What to Expect when Exploring High-Dimensional Data ?

BUCHE Marianne[1], CARRE Yohann[1], LE Sébastien[1]
[1] AGROCAMPUS OUEST – Laboratoire de Mathématiques Appliquées, Rennes
Contact author: sebastien.le@agrocampus-ouest.fr

What is a high dimensional space ?

High-dimensional data can be defined in different ways:
- Data which are impossible to study with traditional analyses;
- Data bigger than in usual analyses;
- Data so big that random effect and structure effect are uneasy to separate.

Today, the technological advances in computer science have lead to the gathering of bigger and bigger data.

p: number of variables
n: number of individuals

Some example of data in high dimension

- p>>n: genetic and biotechnology data
- n>>p: survey or financial data
- p & n in high dimensions: picture and spectra analysis

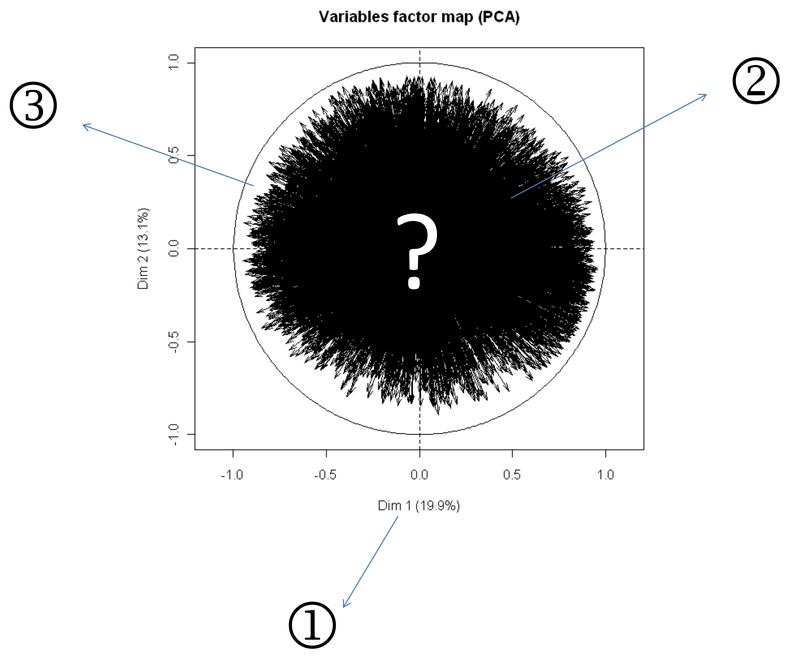
What to expect with the variables in a high-dimensional dataset ?

In this exploratory multivariate framework, we use the PCA to study a quantitative dataset with 12663 variables and 27 individuals.
Three main problems can be noticed when looking at the representation of the variables:

① How can we interpret the inertia values in such a context ?

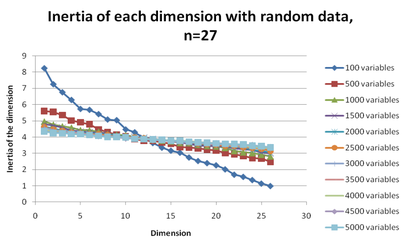
② What to think about the presence of structure / conjuncture variables ?

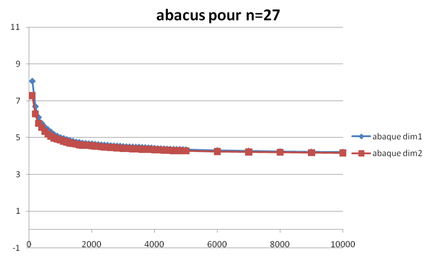
③ Why do the variables never reach the border of the correlation circle ?



① What about inertia ?

Random data involve a fair repartition of inertia on the different axes of factorial analysis: each axe should have 100/(n-1) % of inertia.
But as factorial analysis classify the axes according to their inertia, the first axes have a slightly higher inertia and the last ones have a lower inertia.
It is possible to estimate by simulation the expected values for the inertia of the first axes.

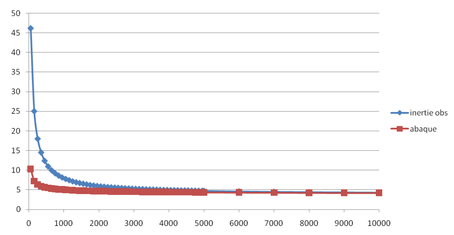




The more the number of variables is important, the more the inertia of the different axes is equally distributed.

② What happens when we add structural data to random data ?

The 50 best projected variables on the first factorial axis of a genomic dataset have been selected, and have been mixed up with random data.
As the 50 variables are highly correlated, they mainly add their inertia to the same axe, involving an overload on one dimension which becomes the first axis.

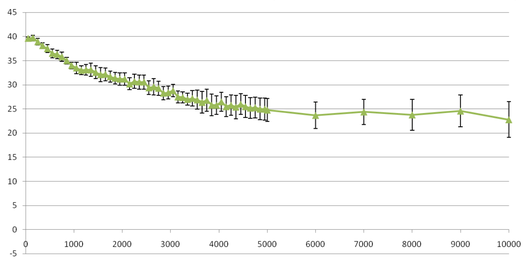


Representation of the inertia of the first axis according to the number of random variables.

The more random data there are, the less the difference of inertia is important, and the less the structure variables have influence.

How to quantify the information ?

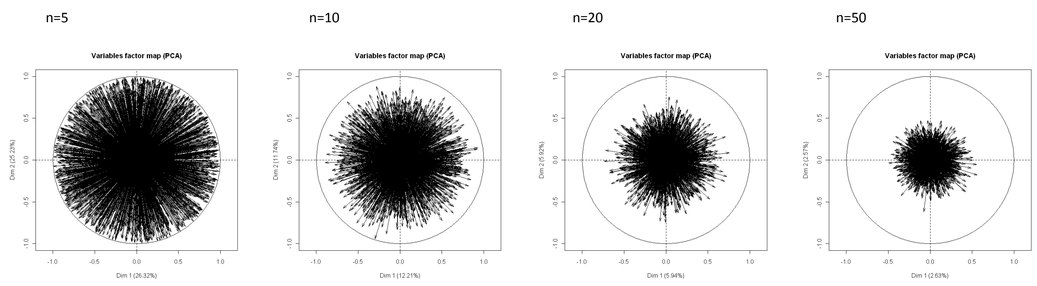
Estimation of the number of structural variables in the dataset (50 structural variables)



An abacus of the expected inertia values with only random data has been built to compare real values to a random situation.
It is thus possible to estimate the number of variables "responsible" for such a difference of inertia, by adding the inertia of the most projected variables one by one on the considered axis, until the inertia difference is reached.
This number of added variables is the wanted estimation.

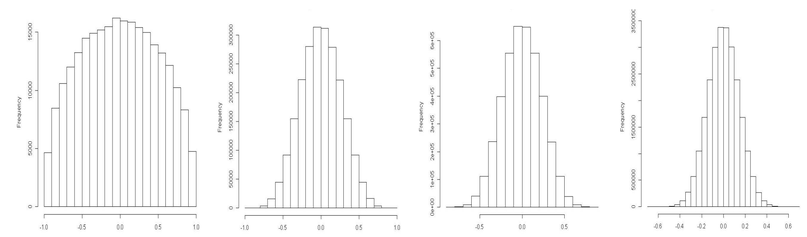
③ The correlation circle never reaches the border:

Representation of the correlation circle for 3000 random variables with various numbers of individuals

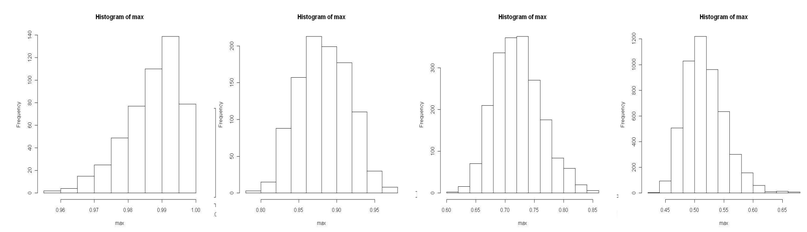


The correlation circle "shrinks" with the increase in the number of dimensions of the space.

How do the correlations evolve with the number of dimensions of the space ?



Histograms of correlations between variables with the same ratio p/n=100



Histograms of the maximum of correlation observed between variables with the same ratio p/n=100

The bigger the space is, the more variables can be « isolated »:
a space with twice as many dimensions is able to afford
more than twice as many variables to observe the same correlation distribution.

Conclusion:

- High dimensional data with p>>n involve that random variables appear significant.
- The correlation circle never reaches the border due to the high dimensions of the space.
- It is not possible to separate entirely structure effect and random effect. But it is possible to estimate the importance of the structure.