Statistical Learning with High-dimensional Data

Pr. Charles BOUVEYRON

Professor of Statistics
Chair of the Institut 3IA Côte d'Azur
Université Côte d'Azur & Inria

charles.bouveyron@univ-cotedazur.fr
@cbouveyron

Preamble

"Ce qui est simple est toujours faux.
Ce qui ne l'est pas est inutilisable."

Paul Valéry

Outline

The AI revolution hasn't happened yet!

Artificial intelligence is a strategic field of research:

- with direct application in most scientific fields (Medicine, Biology, Astrophysics, Humanities)
- and with probably the most impact in innovation and transfer (health, transport, defense).

The AI revolution hasn't happened yet!

Artificial intelligence is a strategic field of research:

- with direct application in most scientific fields (Medicine, Biology, Astrophysics, Humanities)
- and with probably the most impact in innovation and transfer (health, transport, defense).

The recent and impressive NN results should not hide the remaining issues:

- deep learning has impressive results in a few specific cases and with a high-level supervision,
- use of DL techniques in various fields are promising but not well understood.

The AI revolution hasn't happened yet!

Artificial intelligence is a strategic field of research:

- with direct application in most scientific fields (Medicine, Biology, Astrophysics, Humanities)
- and with probably the most impact in innovation and transfer (health, transport, defense).

The recent and impressive NN results should not hide the remaining issues:

- deep learning has impressive results in a few specific cases and with a high-level supervision,
- use of DL techniques in various fields are promising but not well understood.

"*Artificial Intelligence: the revolution hasn't happened yet*"
M. Jordan (UC Berkley)

Open problems of AI

Some open problems are critical:

- reliability of models and algorithms,
- handling data heterogeneity (categorical, functional, networks, images, texts, ...),
- unsupervised learning (clustering, dimension reduction),
- learning from HD and small data (n small / p large),

Open problems of AI

Some open problems are critical:

- reliability of models and algorithms,
- handling data heterogeneity (categorical, functional, networks, images, texts, ...),
- unsupervised learning (clustering, dimension reduction),
- learning from HD and small data (n small / p large),

Combination of statistical theory with deep learning techniques is certainly the future of AI!

AI in France

French policy for AI:

- C. Villani presented in March a recommendation report for AI,
- President Macron announced the creation of a network of AI institutes.

The 3IA institutes:

- 12 french research centers applied for the 3IA call in Sept.,
- 4 projects have been selected in the Spring 2019:
  - Paris, Toulouse, Grenoble
  - and Nice!

A few examples: Cervical cancer detection

Cervical cancer detection:

- it is an important public health field which is currently treated mostly manually,
- screening by human experts is complicated by the amount of cells (20 000/smear),
- and by the very small proportion of cancer cells (less than 1%).

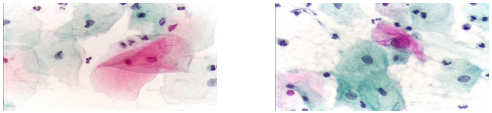
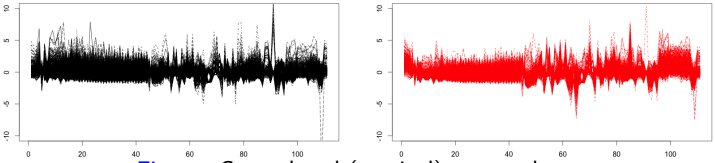


Figure: Normal (left) and abnormal (right) pap smears.

Classification is useful in this context:

- for building supervised classifiers which can select the most likely cancer cells,
- for helping experts in labeling the learning data through weakly-supervised classification,
- for selecting discriminative variables which can be used in a semi-automatic process.



A few examples: Sparse models in Medicine (HEGP)

Problem:

- overcome the curse of dimensionality that occurs in Metabolomics,
- for disease diagnostic and early-stage marker identification,
- metabolomic data fall into the "ultra-high dimensional data" case.

Our solution:

- a Bayesian variable selection technique for PCA,
- that identify the relevant variable for each stage of the disease.

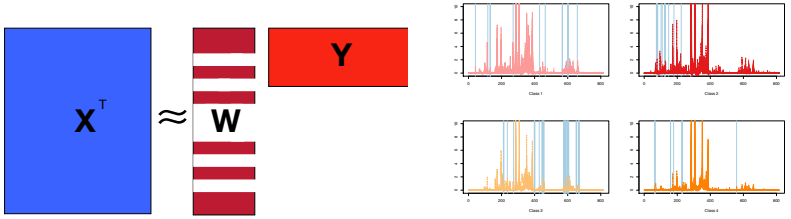


Figure: Functional co-clustering of Linky data (EDF).

Analysis of massive functional data (Linky / EDF)

Problem:

- Linky meters will allow EDF to have access to 27 million of Linky data,
- data are functional data and are measured every 30 minutes -> 17 520 obs./year,
- necessity to summarize those massive data before exploitation.

Our solution:

- a statistical co-clustering technique for functional data,
- that form homogeneous groups of both individuals and days.

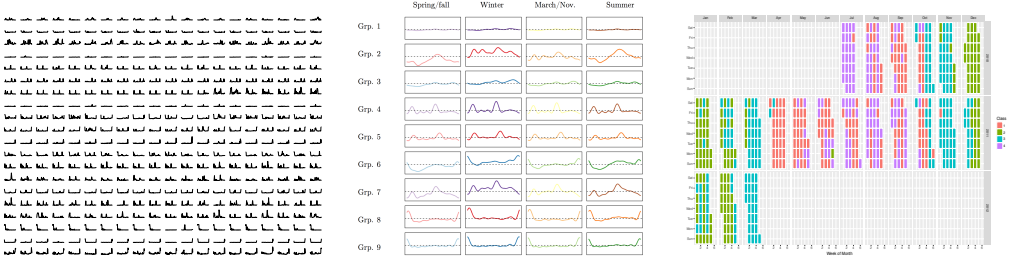


Figure: Functional co-clustering of Linky data (EDF).

Outline of the course

5 first sessions (C. Bouveyron):

- Introduction
- Dimension reduction: PCA, PPCA, FDA
- The GMM and EM algorithms
- Subspace methods for clustering and classification
- Analysis of functional data

5 following sessions (P.-A. Mattei):

- High-dimensional linear regression
- Sparse linear regression and the lasso
- Bayesian linear regression
- Sparsity beyond linear regression 1: classification, recommender systems
- Sparsity beyond linear regression 2: PCA, clustering

Outline

Learning from data...

One task, several families of approaches:

- Statistical learning
- Machine learning
- Deep learning
- ...

Learning from data...

Learning is a two-head problem:

Supervised	Unsupervised

Learning from data...

Methods are specific to each task:

Supervised	Unsupervised

Supervised learning

Supervised learning is also a field with different sub-tasks:

- classification:
- regression:
- time series analysis:
- ...

The supervised learning process

The material: a set of (complete) data

The goal: learn a predictor $f(.)$ from the (complete) data

Measuring the learning performance

One comfortable thing of working in the supervised context is:

- to be able to measure the performance of the learned predictor,

- compare several predictors and pick the most efficient one.

A minimal setup for supervised learning

The minimal setup for building a supervised predictor $f()$ from data is as follows:

Why such a minimal setup?

The goal is to avoid over-fitting when choosing the model or the model parameters:

An advanced setup for supervised learning

Resampling techniques:

- there are several methods (leave-one-out, V-fold cross-validation, bootstrap) depending on the context (sample size, computing time, ...),
- V-fold cross-validation:

Outline

Learning in high-dimensions

Learning in high-dimensions is one of the most important problems nowadays:

A motivating example: cytology

Cytology:

- it is the study of cells in terms of structure, function and chemistry,
- for the diagnosis of disease (we focused on cervical cancer).

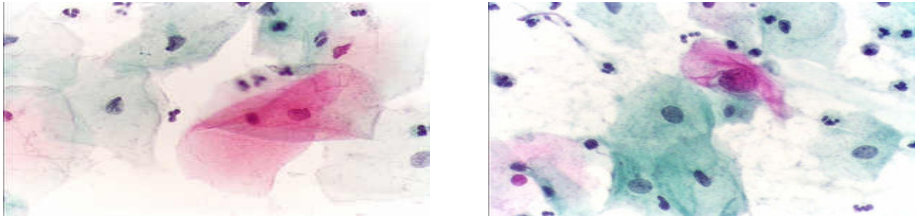


Figure: Normal (left) and abnormal (right) pap smears.

Cervical cancer detection:

- it is an important public health field which is currently treated mostly manually,
- pap smear screening by human experts is complicated by the amount of cells per smear (up to 20 000),
- and by the very small proportion of cancer cells (less than 1%).

A motivating example: cytology

Our data (BC Cancer Agency):

- 20 smears which contains between 4 000 and 10 000 cells,
- each nucleus is described by 111 features (morphological, photometric or texture features),
- only 0.52% of the cells are diseased cells.

Classification is useful in this context:

- for building supervised classifiers which can select the most likely cancer cells,
- for helping experts in labeling the learning data through weakly-supervised classification,
- for selecting discriminative variables which can be used in a semi-automatic process.

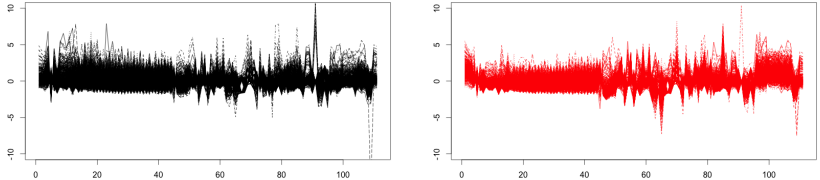
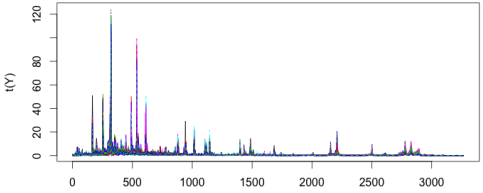


Figure: Control and (cervical) cancer data.

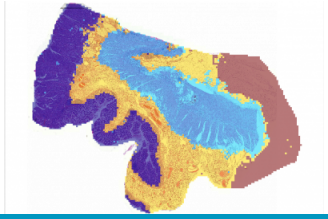
A motivating example: mass spectrometry

Mass spectrometry:

- it is a recent analytical technique that measures the mass-to-charge ratio of charged particles and which aims is to identify the elemental composition of a sample,
- It exist two types of mass spectrometry data:
  - multi-array data which aims to analyze serums or tissue fragments



  - MALDI images which are 2D or 3D MS images of tissues or organs



A motivating example: mass spectrometry

Classification is useful in this context:

- it is used in Medicine for disease diagnostic from blood samples:
  - a supervised classifier is learned from blood samples of healthy and sick patients,
  - the classifier is then used to classify new blood samples.
- a combination of supervised and unsupervised classification can be used to detect errors in the labels

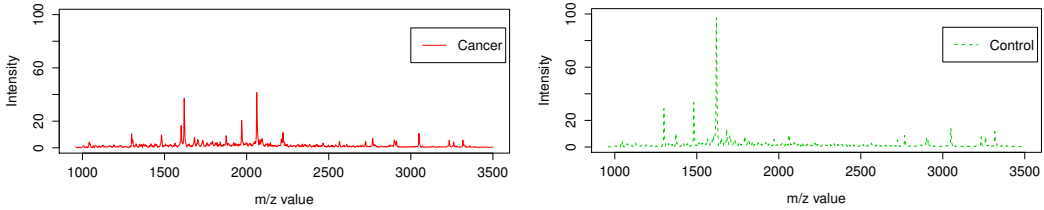


Figure: Control and cancer (colorectal) mass spectrometry spectra.

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 promotate dynamic approaches in programming.

The curse of dimensionality

In the mixture model context:

- the building of the data partition mainly depends on:

$$H_k(x) = -2\log(\pi_k f(x, \theta_k)),$$

- model Full-GMM:

$$H_k(x) = (x - \mu_k)^t \Sigma_k^{-1}(x - \mu_k) + \log(\det \Sigma_k) - 2\log(\pi_k) + \gamma.$$

Consequently:

- it is necessary to invert Σ_k which have a number of parameters proportional to p^2,
- if n is small compared to p^2, the estimates of Σ_k are ill-conditionned or singular and it will be difficult or impossible to invert Σ_k.

The curse of dimensionality

From the estimation point of view:

- let us consider the normalized trace $\tau(\Sigma) = \mathrm{tr}(\Sigma^{-1})/p$ of the inverse covariance matrix Σ^{-1} of a multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$,
- the estimation of τ from a sample of n observations $\{x_1, ..., x_n\}$ conduces to:

$$\tau(\hat{\Sigma}) = \tau(\hat{\Sigma}) = \frac{1}{p}\mathrm{tr}(\hat{\Sigma}^{-1}),$$

$$E[\tau(\hat{\Sigma})] = \left(1 - \frac{p}{n-1}\right)^{-1} \tau(\Sigma).$$

- consequently, if the ratio $p/n \to 0$ when $n \to +\infty$, then $E[\tau(\hat{\Sigma})] \to \tau(\Sigma)$,
- however, if the dimension p is comparable with n, then $E[\tau(\hat{\Sigma})] \to c\tau(\Sigma)$ when $n \to +\infty$, where $c = \lim_{n\to+\infty} p/n$.

The blessings of dimensionality

As Bellman thought:

- all is not bad in high-dimensional spaces (hopefully!)
- there are interesting things which happen in high-dimensional spaces.

First example: volume of the unit sphere is $V(p) = \frac{\pi^{p/2}}{\Gamma(p/2+1)}$,

The blessings of dimensionality

As Bellman thought:

- all is not bad in high-dimensional spaces (hopefully!)
- there are interesting things which happen in high-dimensional spaces.

First example: volume of the unit sphere is $V(p) = \frac{\pi^{p/2}}{\Gamma(p/2+1)}$,

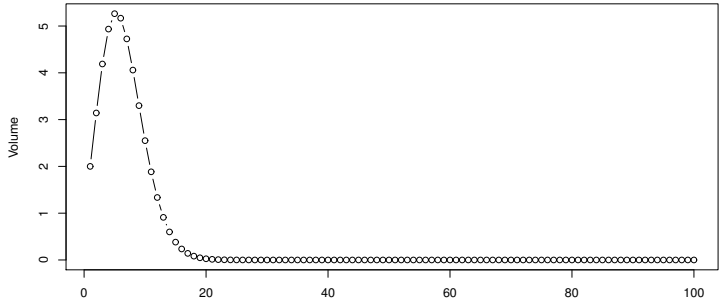


Fig. Volume of a sphere of radius 1 regarding to the dimension p.

The blessings of dimensionality

Second example: probability that a uniform variable on the unit sphere belongs to the shell between the spheres of radius 0.9 and 1 is

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

The blessings of dimensionality

Second example: probability that a uniform variable on the unit sphere belongs to the shell between the spheres of radius 0.9 and 1 is

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

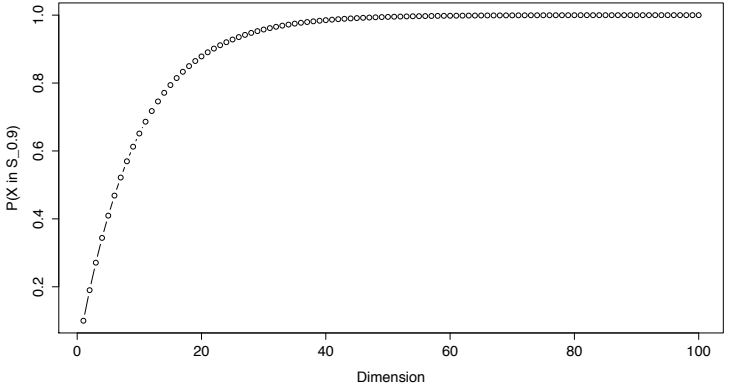


Fig. Probability that X belongs to the shell $S_{0.9}$ regarding to the dimension p.

The blessings of dimensionality

Third example:

- since high-dimensional spaces are almost empty,
- it should be easier to separate groups in high-dimensional space with an adapted classifier,
- a way to observe this is to look at the Bayes classifier behaviour.

The blessings of dimensionality

Third example:

- since high-dimensional spaces are almost empty,
- it should be easier to separate groups in high-dimensional space with an adapted classifier,
- a way to observe this is to look at the Bayes classifier behaviour.

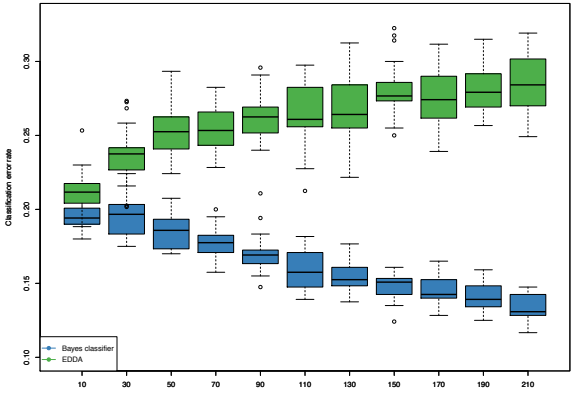


Fig. Classification error rate of the optimal classifier (blue) and EDDA (green) versus the data dimension on simulated data.

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"!

Dimension reduction

A common phantasm about dimension reduction:

- believe that dimension reduction helps for classification,
- this is not true because, most of the time, dimension reduction implies an information loss which would be discriminative.

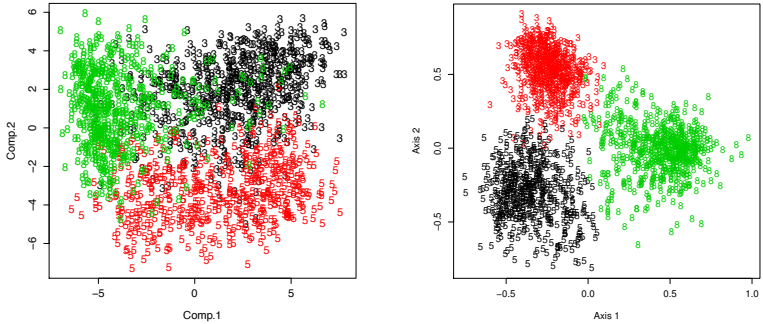


Figure: Projection of the 256-dimensional USPS data with PCA (left, unsupervised) and FDA (right, supervised).

Dimension reduction

Linear dimension reduction methods:

- feature combination: PCA,
- feature selection: ...

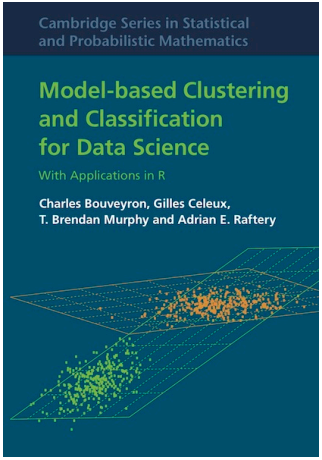
Non linear dimension reduction methods:

- Kohonen algorithms, Self Organising Maps,
- LLE, Isomap, ...
- Kernel PCA, principal curves, ...

Supervised dimension reduction methods:

- the old fashion method: Fisher Discriminant Analysis (FDA),
- many recent works on this topic... but useless in our context.

Want more?



Soon available at Cambridge University Press!