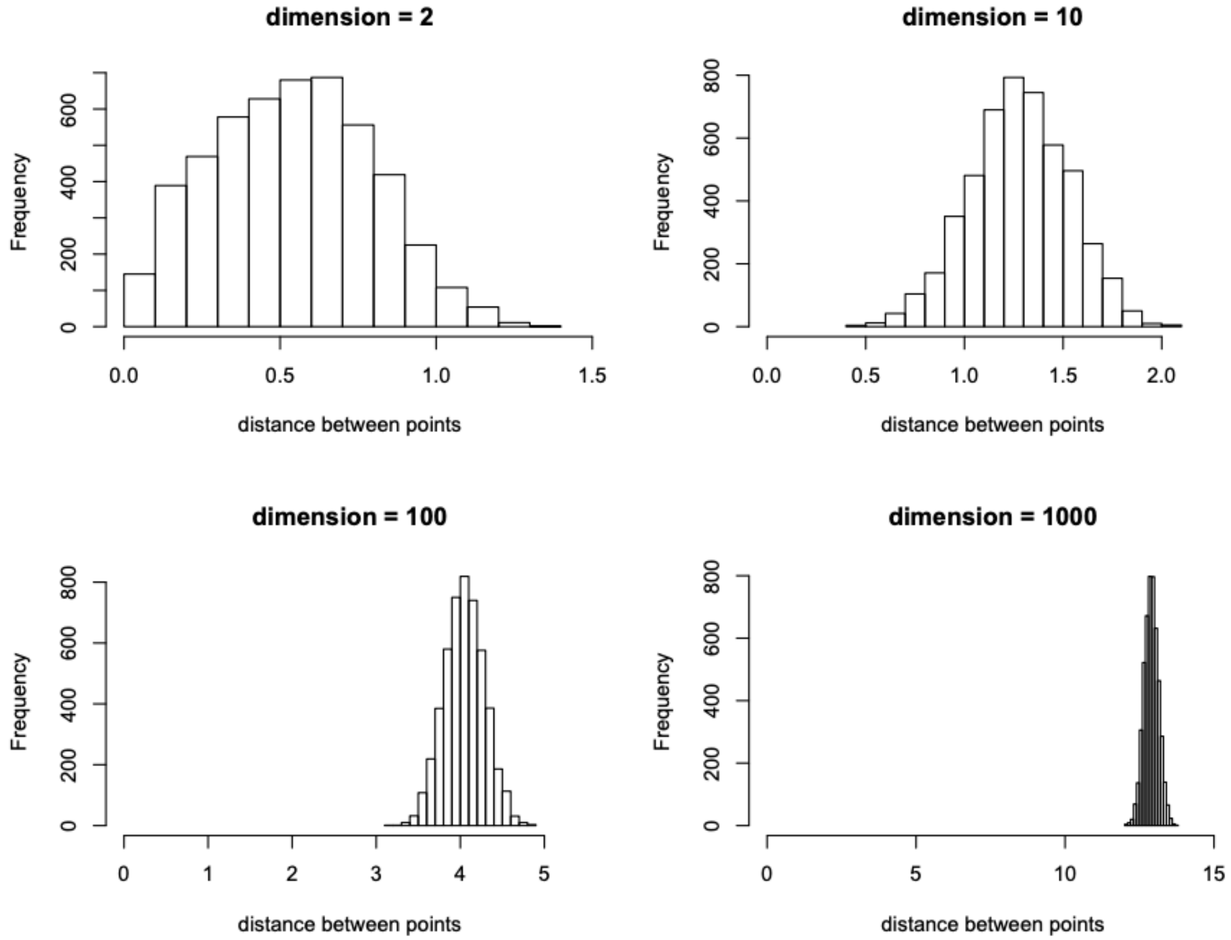
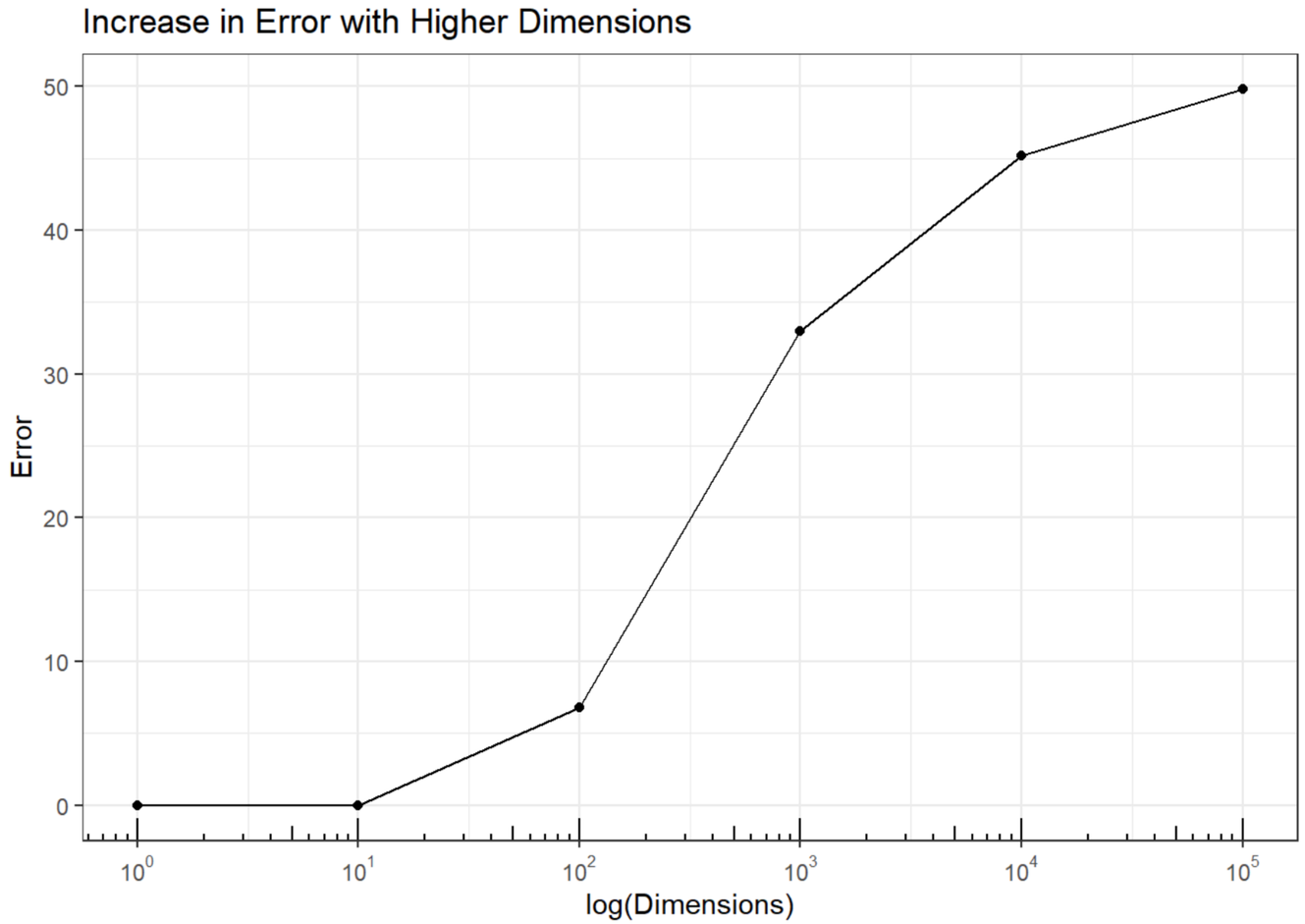


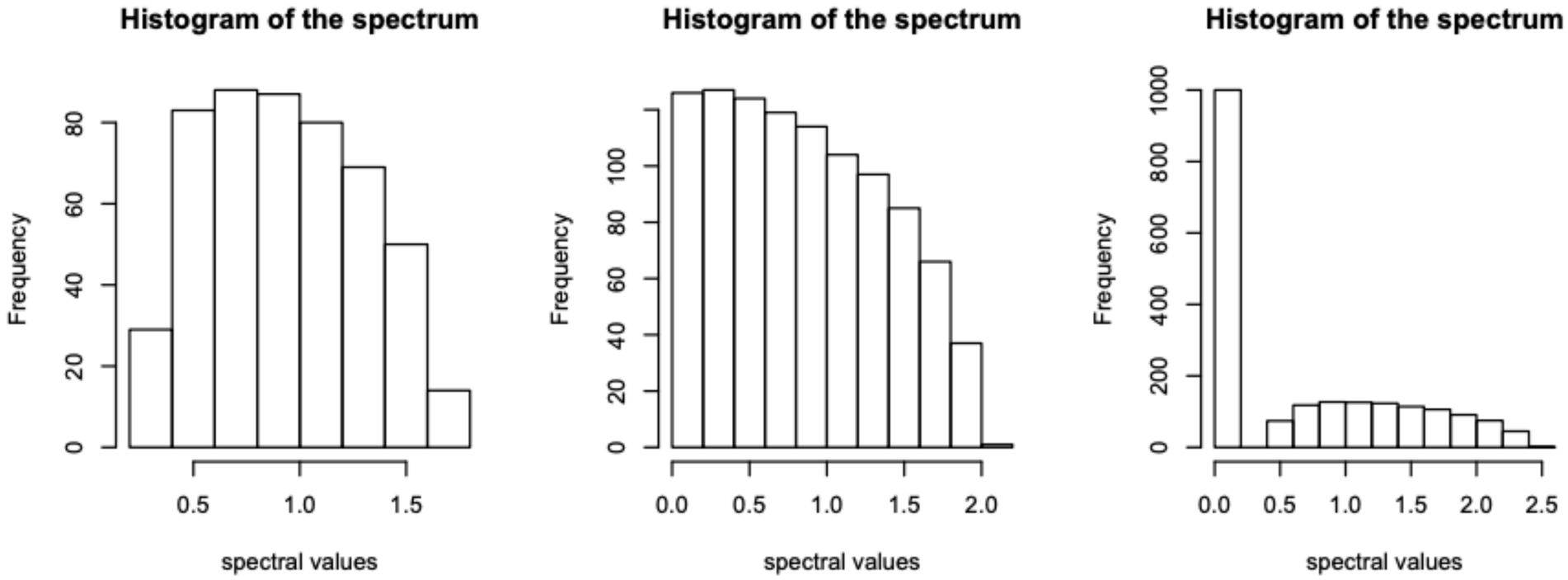
Figure: Histograms of the pairwise-distances between $n = 100$ points sampled uniformly in the hypercube $[0, 1]^p$, for $p = 2, 10, 100$ and 1000.

Impact on Analysis

Nearest neighbors classifiers classify points based on the majority of classes among the nearest points. In this simulation, we sample 100 points from a Gaussian distribution with mean -5 and std 1 and 100 points from a Gaussian with mean 5 and std 1. Uniform noise in [-5,5] is then added.



Impact on Analysis



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).

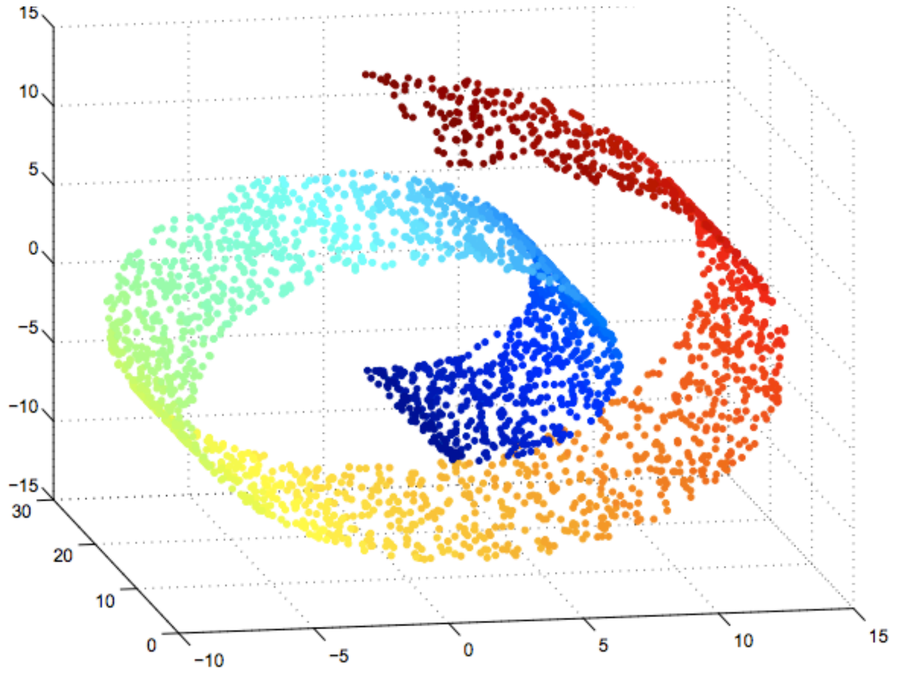
To summarize

- In high dimension, densities are very sparse.
- All the mass is on the edges / corners.
- The distance between all pairs of point becomes the same.

But …

In real life applications, data has structure.

High-dimensional data are usually concentrated around low-dimensional structures reflecting the (relatively) small complexity of the systems producing the data.



Reference : Christophe Giraud’s course.