From Conventional Data Analysis Methods to Big Data Analytics

Gilbert Saporta

From conventional data analysis methods to big data analytics

Gilbert Saporta

1. From data analysis to data mining: exploring and predicting

Data analysis here mainly means descriptive and exploratory methods, also known as unsupervised. The objective is to describe as well as structure a set of data that can be represented in the form of a rectangular table crossing *n* statistical units and *p* variables. We generally consider *n* observations as points in *p* dimensional vector space, which if provided with a distance is an Euclidean space. Numerical variables are vectors of an *n* dimensional space. Data analysis methods are essentially dimension reduction methods that are divided into two categories:

– on the one hand, factor methods (principal component analysis for numeric variables, correspondence analyses for category variables) which lead to new numeric variables, combinations of the original variables, allowing representations in low dimensional spaces. Mathematically, these are variants of singular value decomposition of the data table;

– on the other hand, the unsupervised classification methods or clustering which divide observations, or variables, into homogeneous groups. The main algorithms are either hierarchical (step by step construction of the classes by successive clustering of units), or direct partition searches by k-means.

Many works are devoted to previous methods like [SAP 11].

But data analysis is also an attitude which consists of "letting the data speak" by putting no, or at least very little a priori, on the generating mechanism. Let us recall here the principle stated by [BEN 72]: "The model must follow the data, and not the opposite." Data analysis developed in the 1960s and 70s in reaction to the abuses of formalization, see [ANS 67] regarding John Tukey: "He (Tukey) seems to identify statistics with the grotesque phenomenon generally known as mathematical statistics and find it necessary to replace statistics by data analysis."

Data mining, a movement which began in the 1990s at the intersection of statistics and information technologies (databases, artificial intelligence, machine learning, etc.), also aims at discovering structures in large data sets and promotes new tools, such as association rules. The metaphor of data mining means that there are treasures or nuggets hidden under mountains of data that can be discovered with specialized tools. Data mining is a step in the knowledge discovery process, which involves applying data analysis algorithms. [HAN 99] defined it thus: "I shall define data mining as the discovery of interesting, unexpected, or valuable structures in large data sets." Data mining analyzes data collected for other purposes: It is often a secondary analysis of

databases, designed for the management of individual data, and where there is no concern to effectively collect data (surveys, experimental designs).

Data mining also seeks to find predictive models of a *Y* denoted response, but from a very different perspective than that of conventional modeling. A model is nothing more than an algorithm and not a representation of the mechanism that generated the data. One then proceeds by exploring a set of linear or non-linear algorithms, explicit or not, in order to select the best, that is the one that provides the most accurate forecasts without falling into the overfitting trap. We distinguish regression methods where *Y* is quantitative, supervised classification methods (also called discrimination methods) where *Y* is categorical, most often with 2 modalities. Massive data processing has only reinforced the trends already present in data mining.

2. Obsolete approaches

Inferential statistics were developed in a context of scarce data, so much so that a sample of more than 30 units was considered large! The volume of data radically changes the practice of statistics. Here are some examples:

– any deviation from a theoretical value becomes "significant." Thus a correlation coefficient of 0.01 calculated between two variables on a million observations (and even less, as the reader will easily verify) will be declared significantly different from zero. Is it a useful result?

– the confidence intervals of the parameters of a model become zero width since the latter is generally in $1/\sqrt{n}$. Does this mean that the model will be known with certainty?

– in general, there is no longer a generative model that applies to a large amount of data no more than the rules of choice of model by penalized likelihood that are the subject of so many publications.

It should be noted that the criteria of the type:

$$AIC = -2ln(L) + 2k \quad [2.1]$$

and: $$BIC = -2ln(L) + ln(n)k \quad [2.2]$$

to choose between simple models where *k* is the number of parameters and *L* the likelihood, become ineffective when comparing predictive algorithms where neither the likelihood nor the number of parameters are known, as in decision trees and more complex methods discussed in the next chapter. Note that it is illogical, as is often seen, to use *AIC* and *BIC* simultaneously since they come from two incompatible theories: Kullback-Leibler information for the first, Bayesian choice of models a priori equiprobable for the second.

The large volume of data could be an argument in favor of the asymptotic properties of *BIC*, if it were calculable, since it has been shown that the probability of choosing the true model tends to 1 when the number of observations tends to infinity. The true model, however, must be part of the family studied, and especially that this "true" model exists, which is fiction: a model (in the generative sense) is only a simplified representation of reality. Thirty years ago, well before we talked about big data, George Box declared "All models are wrong, some are useful."

The abuses of the so-called conventional statistics had been vigorously denounced by John Nelder [NEL 85], co-inventor of generalized linear models, in this 1985 text discussing Chatfield's article: "Statistics is intimately connected with science and technology, and few mathematicians have experience or understand the methods of either. This I believe is what lies behind the grotesque emphasis on significance tests

in statistics courses of all kinds; a mathematical apparatus has been erected with the notions of power, uniformly most powerful tests, uniformly most powerful unbiased tests, etc. etc. and this is taught to people, who, if they come away with no other notion, will remember that statistics is about significant differences […]. The apparatus on which their statistics course has been constructed is often worse than irrelevant, it is misleading about what is important in examining data and making inferences."

3. Understanding or predicting?

The use of learning algorithms leads to methods known as "black boxes" that empirically show that it is not necessary to understand in order to predict. This fact, which is disturbing for scientists, is explicitly claimed by learning theorists, such as [VAP 06] who writes "Better models are sometimes obtained by deliberately avoiding to reproduce the true mechanisms."

[BRE 01] confirms this in his famous article of Statistical Science entitled Statistical Modeling: The Two Cultures: "Modern statistical thinking makes a clear distinction between the statistical model and the world. The actual mechanisms underlying the data are considered unknown. The statistical models do not need to reproduce these mechanisms to emulate the observable data." Breiman thus contrasted two modeling cultures in order to draw conclusions from data: one assumes that data is generated by a given stochastic model, the other considers the generating mechanism as unknown and uses algorithms.

In the first case, attention is paid to fitting the model to the data (goodness of fit) and in the second, focus is on forecast accuracy.

[DON 15] recently took up this discussion by talking of generative modeling culture and predictive modeling culture. The distinction between models for understanding and models for predicting was also explicit in [SAP 08] and [SHM 10].

4. Validation of predictive models

The quality of a forecasting model can not be judged solely by the fact that it appropriately fits to the data: it has to provide good forecasts in the future, what is called the capacity of generalization. Indeed, it is easy to see that the more complex a model, for example, a higher degree polynomial, the better it will fit to the data, until it passes through all points, but this apparent quality will degrade for new observations: this is the overfitting phenomenon.

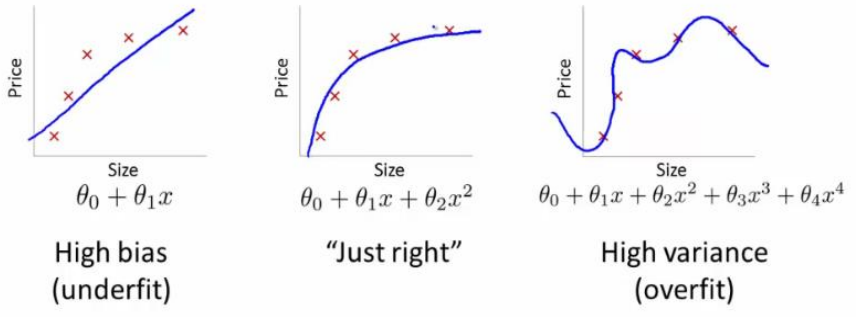

Figure 1. *From underfitting to overfitting (source: available at: http://datascience.stackexchange.com/questions/361/when-is-a-model-underfitted)*

It is therefore appropriate to seek models that behave in a comparable way on available data (or learning data) and on future data. But this is not a sufficient criterion, since for example the constant model: $\hat{y} = c$ verifies this property! Forecasts must also be of good quality.

4.1. *Elements of learning theory*

The inequalities of the learning statistical theory make it possible to find bounds for the difference between learning error and generalization error (future data) according to the number of observations in learning and the complexity of the family of models. Let us illustrate one of these inequalities in the case of supervised classification in two classes. A classifier is then a function of *f(x)* predictors such that if *f(x)> 0* we classify x observation in one group, and if *f(x) <0* in the other group. Points such as *f(x) = 0* define the boundary.

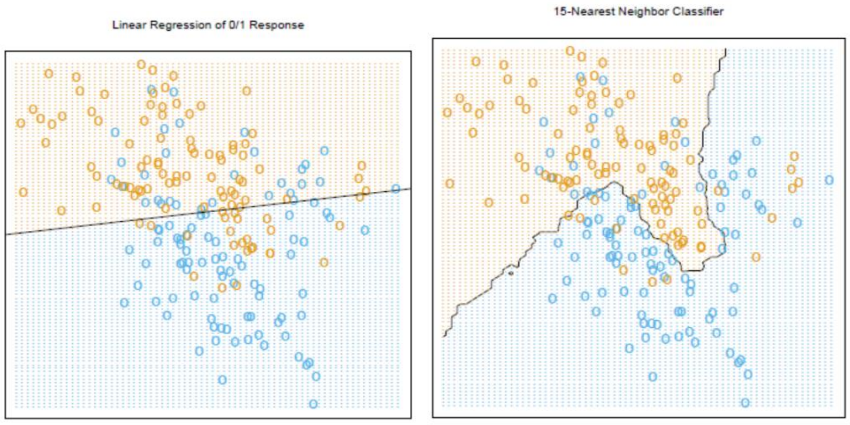

Figure 2. *A linear and nonlinear classifier (according to [HAS 09])*

The classifier error rate, which is a random variable because it depends on the sample, is the proportion of wrongly classified observations. Its expectation is called empirical risk, and denoted R_{emp}. For future observations coming from the same unknown distribution, it will be denoted *R*. Let us consider families of classifiers, such as fixed degree *d* polynomial functions, with or without constraints on the coefficients, or that of the *k*-nearest neighbors (we allocate to the majority class among the *k* neighbors of a "member"). The learning theory has shown that the complexity of these models does not depend on the number of parameters, but on the ability to separate points by the boundary *f(x):* it is VC-dimension or Vapnik-Cervonenkis dimension denoted *h* thereafter. For example, the linear boundaries of $\mathbb{R}^p$ allow us to separate $p+1$ points belonging to different groups but not $p+2$ points: more precisely, there are always configurations of $p+2$ non-separable points, even if there are sometimes configurations of $p+1$ non-separable points. VC dimension is $h=p+1$.

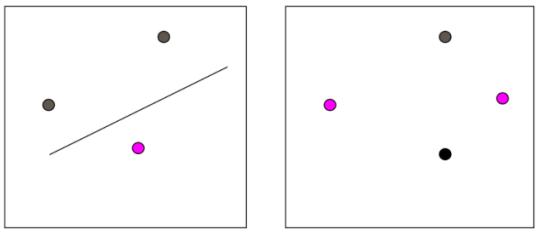

Figure 3. *In the diagram, there are still configurations of 4 non-separable points*

One of the most famous inequalities states that, with a probability 1-α:

$$R < R_{\text{emp}} + \sqrt{\frac{h\left(\ln\left(2n/h\right)+1\right)-\ln\left(\alpha/4\right)}{n}} \qquad [2.3]$$

For fixed n, the increase of h leads R_{emp} to 0 (overfitting) but the radical increases thus the existence of an optimal complexity h^*.

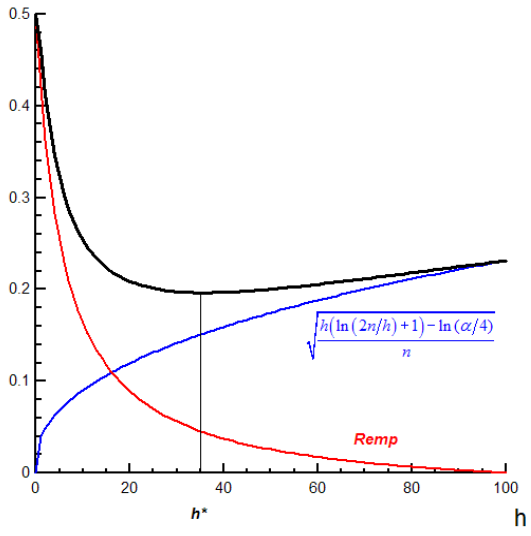

Figure 4. *Optimal VC-dimension*

It should be noted that the gap between empirical risk and risk depends only on the ratio n/h and that if n is increased faster than h, there is convergence. This result shows that the more data we have, the more complex models we can use.

The statistical learning theory abounds with such inequalities, but unfortunately they are not very convenient in practice to choose a model because VC dimension is difficult to obtain. Cross validation methods are therefore indispensable: they consist of setting aside one or more parts of the data in order to simulate the behavior of algorithms or models in the presence of future data.

We must strongly reiterate that the validation of a model or algorithm in big data can only be carried out on "new" data, which make it possible to ensure the reproducibility of results. This is an essential difference from standard statistical practice, although some so-called leave one out methods have been used for a long time in discrimination. Nevertheless, removing an observation when n is large is of little effect.

4.2. *Cross validation*

To choose between several models or algorithms, the practice involves randomly dividing the available data into three subsets including learning, validation and test. Typical values for the proportions of these three subsets are 50%, 25%, 25% [HAS 09]. The learning set is used to estimate the parameters of (or to calibrate) each model. Each of the models is then applied to the validation set to select the best according to the criterion chosen (R^2, misclassification rate, etc.). The best model is then applied to the test set to estimate its performance, which is overvalued in the previous phase since one takes the *sup* of a set. We thus distinguish the evaluation of the performance

of a model, from the choice of this model. Once the model is chosen, it must be re-estimated using all available data before putting it into production.

Ideally, in order to avoid risks due to random splitting in learning, validation and test, it would be necessary to iterate this step, but this is not done for very large datasets. For small size sets, it will be preferable to subdivide the set into 5 or 10 parts of equal number: in a rotating manner, a model is estimated by removing one of the 5 or 10 parts (5 or 10-fold cross validation) and evaluating its performance on the part set side and then averaging the results.

5. Combination of models

Rather than choosing the best among M models or algorithms, it is usually much more efficient to combine them. We then talk of ensemble methods; boosting, bagging, random forests fall into this category, but only combine classifiers or regressors of the same family as trees. The same is true of Bayesian model averaging, which linearly combines submodels of the same family, with as coefficients the posterior probabilities of each model knowing the data. While remaining faithful to data analysis principles, we will not discuss Bayesian model averaging which requires constraining hypotheses in order to be applied.

A particularly well suited method for massive data is stacking, which has yielded excellent results in machine learning competitions, the most famous of which is the one million dollars Netflix prize. In 2009, the two best solutions combined numerous models according to the stacking technique introduced by [WOL 92] and [BRE 96]. Let's start with the context of regression. Let us consider M predictions: $\hat{y}_m = f_m(\mathbf{x}) \quad m = 1,..,M$ obtained using M models or different algorithms, which could be of any type: linear or non-linear, neural networks, regression trees, etc. The very simple idea is to look for a linear combination: $\hat{y} = \sum_{m=1}^{M} w_m f_m(x)$ which provides a sum of squared minimum errors. In the original version, to avoid that the more complex models have more weight because they predict better in learning, the criterion is modified so that the predictions of each y_i are done by removing observation i (predicted residuals):

$$\min \sum_{i=1}^{n} \left(y_i - \sum_{m=1}^{M} w_m f_m^{-i}(x_i) \right)^2 \qquad [2.4]$$

but when n is large, it has little impact.

On the other hand, as shown by [NOC 16], the estimation of weights w_i is made unstable by the fact that the predictions of the different models are highly correlated with one another as soon as these models are efficient. It is therefore necessary to regularize the least squares. One possibility is to carry out a regression of y on m predictions without constant term, under the constraint that weights w_i are positive and of sum equal to 1, as in Bayesian model averaging. A simpler solution is to carry out a PLS regression (see section 2.6.1.2): As the M predictions are positively correlated, a single PLS component is generally sufficient, and ensures the positivity of weights.

Extension to supervised classification is carried out while taking for $\hat{y}_m$ value the probability of belonging to the class of interest. Since the y_i are binary, we will use a PLS logistic regression instead of a PLS regression to estimate the weights.

Extensions of predictors to geometric means have been proposed, as well as the search for areas of competence of each predictor or combinations of some of them. But

in practice stacking proves to be very effective because by construction the optimal linear combination of *M* predictions is necessarily better than each of them.

6. The high dimension case

The data may also be massive in the sense that *p* the number of variables is much greater than *n* the number of observations. This is the case for data from the web or biology, where it is not uncommon to count several thousand of variables. Predictive methods of regression type can not be applied when *p>>n*, since the least square estimator does not exist. If we want to preserve all the predictors, we will resort to regularization methods, if not to sparse methods.

6.1. *Regularized regressions*

They proceed either by projection onto subspaces, or by constraining the coefficients vector. The estimators are biased and properties invariant under change of scale are lost. The data will be centered and reduced prior to the application of methods

6.1.1. *Principal components regression*

This is undoubtedly the oldest method, applied in econometrics by Edmond Malinvaud since 1964 to solve multicollinearity problems. It involves reducing predictors space by using $q < p$ principal components and then regressing *Y* response on these components by ordinary least squares. The principal components being linear combinations of predictors, we ultimately obtain a combination of predictors:

$$\hat{\mathbf{y}} = \mathbf{C}\hat{\boldsymbol{\alpha}} = \alpha_1\mathbf{c}_1 + ... + \alpha_q\mathbf{c}_q = \mathbf{X}\hat{\boldsymbol{\beta}} \qquad [2.5]$$

Coefficient $\hat{\boldsymbol{\alpha}}$ et $\hat{\boldsymbol{\beta}}$ vectors are obtained simply by using the reconstruction formula q (truncated SVD) $\mathbf{X} = \mathbf{C}\mathbf{U}'$ where C is the principal components matrix and U is the principal factors orthogonal matrix:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{+}\mathbf{X}'\mathbf{Y} = \left(\mathbf{U}\mathbf{C}'\mathbf{C}\mathbf{U}'\right)^{+}\mathbf{U}\mathbf{C}'\mathbf{y} = \left(\mathbf{U}\frac{\mathbf{1}}{\mathbf{n}}\mathbf{C}'\mathbf{C}\mathbf{U}'\right)^{+}\frac{\mathbf{1}}{\mathbf{n}}\mathbf{U}\mathbf{C}'\mathbf{y} \\ &= \left(\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}'\right)^{+}\frac{\mathbf{1}}{\mathbf{n}}\mathbf{U}\mathbf{C}'\mathbf{y} = \mathbf{U}\boldsymbol{\Lambda}^{+}\mathbf{U}'\mathbf{U}\frac{\mathbf{1}}{\mathbf{n}}\mathbf{C}'\mathbf{y} = \mathbf{U}\boldsymbol{\Lambda}^{+}\frac{\mathbf{1}}{\mathbf{n}}\mathbf{C}'\mathbf{y} = \mathbf{U}\hat{\boldsymbol{\alpha}}\end{aligned} \qquad [2.6]$$

The symbol + refers to the Moore-Penrose inverse.

Where:

$$\hat{\boldsymbol{\beta}} = \mathbf{U}\hat{\boldsymbol{\alpha}} \quad \hat{\boldsymbol{\alpha}} = \mathbf{U}'\hat{\boldsymbol{\beta}} \qquad [2.7]$$

And:

$$V(\hat{\beta}_j) = \sigma^2\sum_{k=1}^{q}\frac{u_{jk}^2}{\lambda_k} \qquad [2.8]$$

In general, *q* is selected by cross validation, but the regression on principal components has the following drawback: the principal components depend only on the predictors and not on the response, and their ranking does not necessarily reflect the correlations with this response.

6.1.2. *PLS regression*

Developed by H. and S. Wold, PLS regression resembles principal components regression, since data are also projected onto linear uncorrelated combinations of predictors. The main difference is that the PLS components are optimized to be also predictive of *Y*, whereas the principal components only extract the maximum variance of predictors without taking *Y* into account. The criterion used to obtain the first PLS component **t=Xw** is Tucker's criterion:

$$\max_{\mathbf{w}} \operatorname{cov}^2(\mathbf{y}, \mathbf{Xw}) \qquad [2.9]$$

As:

$$\operatorname{cov}^2(\mathbf{y}, \mathbf{Xw}) = r^2(\mathbf{y}, \mathbf{Xw}) V(\mathbf{y}) V(\mathbf{Xw}) \qquad [2.10]$$

we have a compromise between maximizing the correlation between **t** and y (regression) and maximizing the variance of **t** (PCA of predictors).

The solution is as follows: for the first PLS component, the w_j coefficient of each variable is, up to a multiplicative constant, equal to the covariance between $\mathbf{x_j}$ and **y**, which ensures the consistency of signs. The following components are obtained by deflation, that is, by iterating the process on the residuals of *Y* and predictors after regression on **t**. The simplicity of the algorithm, which requires neither diagonalization nor matrix inversion, makes it possible to process massive data. We will refer to [TEN 98] for more details.

6.1.3. *Ridge regression*

Invented by Hoerl and Kennard in the 1970s, this is a particular case of Tikhonov regularization: to avoid unstable coefficients, we add a constraint on their norm:

$$\min \|y - \mathbf{X\beta}\|^2 \text{ under } \|\boldsymbol{\beta}\|^2 \leq c^2 \qquad [2.11]$$

This is equivalent to adding a constant to the diagonal elements of **X'X** to "facilitate" the inversion:

$$\hat{\boldsymbol{\beta}}_R = (\mathbf{X'X} + k\mathbf{I})^{-1} \mathbf{X'y} \qquad [2.12]$$

The constant *k* is determined by cross validation.

6.2. *Sparse methods*

The preceding methods make it possible to obtain a function of all the variables, which becomes a disadvantage when *p* is very large: how can a linear combination of several hundred or several thousands of variables be interpreted? Rather than resorting to stepwise selection techniques, the use of constraints in L_1 norm, effectively solves the problem by enabling both selection and regularization.

6.2.1. *The Lasso*

Lasso or Least absolute shrinkage and selection operator introduced in [TIB 96] consists of minimizing the residual sum of squares, with a bound on the sum of the absolute values of regression coefficients (L_1 penalty):

$$\min \|\mathbf{y} - \mathbf{X\beta}\|^2 \quad \text{avec} \sum_{j=1}^{p} |\beta_j| < c \qquad [2.13]$$

which is equivalent to:

$$\min\left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right) \quad [2.14]$$

When *c* decreases, the regression coefficients reduce and some are canceled due to the use of the L_1 norm. The parameter *c* is generally obtained by cross validation, with the aim of having the best predictor of *Y*.

Many developments followed: sparse variants of PLS regression, and also the group-lasso which applies in the case where the predictors are divided into blocks: the method then helps to eliminates entire blocks of variables.

6.2.2. *Sparse Principal Component Analysis (PCA) and Multiple Correspondence Analysis (MCA)*

In the same vein, sparse versions of principal component analysis have been proposed since the 2000s. There are several versions, but the most widely used is inspired by the Lasso and ridge regression, noting that the SVD can be interpreted as a ridge regression of components on the variables because the main factors are bounded. We obtain "components" that are combinations of a small number of initial variables, which facilitates interpretation, but at the expense of the loss of orthogonality properties of the components and / or factors.

[BER 12] developed a sparse version of multiple correspondence analysis as follows: the MCA being a PCA of blocks of indicators, the authors adapted the group Lasso to sparse PCA previously defined.

7. The end of science?

Big data processing requires new tools (we have briefly presented some), and a new attitude towards models that are just algorithms, based on validation with data set aside.

These new tools can be useful to specialists in a field, as [VAR 14] advises to econometricians.

In a provocative article [AND 08] claimed that the data deluge renders the scientific approach obsolete and declared in essence that: "correlations are enough, we can stop modeling. Let us load the data in larger computers and allow statistical algorithms to find structures where science can not."

It is clear that correlation is not causality: a model that accurately predicts statistically does not necessarily allow for action. It is too often believed that the influence of a variable can be measured by its coefficient in the simple case of a linear model, or by elimination in complex cases: as in sensitivity analysis, we study the variation of a quality criterion (R^2, % of accurately classified observations, etc.) by removing the variable considered. This may be interesting but is still insufficient: On the one hand, to vary a variable "other things being equal" is an illusion, for the modification of a variable can entail modifications on those that are correlated to it, and thus on the response. On the other hand, without a pattern of causality one can not know how the other variables react in view of an intervention.

If we can often forecast without understanding, can we not better forecast if we understand? This subject is discussed in numerous meetings and studies by learning and causality specialists which introduce experimentation in big data, see [BOT 13].

8. Bibliography

[AND 08] ANDERSON C., « The End of Theory : the Data Deluge makes the Scientific Method Obsolete », *Wired Magazine*, 2008. available at : https://www.wired.com/ 2008/06/pb-theory/, last visited on April, 15, 2017.

[ANS 67] ANSCOMBE F.J., « Topics in the investigation of linear relations », *J. Roy. Stat. Soc.*, vol. B 29, p. 1-52, 1967.

[BEN 72] BENZÉCRI J.P., *L'Analyse des Données*, *tome 2*, Dunod, Paris, 1972.

[BER 12] BERNARD A., GUINOT C., SAPORTA G., « Sparse principal component analysis for multiblock data and its extension to sparse multiple correspondence analysis », dans A. COLUBI (ED.), *Compstat Proceedings*, p. 99-106, 2012.

[BOT 13] BOTTOU L. *et al.*, « Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising », *Journal of Machine Learning Research*, n° 14, p. 3207-3260, 2013.

[BRE 96] BREIMAN L., « Stacked Regressions », *Machine Learning*, n° 24, p. 49-64, 1996.

[BRE 01] BREIMAN L., « Statistical modeling: The two cultures », *Statistical Science*, vol. 16, n° 3, p. 199-215, 2001.

[DON 15] DONOHO D., « 50 years of Data Science », Tukey Centennial workshop. available at : http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf, 2015, last visited on April, 15, 2017.

[HAN 99] HAND D., « Why data mining is more than statistics write large », *ISI*, Helsinki. available at : http://www.stat.fi/isi99/proceedings/arkisto/varasto/hand0490.pdf, 1999, last visited on April, 15, 2017..

[HAS 09] HASTIE T., TIBSHIRANI R., FRIEDMAN J., *The Elements of Statistical Learning*, 2e edition, Springer, New York, 2009.

[NEL 85] NELDER J.A., discussion de Chatfield C., « The initial examination of data », *Journal of the Royal Statistical Society*, Series A, n° 148, p. 214-253, 1985.

[NOC 16] NOÇAIRI H., GOMES C., THOMAS M., SAPORTA G., « Improving Stacking Methodology for Combining Classifiers; Applications to Cosmetic Industry », *Electronic Journal of Applied Statistical Analysis*, vol. 9, n° 2, p. 340-361, 2016.

[SAP 08] SAPORTA G., « Models for Understanding versus Models for Prediction », dans I. BRITO (ED.), *Compstat Proceedings*, Physica Verlag, p. 315-322, 2008.

[SAP 11] SAPORTA G., *Probabilités, Analyse des données et statistique*, 3e édition., Technip, Paris, 2011.

[SHM 10] SHMUELI G., « To explain or to predict? », *Statistical Science*, n° 25, p. 289-310, 2010.

[TEN 98] TENENHAUS M., *La régression PLS*, Technip, Paris, 1998.

[TIB 96] TIBSHIRANI R., « Regression shrinkage and selection via the Lasso », *Journal of the Royal Statistical Society*, Series B, n° 58, p. 267-288, 1996.

[VAP 06] VAPNIK V., *Estimation of Dependences Based on Empirical Data*, 2e édition, Springer, New-York, 2006.

[VAR 14] VARIAN H., « Big Data: New Tricks for Econometrics », *Journal of Economic Perspectives*, n° 28, p. 3-28, 2014.

[WOL 92] WOLPERT D., « Stacked Generalization », *Neural Networks*, n° 5, p. 241-259, 1992.