TEST (2019) 28:289–329
https://doi.org/10.1007/s11749-019-00651-9

INVITED PAPER

Data science, big data and statistics

Pedro Galeano[1] · **Daniel Peña**[1]

Published online: 8 April 2019

Abstract
This article analyzes how Big Data is changing the way we learn from observations. We describe the changes in statistical methods in seven areas that have been shaped by the Big Data-rich environment: the emergence of new sources of information; visualization in high dimensions; multiple testing problems; analysis of heterogeneity; automatic model selection; estimation methods for sparse models; and merging network information with statistical models. Next, we compare the statistical approach with those in computer science and machine learning and argue that the convergence of different methodologies for data analysis will be the core of the new field of data science. Then, we present two examples of Big Data analysis in which several new tools discussed previously are applied, as using network information or combining different sources of data. Finally, the article concludes with some final remarks.

Keywords Machine learning · Sparse model selection · Statistical learning · Network analysis · Multivariate data · Time series

Mathematics Subject Classification 62A01 · 62H99

This research has been supported by Grant ECO2015-66593-P of MINECO/FEDER/UE.

This invited paper is discussed in comments available at: https://doi.org/10.1007/s11749-019-00639-5, https://doi.org/10.1007/s11749-019-00640-y, https://doi.org/10.1007/s11749-019-00641-x, https://doi.org/10.1007/s11749-019-00642-w, https://doi.org/10.1007/s11749-019-00643-9, https://doi.org/10.1007/s11749-019-00644-8, and https://doi.org/10.1007/s11749-019-00646-6, https://doi.org/10.1007/s11749-019-00647-5, https://doi.org/10.1007/s11749-019-00648-4.

✉ Pedro Galeano
pedro.galeano@uc3m.es

Daniel Peña
daniel.pena@uc3m.es

[1] Departamento de Estadística and Institute of Financial Big Data, Universidad Carlos III de Madrid, 28903 Getafe, Madrid, Spain

1 Introduction

At the end of last century the two main ingredients of the digital society were created: the World Wide Web in the CERN in Geneva, that allows fast and simple communication in the internet, and the smart phones in USA, which offer high computing power and new ways of receiving and transmitting information to an increasing number of persons. Both advances have modified the way we work, how we interact with others and the use of our free time. They have made people generators of social data that have been found to be of economic value in many different environments. Social networks, and surfing in the www using the smart phones, are producing large amount of information that adds to the huge data banks generated in an automatic way by sensors monitoring industrial, commercial or services, activities. For the first time in history we have data everywhere, the now called Big Data. These data are a mixture of structured and unstructured information; they grow exponentially and are produced with very small cost. Also, the cost of storing data is continuously decreasing and the speed of processing is growing very fast.

Statistics as a scientific discipline was created in a complete different environment. The statistical methods that are still taught today were developed for a world in which data were very scarce, and we have to supplement this lack of information with models based on simplifying assumptions in order to draw conclusions from small data sets. Merging experimental data and casual statistical models has been the backbone of the scientific method to advance our knowledge in many disciplines. However, the main paradigm in statistics, we have a random sample from some population, and we want to use this sample to make inference about the parameters of the population, is not well suited to the new problems we face today: large heterogeneous databases sometimes unstructured, which may include texts, images, videos, or sounds, from different populations and as many (or even more) variables than observations. Also, the standard way of comparing methods of inference in terms of efficiency is not very relevant when the data coincide almost with the whole population. On the other hand, the idea of robustness become increasingly important, although in a more broad meaning that is normally used in standard robust statistics. The computing capabilities, that imposed a strong limitation for many years in the development of statistical methods, have increased so much that many of the usual assumptions are no longer needed. For instance, the hypothesis of linearity is seldom true in large data sets, and it is not required with the estimation power of nowadays computers. Also, new criteria should be used when the number of variables is larger than the number of observations. Finally, we need automatic procedures able to extract the information in large and dynamic contexts in which the data are produced continuously.

Several works have analyzed the changes that this Big Data world is producing in statistical data analysis. Efron and Hastie (2016) is an excellent reference on these changes; see also Bühlmann and van de Geer (2011) for the analysis of high-dimensional data. Fan et al. (2014) includes an interesting presentation of several statistical procedures that are not longer optimal with Big Data and discusses, among other problems, the effect of endogeneity, that is usually forgotten in standard statistical analysis. Chen and Zhang (2014) presents an overview of these problems, mostly from the computer science perspective. Donoho (2017) analyses data science

as a new area that includes statistics but has a broader perspective. Hall et al. (2005) were among the first that used asymptotics in which the dimension tends to infinity, while the sample size remains fixed, and found a common structure underlying many high-dimension, low-sample-size data sets. Other useful references on this field are Bühlmann et al. (2016), Bühlmann and van de Geer (2018), Cao (2017), Dryden and Hodge (2018), Gandomi and Haider (2015), Härdle et al. (2018), Peña (2014), Riani et al. (2012), and Torrecilla and Romo (2018), among many others.

The article is organized as follows. In the next section, we discuss how Big Data is changing statistics and modifying the way we learn from data. This is illustrated by discussing seven areas which have been shaped by the use of increasingly large and complex data sets. Our approach complements in some topics to Fan et al. (2014), and includes others, as network models, that are not discussed in previous works. Next, we compare the statistical approach with these of computer science and machine learning, argue that the new Big Data problems are a great opportunity to expand the scope of statistical procedures, and discuss the emergence of data science as the field that studies all the steps in data analysis with a convergence of different methodologies. In this part we supplement the excellent article on data science by Donoho (2017), emphasizing aspects that are not considered in his review of this field. Then, we present two examples of Big Data analysis in which several of the ideas discussed in the previous section are used to provide new data insights. Finally, the article concludes with some final remarks.

2 Changes in statistics for big data

We have selected seven areas in which the availability of increasing large and complex data sets have changed the traditional statistical approach. Also, these fields are expected to be transformed further for the new opportunities provided by Big Data. They are: (1) analyzing new sources of information: texts, images, videos, audios and functions; (2) data visualization in high dimensions; (3) analyzing heterogeneous data; (4) multiple hypothesis testing and the false discovery rate; (5) automatic procedures for model selection and statistical analysis; (6) estimation procedures in high dimension with sparse models; (7) analyzing networks and incorporating network information into statistical models.

2.1 Analyzing new sources of information: texts, images, videos, audios and functions

Until very recently, in statistics data were a set of observed values of one or several variables. The values can represent a sample at a given time, a sequence over time of one or several time series, or a sequence of spatial data in different locations. It is assumed that these data are represented by numbers (for numerical variables) or letters (for attribute variables), and are summarized in a table or in a matrix. Also, it is assumed that the data have been recorded by some given objective, and they represent a sample from some well-defined population. However, now, data are often generated in

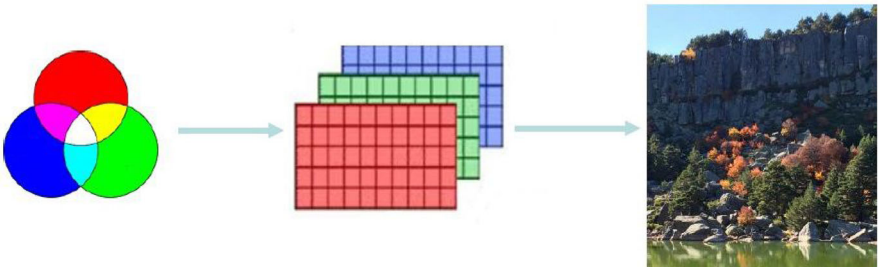

Fig. 1 Process of transforming pixels in images into images

an automatic way, by sensors, webs, social networks, and other devices with different frequency and periodicity, and include not only structured variables, as before, but also texts, images, videos, audios or functions, that should also be considered as data. Thus, a central problem is combining information from different sources. For instance, in Medicine, we want to merge all sorts of information coming from different hospitals and care units in order to learn at a faster rate from new illnesses. In fact, many of the huge advances in clinical treatment in the last years are mostly due to this process of combining many sources of information.

The analyses of text data has a long tradition in statistics. For instance, Mosteller and Wallace (1963) used text analysis to decide the authorship of the disputed Federalist Papers in USA. However, the large textual information in social networks and in the web, the advances in speech recognition and the increase in computer power have led to the computerized text analysis of natural language and the research field of sentiment analysis (see Tausczik and Pennebaker 2010). Sentiment analysis deals with the computational treatment of opinions and sentiments, considering people subjectivity as an important field of empirical research. This area is being mostly developed in the computer science literature, although using many tools of classical multivariate analysis, such as discrimination, clustering or multidimensional scaling. In addition, the merging of information coming from sentiment analysis with network information (see Sect. 2.7) is a powerful tool for social science analysis, see, for instance, Pang and Lee (2008).

The second type of new data we discuss are images. The digital computer made possible incorporating images as new sources of information. It was known since the second half of the 19th century (due to Young and Helmholtz) that any color can be well represented by merging three monocolor filters: red, green and blue, the RGB representation. This idea was incorporated in the computers in the 1990s to manage colors, with the enhanced graphics adapter (EGA), that represents an RGB image by three matrices of numbers (pixels) that when combined produce the image, as indicated in Fig. 1. This representation opens the way to image analysis, initiated in the field of computer science in groups of artificial intelligence and robotics in USA, mostly with medical applications. The advances in this field in the 1980s are presented in Jain (1989). A pioneering work of statistical analysis of images was Besag (1986), but the most important developments in this field, as computer vision, have appeared outside statistics. Only recently, statisticians are considering images as a new source of useful data for statistical analysis. See, for instance, Lu et al. (2014).

A similar situation occurs with video analysis. Videos can be seen as images collected through time and are frequently used in diverse areas including climatology, neuroscience, remote sensing, and video surveillance, among others. Due to the dynamic nature of videos, change point detection is an important problem in video analysis. For instance, we can be interested in detecting the birth of a hurricane, the presence of brain activity, or the presence of a thief in a building. Radke et al. (2005) present a survey of the common processing steps and core decision rules in change point detection algorithms for videos. Dimension reduction techniques are frequently used for video compression, see, for instance, Majumdar (2009). Clustering methods are used to motion segmentation and face clustering problems in computer vision, see, for instance, Vidal (2011).

Audio analysis has been mostly developed in electric engineering, often using statistical ideas as, for instance, hidden Markov models in speech recognition, see Rabiner (1989). Some nonlinear time series research have used time series of sounds as examples for modeling, but the advances in this field have not stimulated research published in statistical journals. Some exceptions are Bailey et al. (1998) and Irizarry (2001), among others. More recently, Pigoli et al. (2018) used a time–frequency domain approach to explore differences between spoken Romance languages using acoustic phonetic data.

Both images and audio signals have been recently part of the interest of functional data analysis, a field of statistics that has grown fast in the last two decades. Functional data arises when the variables of interest can be naturally viewed as smooth functions. For instance, sensors measuring human vital signals, such as body temperature, blood pressure, and heart and breathing rates, or human movements, such as hip and knee angles, are able to provide almost continuous measurements of all theses quantities. This leads to data sets of several terabytes, such as those encountered in fMRI (functional magnetic resonance imaging). It is true that functional data samples have an inherent limitation since the functions can only be observed at discrete grids. However, smoothing techniques are able to reproduce the unobserved functions that allow researchers to use the underlying infinite-dimensional and functional characteristics of the data. There is an increasing amount of methods developed for functional data sets, including dimension reduction techniques for handling the infinite-dimensional data structures, regression models in which the predictors and/or the response are functional, supervised and unsupervised classification methods, and functional time series, among many others. See Ramsay and Silverman (2005), Cuevas (2014) and Kokoszka and Reimherr (2017), for overviews in functional data analysis covering all these aspects, and Shi and Choi (2011) for an overview on regression analysis for functional data.

A traditional way of combining information about variables of different frequency or location is Meta-Analysis (see Brockwell and Gordon 2001), that has had many applications in Medicine and Social research. In other fields, as in Economics, the need of merging information has led to combine time series of different periodicity to improve prediction, a field now called nowcasting, name borrowed from the field of Meteorology (see Giannone et al. 2008). However, new methodologies are needed to combine in an effective way data from new sources, as texts or images, with more conventional sources of data to improve statistical analyses. For instance, Chen et al.

(2014) combine standard information with text information obtained by computerized searching of financial webs, to forecast the stock market. The increasing availability of new information from new sources will stimulate a broader Meta-Analysis methodology. For instance, in marketing research, we may want to combine the classical information we have about a customer with image analysis of his movements in the shop, as recorded by cameras, face analysis of his/her reaction to different stimulus and audios of the conversations between the customer and the shop attendant.

2.2 Data visualization in high dimensions

Visualization of large dimensional data sets is a difficult problem. The most often used approach is to make a graph in which (1) different variables are plotted against categories or (2) the relation between a set of variables is shown by scatter plots of each pair, as done in most popular statistical programming languages such as R or Matlab. A survey restrained to table data is de Oliveira and Levkowitz (2003), mostly from the computer science point of view, and emphasizing cluster analysis results. See also Munzner (2014), that covers visualization of tables, networks, and spatial fields. The use of videos to display quantitative information over time has become very popular after the pioneering work of Hans Rosling. His videos on TED talks (see https://www.gapminder.org/) are wonderful examples of the use of video animation to explain is a simple way complex problems. The area of visuanimation, see Genton et al. (2015), will have an increasing importance. See, for instance, Benito et al. (2017) for a video example of the performance of a classification rule to identify gender.

Two important ideas to display multivariate data are Grand Tour and Projection Pursuit. Grand Tour (Asimov 1985) builds a sequence of low-dimensional projections, like a dynamic movie, of the data. Projection Pursuit tries also to find low-dimensional projections being able to show interesting features of the high-dimensional data by maximizing a criteria of interest. For instance, Peña and Prieto (2001a, b) proposed the kurtosis coefficient as an effective way to find clusters and outliers in high dimensions. Cook et al. (1995) proposed using a sequence of plots that are selected by Projection Pursuit criteria. In large data bases the space of all possible views is extremely large and a way to reduce the space to search is to define the objective that we would like to find. This is the approach followed by Targeted Projection Pursuit, in which the ideal view of the data is specified and the objective is to find a view as close as possible to this objective, see Faith et al. (2006). Criteria for judging different visualization of high-dimensional data are discussed by Bertini et al. (2011).

A useful way to represent data in statistics is to use quantiles. Tukey (1970) introduced the boxplot, based on the sample quantiles of univariate continuous distributions. The extension of this plot for multivariate data requires a definition of the multivariate quantiles and this can be done in many ways because the lack of a canonical ordering in $\mathbb{R}^m$, for $m > 1$. Small (1990) presents a survey of the field and we refer to Chernozhukov et al. (2017) for a recent approach on multivariate quantiles based on measure transportation that includes many references. For functional data, quantiles are related to the concept of depth, and López-Pintado and Romo (2009) introduced a useful way to make this connection. Given a set of m functions $x_i(t)$, where t belongs

to a closed interval in the real line, they define the band of order r (created by a subset of size r of these functions, $2 \leq r \leq m$) as the space between the two functions obtained at each point t by taking the minimum and maximum at t of all the functions in the subset. Then, the (total) band depth of order r for a member of the set, $x(t)$, is defined as the proportion of all the possible bands of order r that includes at all times the function $x(t)$. A (modified) band depth of order r for a member of the set, $x(t)$, is also defined as the average in all the possible bands of order r of the proportions of times in which $x(t)$ is included in each possible band. The band depth leads to a natural ordering of the functions and calling $x_{[1]}(t)$ to the function with the largest depth and $x_{[m]}(t)$ to the one with the smallest value, the sequence $x_{[1]}(t), \ldots, x_{[m]}(t)$ can be treated as order statistics and used to compute quantiles of the data set. This idea was used by Sun and Genton (2011) to propose functional boxplots, that have the median, or deepest function, in the middle, a central band defined by the band formed by the set $\left(x_{[1]}(t), \ldots, x_{[m/2]}(t)\right)$, where $[m/2]$ is the smallest integer equal or greater than $m/2$, and the limits or whiskers and the outliers of the functional boxplot are computed as in the standard one by taking the central band as the interquartile range. These methods have been generalized for images with the surface boxplots; see Genton et al. (2014) and for outlier analysis, see Arribas-Gil and Romo (2014).

Quantiles in time series have had a limited application. For stationary time series, the population quantiles are constant lines with values determined by the common marginal distribution function. For non-stationary time series, the quantiles will be time series that follow the changes in the marginal distributions and are more informative. They can be estimated by locally smoothing as shown by Zhou and Wu (2009). However, quantiles have not been shown to be useful for visualization of large sets of time series. Peña et al. (2019b) proposed to define the empirical dynamic pth quantile for a set of possible non-stationary time series $\mathbb{C} = \{x_{it}, 1 \leq i \leq m, 1 \leq t \leq T\}$ as the series of the set $\mathbb{C}$ that verifies

$$q_t^p = \underset{q_t \in \mathbb{C}}{\operatorname{argmin}} \left[\sum_{t=1}^{T} \left(\sum_{x_{it} \geq q_t} p \, |x_{it} - q_t| + \sum_{x_{it} \leq q_t} (1-p) \, |x_{it} - q_t| \right) \right]. \tag{1}$$

For instance, the empirical median minimizes the L_1 distance to all the series. It is shown that the minimization of (1) is equivalent to finding the time series in the set, x_{jt}, that is as close as possible to the pointwise quantiles, q_t^{*p}, in some weighted L_1 metric:

$$\sum_{t=1}^{T} c_{jt} \; \left| q_t^{*p} - x_{jt} \right| . \tag{2}$$

Note that solution of (1) grows with m^2 but the one of (2) is linear in m. Thus we can compute empirical dynamic quantiles for large sets of time series and use these quantiles to make plots of the series. For instance, Fig. 2 shows the stock prices of the 99 most important markets in the world. This plot is not useful to see the general structure of the set. In Fig. 3 we see the plot of the three quartiles of the set of time series, which give a more useful idea of the general evolution of the set of time series. These three quartiles are also plotted in Fig. 2, but they are not useful for the problem of scale and it is useful to standardize the series before plotting them.

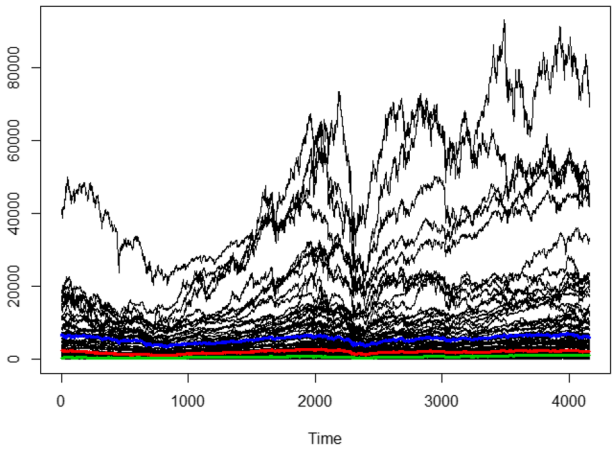

Fig. 2 World series of prices in 99 Stock markets in the period 2000 to 2015

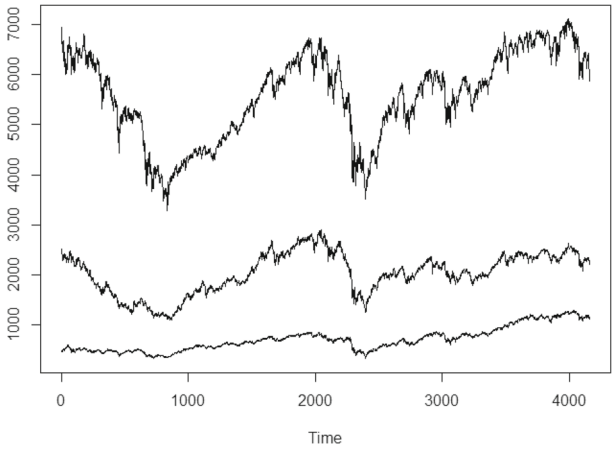

Fig. 3 Three quartiles of the World Stock Prices. The first is the Rusell 2000 index of USA, the second the MSCI index of the Pacific Zone, and the third the FTSE 100 of London

Besides scatterplots and three-dimensional scatterplots, heatmaps are graphical representations of the sample correlation matrix that associates different colors with low and large correlations. Hierarchical clustering using these correlations as distances between variables can be used to identify groups of variables highly correlated. Parallel coordinates plots are another useful tool to visualize a large number of variables. The idea is to map the data into a two-dimensional plot with two axes where the variables are mapped onto the horizontal axis, while the observed values of each variable are mapped onto the vertical axis. Note that parallel coordinates plots are very sensitive to the order of the variables. The Andrews plot results from representing the observations in terms of Fourier series. The resulting plot is a graphical representation of the curves obtained that are expected to have a certain common behavior if the variables in the data set are related. A very popular representation of texts and documents is the word

cloud. Essentially, the cloud shows the most frequent words in the texts that are shown with different sizes and colors in terms of their importance. In some sense, a word cloud can be seen as a kind of barplot for data taken from texts and documents. We refer to Cairo (2016) and Evergreen (2016) for recent books on data visualization with special emphasis on statistical graphics.

2.3 Multiple hypothesis testing and the false discovery rate

The traditional testing approach in statistics, and the one that is still usually taught in textbooks, is that we want to test some scientific hypothesis, then we collect the data and use it to test the hypothesis. However, the situation now in most applied studies is that we have a very large data set, then we imagine a possible set of hypothesis and then test all of them using some procedure. This change of paradigm creates two problems. The first one is the selection bias, i.e., the bias introduced in the analysis when the observed sample is not representative of the population under scrutiny. The second one is the multiple testing problem, i.e., the rejection of true null hypotheses when a large set of hypotheses are tested simultaneously. It is well known that if we want to test n hypotheses simultaneously and each hypothesis is tested separately using some significance level α, the probability of wrongly rejecting at least one null hypothesis is $1-(1-\alpha)^n$, that goes to one very fast with n. Consequently, when the number of hypotheses to test is large, we will wrongly reject at least one null hypothesis almost surely.

There are several ways to try to avoid the multiple testing problem. A common method was to use the Bonferroni bound that sets the significant level of the n tests at α/n. In this case, the probability of wrongly rejecting at least one null hypothesis is $1-(1-\alpha/n)^n$ that converges very fast to $1-e^{-\alpha}$, that is approximately α, for α small. Consequently, the Bonferroni bound is able to control the wrong rejections. However, it is very conservative, because one false null hypothesis will only be rejected if the associated p value is smaller than α/n, which will be very small if n is large.

Alternatively, Benjamini and Hochberg (1995) proposed a procedure, posteriorly extended, see Benjamini (2010), to control the false discovery rate (FDR). The FDR is defined as $E[V/R]$, where R is the number of rejected null hypotheses and V is the number of wrongly rejected null hypotheses. As the FDR is unobservable in practice, they proposed a procedure to ensure that $FDR \leq q$, if the test statistics are independent. Let $p_{(1)} \leq \cdots \leq p_{(n)}$ be the ordered p values corresponding to the n null hypotheses tested. Then, if k is the largest i for which $p_{(i)} \leq \frac{i}{n}q$, the Benjamini–Hochberg (BH) procedure rejects all the null hypotheses associated with the p values $p_{(1)}, \ldots, p_{(k)}$. Note that $q = \frac{n}{i}p_{(i)}$ is the q value of the test, i.e., the minimum FDR at which the test may be called significant. The BH procedure is more powerful than the Bonferroni method but the cost is to increase the number of Type I errors. It can be shown that the BH procedure also controls the FDR at level q under positive dependence assumptions. Otherwise, it would be necessary to find the k that is the largest i for which

$$p_{(i)} \leq \frac{i}{n \sum_{j=1}^{n} \frac{1}{j}} q.$$

The BH procedure, and other alternatives with the same objective, has become very popular, with large number of tests. This can happen in several problems such as outlier detection and outlier-free goodness-of-fit testing, variable selection and the determination of local covariance structures, among others, when the dimension of the data set is very large. For instance, Riani et al. (2009) proposed a method for outlier detection in multivariate data sets with a robust estimation of the Mahalanobis distances between the observations and a robust estimate of the center of the data. The test statistic is computed for subsets of observations, and these authors proposed a controlling method to avoid the false detection of outliers. Cerioli et al. (2013) proposed a robust method to test multivariate normality relying on Mahalanobis distances. For that the authors introduced a way to control the error rate when removing outliers of the observed sample based on the FDR. Barber and Candès (2015) studied how to identify a subset of relevant explanatory variables in regression models with a large number of regressors by controlling the FDR with knockoffs, which are new variables, obtained from the original ones, with similar correlations among then than the original ones. These knockoff variables are used as control to help in the selection of the relevant variables to predict the response. The method was extended to the case of arbitrary and unknown conditional models of any dimensions in Candès et al. (2016). Also, Sesia et al. (2018) extended this methodology to a rich family of problems where the distribution of the covariates can be described by a hidden Markov model (HMM). Cai (2017) reviews several papers dealing with multiple testing for high-dimensional covariance structures in large-scale settings including Cai and Liu (2016), who proposed an algorithm for simultaneous testing for correlations that has better performance than the BH procedure, and Liu (2013) and Xia et al. (2016), who proposed approaches for simultaneous testing for the existence of edges in Gaussian graphical models, and differential networks, respectively.

The use of controlling methods to avoid multiple testing problems is also very popular for the analysis in large-scale microarray data with number of variables going from thousands to millions. For instance, Tzeng et al. (2003) proposed a matching statistic for discovering the genes responsible for certain genetic disorders. The test statistic is computed for many regions across the genome and these authors used the BH procedure to control the false association of genes and disorders. Problems in this area are the identification of differentially expressed genes in mapping of complex traits, based on tests of association between phenotypes and genotype, among other experiments. All of them share some general characteristics such as thousands or even millions of null hypotheses, inference for high-dimensional multivariate distributions with complex and unknown dependence structures among variables, and broad range of parameters of interest, such as regression coefficients in nonlinear models, measures of association, and pairwise correlation coefficients, among others. In these Big Data circumstances, the use of multiple testing procedures controlling false discoveries seems essential.

2.4 Analyzing heterogeneous data

The possibility of fast and parallel computing is changing the way statistical models are built. Large data sets can be broken down into blocks to be processed and a central problem is to decide if they are homogeneous, so that the partial results obtained with each block can be combined in a single model, or are heterogeneous, and a mixture of models is required. Heterogeneity was usually considered in statistics as a two-model problem. Huber (1964), Box and Tiao (1968) and Tukey (1970) assumed that the data have been generated by a central model with some possible fraction of outliers coming from a different distribution, that is

$$x \sim (1-\alpha)\, F\,(x) + \alpha G\,(x)$$

where F is the central distribution, usually normal, and G is an arbitrary contaminating distribution. The large literature on diagnosis and robust statistics has been very useful to find outliers in large data sets, and it will continue to be important in the future. For instance, many communications and controlling devices automatically collect data using wireless sensor networks. However, sensor nodes sometime fail to record the data correctly (see Paradis and Han 2007, for a survey of this problem) due to depletion of batteries or environmental influence, and congestion in communication may lead to packet loss. These failures will produce outliers in the data generated by these sensors and some data cleaning method should be applied before building any model for the data, as it is well known that outliers can modify completely the conclusions obtained from statistical analysis. See Rousseeuw and van den Bossche (2018) for a recent analysis of finding outliers in data tables and Maronna et al. (2019) for an overview of robust statistics. This problem is also important in dynamic situations and Galeano et al. (2006) and Galeano and Peña (2019) have studied the detection of outliers in large sets of time series.

Although the idea of a central distribution is useful, it is too restrictive for many of the usual large data sets. A more appropriate representation is to assume that we have a mixture of models and, for that reason, cluster analysis is becoming a central tool in the analysis of Big Data. Many useful procedures are available for clustering. Partitioning algorithms, such as K-Means, see MacQueen (1967), PAM or K-Medoids, see Kaufman and Rousseeuw (1990), MCLUST, see Banfield and Raftery (1993), TCLUST, see Cuesta-Albertos et al. (1997), extreme kurtosis projections, see Peña and Prieto (2001a), and nearest neighbors medians clustering, see Peña et al. (2012), are useful for small data sets, but they have limitations when p and n are large. Some alternatives for large data sets have been proposed in the computer science literature; see Kriegel et al. (2009) for a review. Hierarchical methods can also be very useful, but they need to be adapted for large data sets. Two key problems in clustering high-dimensional data are: (1) the presence of irrelevant variables for clustering, because they negatively affect the efficiency of proximity measures; and (2) the dimensionality curse, which produces a lack of data separation in high-dimensional spaces. The first problem has been tackled by variable selection and the second by dimension reduction.

Variable selection can be carried out by adding in the estimation criterion some penalty function, as in the Lasso method. For instance, in model-based clustering, we

can maximize the likelihood of the mixture of normals adding some penalty function to introduce variable selection (see Pan and Shen 2007; Wang and Zhu 2008). Also, we can select variables as a model selection problem, as proposed by Raftery and Dean (2006). Other variable selection approaches is due to Fraiman et al. (2008), who proposed an interesting method to detect the noninformative variables in clustering. Witten and Tibshirani (2010) developed a cluster algorithm that can be applied to obtain sparse versions of K-means and hierarchical clustering. Some comparisons of these methods and other related references can be found in Bouveyron and Brunet-Saumard (2014), who present a review of model-based clustering for high-dimensional data, and in Galimberti et al. (2017).

Dimension reduction is carried out by identifying some subspace which includes the relevant information for clustering. See Johnstone and Titterington (2009), for interesting insights on this problem, and Bouveyron and Brunet-Saumard (2014), for a survey of the field. See also Cook (2018) for dimension reduction in other problems.

Clustering time series is becoming an important tool for modeling and forecasting high-dimensional time series. See Aghabozorgi et al. (2015) and Caiado et al. (2015) for recent surveys of the field. High-dimensional time series are usually analyzed by Dynamic Factor models (see Peña and Box 1987; Stock and Watson 2002; Forni et al. 2005). However, these factor models have often cluster structure, that is some factors are general and others are group specific and finding clusters in time series that have a similar dependency will be an important objective. Some recent works in this field are Ando and Bai (2017) and Alonso and Peña (2018).

The idea of heterogeneity has been extended to all branches of statistics, by assuming different models in different regions of the sample space. For instance, in regression problems, we may assume the model

$$y_i|\mathbf{x}_i \sim \sum_{g=1}^{G} \alpha_g N \quad \mathbf{x}_i'\boldsymbol{\beta}_g, \sigma_g^2 \quad , \tag{3}$$

where $\alpha_g \geq 0$ and $\sum_{g=1}^{G} \alpha_g = 1$. This model has been studied extensively both from the Bayesian and the likelihood points of view. When the number of groups is known, we have some reasonable initial estimate for the parameters in the different regimes the model can be estimated by MC^2 methods. However, when this information is not available, the estimation of this model is a difficult problem. See Frühwirth-Schnatter (2006) and Norets (2010). In time series, Tong and Lim (1980) introduced the threshold autoregressive models, which have been very useful for modeling nonlinear time series (see Tong 2012; Tsay and Chen 2018).

With Big Data heterogeneity, instead of being a particular aspect of the data, should be the standard assumption. Thus, there is a need to reconsider the classic set up followed in most basic statistical courses and emphasize mixture models and cluster analysis from the first week of teaching.

2.5 Automatic procedures for model selection and statistical analysis

The first change that large data sets have introduced in statistics is the need for automatic procedures for model selection. We can fit a regression model with the care of a craftsman, checking for the best combination of the explanatory variables, using the residuals to do diagnosis and identify nonlinearities and monitoring carefully the forecast performance. However, when we need to fit thousands of possible nonlinear regression models we cannot rely on these careful step by step strategies to build each of these models and we have to use automatic procedures. The statistical methods developed first by Pearson and Fisher in the first half of the 20th century and later by Box, Cox, and Tukey, among others, in the second half of the previous century, were thought for small data sets and emphasized the detailed analysis in each particular problem. A breakthrough in building models was the automatic criterion proposed by Akaike (1973) to select the order of an autoregressive process. His criterion, AIC, provides a general rule to select among complex models. It can be said that AIC was the first step toward artificial intelligence in statistics. A few years later, Schwarz (1978), from a Bayesian approach, proposed the now called BIC criterion for model selection.

Suppose that we have a data matrix $\mathbf{X}$ of n observations and p variables and that we have fitted different models $f_i\left(\mathbf{X}|\boldsymbol{\theta}_i\right)$ which depend on a vector of parameters $\boldsymbol{\theta}_i$ and let $c_i = \dim\left(\boldsymbol{\theta}_i\right)$. Calling $f_i\left(\boldsymbol{\theta}_i\right)$ to the maximum value of the likelihood function, these criteria select the model that minimize

$$M = -\log f_i\left(\boldsymbol{\theta}_i\right) + P\left(n, c_i\right),$$

where $P\left(n, c_i\right)$ is a penalty function that may depend on n and c_i. For AIC, $P\left(n, c_i\right) = c_i$, and for BIC, $P\left(n, c_i\right) = \left(c_i \log n\right)/2$, and it is well known that the AIC criterion is asymptotically efficient, i.e., selects the model with minimum out of sample expected error, whereas BIC is consistent, i.e., selects asymptotically the true model with probability one (see Yang 2005, for an analysis of these properties). It is well known that the results of model selection can be very different from the ones obtained with significant tests. As an example, suppose we compare two regression models: the first contains p variables and the second includes an additional variable, so that it has $p + 1$ variables. The BIC criterion will select the first model if:

$$BIC_p = n \log \sigma_p^2 + p \log n < BIC_{p+1} = n \log \sigma_{p+1}^2 + (p+1) \log n, \tag{4}$$

that is when $\sigma_p^2/\sigma_{p+1}^2 < n^{1/n}$, where σ_p and σ_{p+1} are the estimated residual variances with p and $p + 1$ variables, respectively. A significant F test at α significance level will check the coefficient of the additional variable in the second model. This test is equivalent to the standard t test for the significance of the coefficient of the new variable and the F statistic can be computed as

$$F = (n - p - 2)\left(\frac{\sigma_p^2}{\sigma_{p+1}^2} c_{n,p} - 1\right). \tag{5}$$

where $c_{n,p} = (n - p - 1)/(n - p - 2)$. Then, the simplest model will be chosen if this value is smaller than the selected critical value $F_{1,n-p-2,\alpha}$. Usually, the value of α is chosen a priori, a common level is $\alpha = 0.05$, and the simplest model will be accepted, or the additional variable will be rejected, if $F < F_{1,n-p-2,\alpha}$. Therefore, both procedures check the value of $\sigma_p^2/\sigma_{p+1}^2$, but the decision with the BIC criterion depends strongly on the sample size, whereas the one with the significance test will depend mostly on α. For instance, if the sample size is larger than 100, p/n is small and $\alpha = 0.05$, then $F_{1,n-p-2,\alpha} \approx 3.85$. Note that the value of the F statistic that makes $BIC_p = BIC_{p+1}$ is, assuming $c_{n,p} = 1$,

$$F^*(n) = (n - p - 2) \quad n^{1/n} - 1 \tag{6}$$

and we have that $F^*(100) \approx 4.5$, whereas F^* (100,000) ≈ 11.5, for p/n small. Thus, with the significant test we usually reject the additional variable if $F < 3.85$, whereas with the BIC criterion we will reject it for $F < 4.5$, if $n = 100$, and for $F < 11.5$, if $n = 100{,}000$.

The BIC criterion can be interpreted as a significance test where the α level decreases when the sample size increases. Thus, in practice, the results of model selection criteria for large sample size are very different from those of significant tests in comparing models with different number of parameters. As it has been discussed in Subsect. 2.3 statistical tests were not designed to be applied with very large data sets, or to compare models with different number of parameters. For this reason, model selection procedures are more useful for selecting models with Big Data. For instance, we have checked that the top 10 articles in the list of the 25 most cited statistical articles (Ryan and Woodall 2005) have increased their cites between 2005 and 2015 by a factor around 2, and the most cited article in statistics, Kaplan and Meier (1958), had gone from around 25,000 cites in 2005 to 52,000 in 2015. However, the two seminal articles that introduced automatic criteria for model selection have multiplied their cites by more than 10 times in this period. More precisely, from 2005 to 2015, Akaike (1974) has gone from 3400 to 38,000, and Schwarz (1978) from 2200 to 33,000. The existence of these criteria for model selection has stimulated statistical automatic modeling in many fields. For instance, Gómez and Maravall (1996) developed the programs TRAMO and SEATS for automatic modeling and forecasting of time series that have become very popular in economic and business applications. The very popular book on statistical learning by Hastie et al. (2009) illustrates the usefulness of automatic modeling in many different statistical problems.

The AIC and BIC criteria were derived as asymptotic approximations when the sample size goes to infinity. They are less useful when the number of variables, p, is very large, even greater than the sample size, n, and new criteria for model selection have been proposed when both p and n go to infinity. In fact, the emergence of Big Data has created new asymptotic theories when both p and n are large. For instance, Chen and Chen (2008) generalized the BIC penalty term for situations, as in gene research, in which we have much more variables than observations, $p > n$. This problem also appears in large panels of time series in which we have also the number of series, m, can be much large than the number of observations in each time series, T. Bai

and Ng (2002) have proposed three consistent criteria for these problems, where the penalty term depend on both p and n. Suppose we compare dynamic factor models with different number of factors, then the first modified BIC criterion proposed by Bai and Ng is

$$IC_1(p) = Tm \log \sigma_p^2 + p(m+T) \log\left(\frac{mT}{m+T}\right), \tag{7}$$

where p is the number of factors, and σ_p^2 is the average residual variance of the fitted factor model. Note that the number of observations is $n = mT$ and, the number of estimated parameters is pT, for the factors, and mp, for the loading matrix. Therefore, comparing (4) and (7), we see that they have the same form but the penalization is different, instead of $\log(mT)$, that will be the equivalent to n in (4), we have $\log \frac{mT}{m+T}$. For instance, for $T = m$, the criterion is

$$IC_1(p) = T^2 \log \sigma_p^2 + 2pT \log\left(\frac{T}{2}\right),$$

and the penalty for an additional factor is smaller than that with standard BIC criterion (4). This is reasonable because the minimum value of m and T fix the rank of the system: if $m > T$, the rank of the covariance matrices is T, while if $T < m$, the rank is m. See Peña et al. (2019a) for the use of these criteria to build automatic forecasting procedures with dynamic principal components for large sets of time series.

An alternative method to derive automatic model selection procedures is cross-validation (CV), introduced by Stone (1974), as a universal nonparametric rule for model selection. Suppose we have data $(y_i, \mathbf{x}_i)$, for $i = 1, \ldots, n$, and we want to compare several predictor models for y based on the covariates $\mathbf{x}$. We use the data to estimate different models, $j = 1, \ldots, M$, leading to predictions of the form $y_i(j) = g_j(\mathbf{x})$ with estimation error $\sum_{i=1}^{n} (y_i - y_i(j))^2$. To compare these models, we would like to obtain independent estimates of the forecasting error. To do so, we divide the data into two parts, an estimation or training sample and a validation or prediction one. Then we use the first part to estimate the model and the second to check the out-of-sample performance of each prediction rule. The model selected will be the one with best out-of-sample forecasting performance. In practice, there is no clear criterion to split the sample into the estimation and validation parts and Stone thought of a way of estimating the validation error with the maximum number of points. He defined leave-one-out cross-validation (LOOCV) as a procedure in which we estimate the model in a sample of size $n - 1$ and forecast the deleted observation. This out-of-sample forecast can be applied to the n observations in the sample to compute a cross-validation forecasting error with all the sample points. See also Geisser (1975) for a similar approach. Multifold cross-validation, leaving n_0 observations out, training the procedure in $n - n_0$ data, and forecasting n_0 has been found to work better than LOOCV in many settings (see Zhang 1993; Shao 1993). As this requires to compute all $\binom{n}{n_0}$ samples, and for large n, this number is huge, some approximations are made. See Arlot and Celisse (2010) for a survey of this field.

Model selection and cross-validation are very related. Stone (1977) was the first to prove the asymptotic equivalence of cross-validation and AIC and many articles have compared these two approaches, see Arlot and Celisse (2010), for many references. The main argument for CV is its generality: it has been applied in many problems, including regression, discriminant analysis, cluster analysis, and density estimation, among others. On the other hand, the computational cost is usually higher, and when there is a clear family of models to be tested, the model selection approach works usually better. Also, cross-validation was derived under the assumption of independent data, whereas model selection has no this limitation. Thus, it is not obvious how to apply cross-validation to time series, spatial data and other dependent data. A few works have tried to extend these ideas to correlated data. For instance, Peña and Sánchez (2005) proposed a multifold validation procedure for ARIMA time series, and Bergmeir and Benítez (2012) compared different procedures in real data sets. However, this problem requires further research.

2.6 Estimation procedures in high dimension with sparse models

Many statisticians were puzzled when James and Stein (1961) proved that for $p \geq 4$, the maximum likelihood estimator of the vector of population means, $\boldsymbol{\mu}$, the sample vector mean, $\overline{\mathbf{x}}$, is inadmissible: It has always larger mean-squared error than the shrinkage estimate

$$\boldsymbol{\mu} = \alpha \overline{\mathbf{x}} + (1 - \alpha)\, \mathbf{1}_p \overline{x}, \tag{8}$$

where $0 < \alpha < 1$, $\overline{\mathbf{x}}$ is the sample mean vector, and $\overline{x} = \frac{1}{p}\mathbf{1}'_p \overline{\mathbf{x}}$ where $\mathbf{1}'_p = (1, \ldots, 1)$. This implies that we can improve the ML estimate by giving some arbitrary weight to the vector of means computed assuming that all the components of the vector have the same mean. Note that this will also be the Bayesian estimate for this problem assuming a common prior $\mu_0 \mathbf{1}_p$ for the distribution of $\boldsymbol{\mu}$. The shrinkage coefficient, α, depends on the variability among the components of $\overline{\mathbf{x}}$. A similar results was discovered a few years later for the least squares estimate $\boldsymbol{\beta}_{LS}$ in the regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}$. It can be shown that when we have a large number of predictors, k, we can always improve the least-squares estimator by using the shrinkage estimate

$$\boldsymbol{\beta}^R = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{A})^{-1} \mathbf{X}'\mathbf{Y}$$

where $\lambda > 0$ and $\mathbf{A}$ is a positive definite matrix. Taking $\mathbf{A} = \mathbf{X}'\mathbf{X}$ we obtain $\boldsymbol{\beta}_R = (1 + \lambda)^{-1} \boldsymbol{\beta}_{LS}$, which is a James-Stein estimator that shrinkage the LS estimate toward zero. Taking $\mathbf{A} = \mathbf{I}$, the Ridge regression estimate introduced by Hoerl and Kennard (1970) is obtained. An interesting property of this estimate is that it can be obtained as a solution to the problem

$$\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|^2 \right\} \tag{9}$$

where $\|\cdot\|$ represents the Frobenius norm of a matrix or the Euclidean norm for a vector. If we have many predictors we expect that several of them will have a small effect

in the response and then their regression coefficients will be close to zero. Imposing a penalization on the norm of the vector can improve the accuracy of the estimation. With many predictors, a better penalized function is

$$\min_{\boldsymbol{\beta}} \ \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\ ^2 + \lambda\ \boldsymbol{\beta}\ _{L_1} \tag{10}$$

where now the penalty depends on the L_1 norm. The advantage of (10) with respect to (9) is that it will force small coefficients toward zero, improving the sparsity of the final estimate. This is the lasso estimate introduced by Tibshirani (1996). This idea of imposing a penalty function in the estimation of the parameter is usually called regularization and has many applications in statistics for sparse data, that is, when only a relatively small number of parameters are required to explain or forecast the data. In these cases, we can estimate these parameters effectively, using the lasso in an equation with all the possible parameters. Problem (10) is convex and the solution for a given λ can be easily found (see Hastie et al. 2015). The parameter λ is usually chosen by cross-validation. For that, the sample is split into h groups with $h > 1$. Then we take the first group as test or estimation group and the remaining $h - 1$ groups as validation or training sample. The model is estimated in the estimation group for a range of values of λ. Then this fitted model is used to predict the responses in the $h - 1$ validation groups and computing the mean-squared prediction errors for each value of λ. The same process is repeated for the 2nd,…,hth group, obtaining h different estimates of the prediction error and λ is chosen as the value with smallest average prediction error. The lasso approach and its generalizations, such as elastic net, group lasso, and fused lasso (Hastie et al. 2015), have been applied to many sparse problems such as sparse covariance matrix estimation, see Friedman et al. (2008), Bickel and Levina (2008), Cai and Liu (2011), and Cai and Zhuo (2012), among others, sparse principal component analysis, see Shen and Huang (2008) and Candès et al. (2011), and canonical correlation analysis, see Witten et al. (2009). Hastie et al. (2015) includes applications of regularization methods in logistic regression, generalized linear models, support vector machines, discriminant analysis and clustering.

Also, time series shrinkage estimates have been found useful in improving forecasts. García-Ferrer et al. (1987) showed that the univariate forecasting of macroeconomic variables can be improved by using pooled international data. This is a similar result to (8), and has the form

$$\mathbf{y}_t^P = \alpha \mathbf{y}_t + (1 - \alpha)\, \mathbf{1}\overline{y}_t$$

where $\mathbf{y}_t$ is a vector of forecast of time series computed by using linear univariate models and $\overline{y}_t = \frac{1}{p}\mathbf{1}'_p\mathbf{y}_t$ is the mean of the forecasts. Peña and Poncela (2004) showed that this class of pooling forecast can be generated by a dynamic factor model. Lasso estimation has been also applied to times series, see, for instance, Basu and Michailidis (2015).

Another approach in which the L_1 norm is used is compressive sensing, a signal processing tools in which we want to find linear combinations of many variables that keep all the relevant information. Donoho (2006a, b) proved that the minimal

L_1 norm is the sparsest solution in many of these problems of linear data reduction. See also Candès et al. (2006) and Candès and Tao (2006). This approach has opened the way to computing random projections in large dimensional spaces to find new variables, linear combinations of the original ones with good explanatory power. See, for instance, Guhaniyogi and Dunson (2015) for a Bayesian application to regression problems.

2.7 Analyzing networks and incorporating network information into statistical models

The extreme popularity in recent years of social networks, such as Facebook, Twitter, Linkedin, and Instagram, has placed the focus of many researchers and companies on the analysis of network data. Networks can be found in many diverse fields. For example, technological networks, which include transport networks, such as air routes, energy networks, such as electricity networks, and communication networks, between interactive communication devices. Biological networks represent biological systems, such as networks of neurons, or information networks, that describe relationships between information elements, such as citing networks of academic articles. The information contained in a network is very rich in itself and has led to what is called network science, see Kolaczyk (2009) and Barabási (2016). This information can also be of tremendous utility for the enrichment of usual statistical models. Before discussing this new field, we briefly describe some network features and problems that can be relevant for this goal.

The mathematical basis behind network analysis is graph theory that dates back to 1735 when Leonard Euler solved the famous problem of the seven bridges of Königsberg. Graphs offer a common framework to analyze networks that may have many different characteristics in terms of form and size, among many other features. Essentially, a graph consists of a list of elements usually called vertices or nodes, and the connections between them, usually called edges or links. The edges of a network can be directed or undirected, depending on whether they have a direction, and/or a weight, that somehow measures the strength of the edge. Two relevant problems in network analysis are vertex centrality and community detection. On the one hand, measuring vertex centrality is important to identify the key vertices in the network. For instance, in social networks, the most important vertices are used to identify the network influencers. Probably, the easiest way to measure vertex centrality is through the vertex degree. Other alternatives are the closeness centrality, the betweenness centrality, and the eigenvector centrality. On the other hand, community detection is important to identify set of vertices well connected among them, and relatively well separated from vertices in other sets. The two main community detection algorithms are hierarchical clustering and methods based on network modularity. See Kolaczyk (2009) for a complete overview on network features and problems.

Vertices and edges have certain characteristics called attributes. For instance, in a social network in which the vertices are people and the edges friend relationships, attributes of the vertices can be age, gender, marital status, and personal likes, among others, while attributes of the edges can be the number of private messages between

them and the duration of the relationship, among others. Now, if we want to classify new members of the network in terms of a certain variable, the inclusion of network information, such as vertex centrality and/or communities, will improve the classification power of standard methods. Section 4.1 below presents an example in which the inclusion of network information improves the power of several statistical methods used to solve three different problems regarding bank customers.

There is a recent growing interest in the interaction between statistical methods and network analysis. For instance, Gaussian graphical models are frequently used for modeling the conditional dependence structure of large dimensional systems. This is because the structure of an undirected Gaussian graph is characterized by the precision matrix of the distribution of the random variables, see Lauritzen (1996). Accurate estimation of high-dimensional covariance, correlation and precision matrices under Gaussian graphical models and differential networks have been carried out by several authors including Meinshausen and Bühlmann (2006), Cai et al. (2011), Zhao et al. (2014), Ren et al. (2015), and Cai (2017), among many others. It is important to note that all these papers consider regularization methods, such as the Lasso mentioned in Sect. 2.6, to determine the existence of relationships between variables, or equivalently, the existence of edges between nodes in the associated graph. In the time series setting, Zhu et al. (2017) proposed network vector autoregressions to analyze the dynamic behavior of networks evolving over time. These network vector autoregressions resemble the vector autoregression models, where a vector of time series is explained in terms of its past, some covariates and independent noise. The idea is to explain some attribute in terms of past information of the nodes and their neighbors, as well as certain covariates and independent noise. Additionally, Wei and Tian (2018) have considered a similar approach in regression problems by proposing a network regression model. The idea is to understand or predict the effects of network systems on certain response variables. Estimation of network vector autoregressions and network regression models can be carried out with the combination of least squares and regularization methods. As the previous papers suggest, there is a wide field of analysis of the interaction between classical statistical models and networks that can be very useful for improving the analysis in both fields of interest.

3 The emergence of data science

During most of the last century statistics was the science concerned with data analysis. Once the objective of the study was defined, statistics has a role in all the steps of data analyses: (1) collecting the data, by sample surveys or designing experiments; (2) describing the data, by plots and summary statistics, and selecting a possible model or a set of models; (3) estimating the model parameters, by maximum likelihood or Bayesian estimation, and making validation of the model or model selection; and (4) interpreting the results. Only in the first and last part of this process, defining the problem and interpreting the result, the main role correspond to people from the subject matter field of the application. The emphasis of this methodology was on model building and understanding the relation among the variables involved. The growth of data availability in the last part of the last century stimulated the need of solving

prediction problems in many areas that cannot be solved by the standard statistical methods.

Breiman (2001), who opened new ways for classification with CART and random forests, explained the two cultures of data analysis that were emerging at the end of the 20th century: Modeling, the core of statistical courses, and Forecasting, that was required in many fields with new types of data. In fact, the growth of data availability has been reducing the role of statistics in the data analysis process. First, as explained in Sect. 2.1, new types of information in engineering and computer science have been considered, requiring new tools for classification and prediction with a different philosophy to standard statistical methods. These new approaches, as neural networks, are providing solutions in the analysis of images or sounds where classical statistics have had a limited role. Second, when data are generated continuously with sensors or people activity recorded in an automatic way, the problems of data storing, handling and processing become very important, and scientists from computer science are not only taking an important role in making the data available for analysis, but also in developing new tools for analysis. For instance, the field of recommendation analysis, that uses previous people choices to forecast future choices, have been mostly developed in computer science. Third, new optimization requirement from the new problems, from support vector machines to Lasso, as well as the growing importance of network data has led to a closer collaboration of statistics and operations research, a field that splits from statistics in the second half of the 20th century and that has developed procedures very relevant for the needs of Big Data. For instance, linear programming to solve the L_1 optimization problems that often appear in finding sparse solutions in statistics. These changes have expanded the field of data analysis to create what is called data science, as the integration of ideas from statistics, operations research, applied mathematics, computer science and signal processing engineering. Donoho (2017) and Carmichael and Marron (2018) present very interesting discussions of the evolution of this field.

The idea behind artificial intelligence is that the process of human thought can be mechanized. This is a broad concept that leads to many different research areas. In particular, machine learning is the part of the artificial intelligence that allows machines to learn from data by means of automatic procedures. Probably, the first paper mentioning the term machine learning was Samuel (1959) who wrote a program to play the game of checkers. The program improved the results by analyzing the moves that leads to winning strategies. Just one year before, Rosenblatt proposed the perceptron, i.e., the first neural network for computers that simulates the thought processes of the human brain. Since them, many machine learning researchers have proposed data-driven procedures to learn from data in an automatic way. The main focus of these analyses are on supervised and unsupervised learning problems, known in statistics as discrimination and clustering problems, respectively, and on dimension reduction techniques.

Some of the most popular tools for supervised classification in the machine learning area along the years includes the perceptron (see Rosenblatt 1958), the k-nearest neighbors (k-NN) algorithm (see Cover and Hart 1967), the classification trees (see Breiman et al. 1984), the feedforward neural networks (see Hornik 1991), the support vector machines (see Cortes and Vapnik 1995), the naïve Bayes classifiers (see

Domingos and Pazzani 1997), the random forests (see Breiman 2001), and the deep learning methods (see LeCun et al. 2015), among others. See also Genton (2001) for an overview of kernels, that are frequently used in machine learning methods for supervised classification, from a statistical perspective and Lam et al. (2018) for an efficient implementation of support vector machines in high dimension low sample size settings. On the other hand, some popular unsupervised classification methods in machine learning are subspace clustering, pattern-based clustering, and correlation clustering methods, see Kriegel et al. (2009), for a review. Finally, Kernel principal component analysis (KPCA), see Schölkopf et al. (1997), independent component analysis (ICA), see Hyvärinen and Oja (2000), and partial least squares, see Cook (2018), are popular approaches to the dimension reduction problem.

The success of machine learning methods is the integration of some useful methods developed for large data analysis with the ones created in statistics, operations research and applied mathematics. For instance, the support vector machines and the regularization methods heavily rely on solving more or less complex optimization problems. Also, many methods of network analysis, such as community detection, involve the intersection of these areas. Computational efficient implementation of all these methods in large-scale settings is an important issue. As a consequence, a procedure that may not be particularly attractive from a theoretical point of view, may have its space if it allows to solve a problem that otherwise would not have an easy solution. For instance, the naïve Bayes classifiers, a family of procedures very little appreciated in the statistical community, are very popular in large-scale supervised classification, where other more theoretical attractive methods are not applicable or may have worse performance than expected in such large-scale settings.

The range of applications of machine learning is somehow broader than that of statistics, mostly restricted to well structured data sets in the form of tables. For instance, texts and documents classification, image, video and speech recognition, natural language understanding and language translation, among other issues, are the natural domain of applications in the artificial intelligence and machine learning areas. Many of the advances on these areas comes from substantive real problems such as automated brain tumor detection from images. However, statistics can be very useful when the objective is to understand the relationship between the variables involved and to make models able to describe the problem and generate forecast in these situations. This explains why statistics is the support of many sciences such as demography, economics, environmental science, medicine, and psychology, among many others. Statistics offers a rigorous process for analyzing data that includes important steps such as data sampling, exploratory and descriptive analysis, inference, prediction, measurement of uncertainty, and interpretation. Many of these steps are usually ignored by the machine learning community, mainly focused in obtaining automatic predictions from data.

We believe that we are going to see a convergence of these different approaches of data analysis under the data science umbrella and that this process will stimulate scientific advances in all areas of knowledge. Statistical analysis will continue to be the core of scientific modeling with well structured data, but machine learning and artificial intelligence will create new forecasting procedures in problems where the relationship between the output and the inputs available for its prediction is not

well understood. On the other hand, statistical ideas will be used to decompose and understand the forecasting rules created in other areas, to identify the importance of the more relevant variables and to split the signal from the noise. All these advances will be the subject of data science.

4 Two examples of big data analysis

In this section, we will present two examples of analyzing Big Data using several of the procedures explained in the previous sections. They use data sets of several millions of records and both were carried out on demand of a private company. The first example analyzes the network of more than five millions of customers of Bank of Santander (BS) in Spain. In this project, in addition to building the network and using network variables for improving the performance of forecasting models, we have required new data visualization tools for networks, heterogeneity and cluster analysis, automatic model building, high-dimension estimation, multiple testing and outlier analysis. The second application is concerned with forecasting customer loyalty, using data of more than eight millions of customers of a chain of supermarkets in Spain, DIA. In this study we have developed new ways of visualizing large sets of time series, built forecasting procedures combining cross-section and dynamic information, estimated many models using automatic procedures and dealt with several sources of heterogeneity, including cluster and outlier analysis. Both studies have been carried out with other members of the Institute UC3M-BS of Financial Big Data (IFiBiD), which are listed in the acknowledgements, and in close collaboration with teams in BS and DIA.

4.1 Customers network analysis

4.1.1 The problem and the data

The main objective of the project was to investigate whether the information contained in the BS data base of its customers can be analyzed as a network useful to guide BS future actions and policies. Specifically, the project focused on solving three relevant issues: (i) to build the BS customer network and use it to analyze the intensity of economic relations between customers, the groups formed by similar clients and the centrality and importance of each customer; (ii) to develop a decision support system for helping BS managers to decide the sequence of customers to contact to reach a designed target; and (iii) to develop statistical models to explain the entry and exit in default of different types of BS customers (companies, freelancers and individuals). For space limitations, we will focus mostly in the third issue and give a very brief summary of the results on the first two issues.

To carry out the project, BS allowed us to access to totally anonymized information on several millions of customers and a total of 81 millions of transactional relationship between them in three periods of time: December 2014, June 2015 and December 2015, respectively. The whole set of information was split in three categories: (1) customer profiles, including age, type of consumer, relationship with BS, products and

services contracted with BS, such as payrolls, credit cards, receipts,…, and resources, such as amount of money in accounts, savings insurance, deposits or funds; (2) relationship between customers, including different types of relationship categories, the direction of the relationship, and indicators of the relationship intensity; and (3) customer default status, including the amount of the default, if any. A careful treatment of this information led to the identification of many outliers that correspond mostly to changes in the way the data was recorded, typing errors or other mistakes. As a result of this cleaning, three structured databases of debugged and reliable BS customers were constructed corresponding to each of the time periods considered. These databases, corresponding to almost 5 millions of customers and 6, 3 millions of relationships, were used to build a customer network to analyze the three issues considered.

4.1.2 Network analysis

The first step of the project was to analyze the structure of the BS customer network. For that, we constructed a graph formed by vertices and edges, where each vertex represents a BS customer (companies, freelancers and individuals), and each edge represents at least one relationship or flow between two customers. As two costumers can be related in many ways, all possible edges are summarized in a single one, that has as attributes all types of existing relationships. In addition, each edge is valued by a weight function taking values in the interval $[0, 1]$, to represent the strength of closeness between the customers that it unites. That is, a weight value close to 1 represents the largest closeness between the two customers. We focused on determining the topology of the network to understand the mechanisms underlying the aggregation of new nodes in the network. For this, several characteristics were used, including measures of the centrality of the customers in order to quantify the relationships of power, protagonism, trust, etc…, and the detection of specific communities that may have interesting characteristics. This is important to determine which customers and communities are the most relevant within the network. On the one hand, we used measures such us the vertex degree, the eigenvector centrality and the concentration degree (see Kolaczyk 2009) that allowed us to find the most influential customers through their connections in the network. The main conclusion of this analysis is that new customers often relate to highly connected customers that represent centers of influence in the network. Therefore, the maintenance and strengthening of these influential customers are of primary importance for the preservation of the structure of the network and its expansion. On the other hand, we used community detection algorithms, such as the one proposed by Blondel et al. (2008), specially suited for very large networks, to find groups of customers with a strong mutual relationship. A total of approximately $120,000$ communities were detected. The vast majority of communities have a very small size. To have an idea, the three largest communities have approximately $250,000$, $156,000$ and $94,000$ customers. An in-depth study of the most important communities allowed us to identify common characteristics among the customers that compose them that helped BS to design strategies and products specifically addressed to these groups. Also, the network offered new insight about the importance of the customers for BS. For example, commercial banks usually classify their customers for the amount of assets or deposits in the bank. However, a better

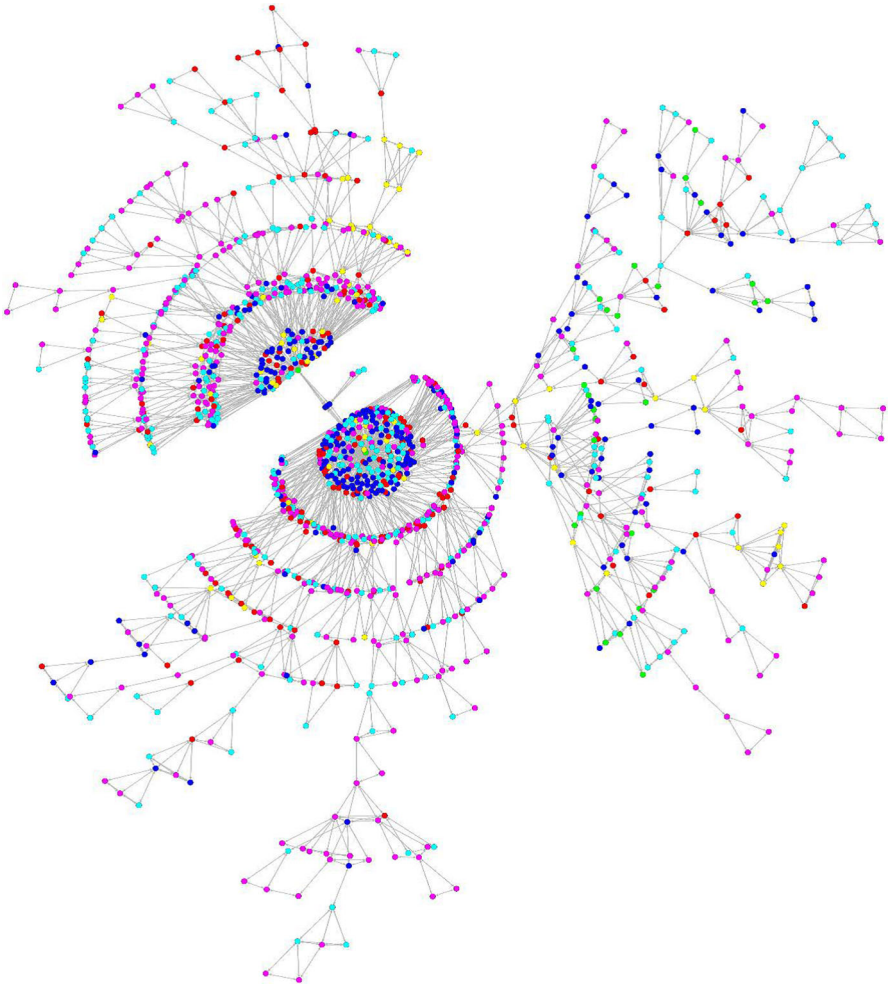

Fig. 4 Two large communities connected by a key customer in the BS network

classification can be obtained by thinking in the loss for the bank if a customer moves to another bank. To estimate this effect we need to consider, in addition to the assets, the relation that this client has in the network and the effect that leaving it can have in other clients, that depend on his/her connections in the network. Figure 4 illustrates this situation, where two large communities are connected by a key customer.

The second step of the project was to develop a methodology to determine the best sequence of customers that a BS manager should contact to reach a target starting from any client in the manager's portfolio. With this methodology, the BS would acquire new customers or sell additional products to existing ones. For that, a decision support system (DSS) was developed that provides managers of BS with possible paths to follow to attract new customers sorted by probability of success. These probabilities were determined with the information given by the customers' information as well as

with the network information. A complete description of the DSS, including the way in which the probabilities of success of different paths are obtained, can be found in Quijano-Sánchez and Liberatore (2017).

4.1.3 Improving prediction with network variables

The third step of the project was to analyze the default status of different types of BS customers. For that, we developed a set of statistical models for investigating the entry and exit in default of different types of BS customers. Models were constructed to explain the customers' default status in two temporary moments of the year 2015 (June and December, respectively) using a wide set of explanatory variables that include: (i) services and products contracted with BS; (ii) resources in BS; (iii) situation and connections within the BS customer network; and (iv) changes in all these variables with respect to the previous period. Importantly, note that we are using information from the customers themselves but also new information on the situation of the customers within the BS customer network. For instance, whether the customer has direct or indirect connections with default customers. Additionally, the models were built for different groups of customers that result from segmenting them in terms of three types of customers, i.e., companies, freelancers and individuals, and four types of linkages with BS, i.e., very strong, strong, weak, and very weak. Consequently, we needed to built a total of 24 models, resulting from the combination of 2 periods of time, 3 types of customers, and 4 types of linkages with BS. The number of customers in each group ranges from about 50, 000, for freelancers with very weak linkage with BS in June 2015, to around 3 millions, for individuals with weak linkage with BS in June 2015.

The generic model chosen to explain the customer default is logistic regression for two main reasons. First, as we will see, logistic regression allows to determine the importance of each of the variables used to explain the default status. This is an important advantage over alternative models because we can identify the variables that best explain the default of BS clients and measure their effects in terms of default probability. Second, this model has proven their effectiveness for prediction in many different contexts. The default status will be the variable to explain, denoted by y, taking values 0 and 1, to represent the no default and the default status, respectively. The proportions of default customers in the 24 classes considered, ranges from 0.02, for individuals with strong linkage with BS in June 2015, to 0.4507, for freelancers with very weak linkage with BS in December 2015. The explanatory variables are classified in three blocks. The first block includes variables that measure the use of products and services contracted by the customer, as well as some customer descriptive variables such as whether the customer is active or retired, among other things. We consider 18 categorical variables describing the customer and the use or not of a product or service offered by BS, as well as another set of 32 dynamic variables describing the changes experienced with respect to the previous observed period, i.e., December 2014 in the case of June 2015, and June 2015 in the case of December 2015, respectively. The second block of explanatory variables includes the available resources of the customer in BS. We consider 9 quantitative variables describing the resources of the customer, such as payrolls, deposits, etc…, as well as another set of 9 variables describing the changes experienced with respect to the previous observed

period. Due to their high skewness, all the variables in this second block have been transformed using a logarithmic transformation. The third block includes 15 network variables: 6 variables measuring the proportion of direct neighbors or second-level neighbors (that is, neighbors of neighbors) that are default customers in BS, divided in companies, freelancers and individuals, and another set of 9 variables describing the changes in these proportions experienced with respect to the previous observed period. In summary, we considered 80 explanatory variables for each of the 24 models constructed. We do not give a full description of all the variables here for easiness in exposition and prefer to focus in the most important ones to explain the default status once the model parameters have been estimated.

Call $\Pr(y = 1|\mathbf{x}_0)$ the probability that a given customer with p explanatory variables $\mathbf{x}_0 = (x_{01}, \ldots, x_{0p})'$ is in default. Assuming a logistic regression model, the odds ratio is given by:

$$O(\mathbf{x}_0) = \frac{\Pr(y = 1|\mathbf{x}_0)}{\Pr(y = 0|\mathbf{x}_0)} = \exp(\beta_0) \prod_{j=1}^{p} \exp(\beta_j x_{0j}),$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ is the vector of parameters of the model and $p = 80$ is the number of explanatory variables. In order to understand the coefficients β_j of the model suppose that we increase the value of a continuous variable x_{0j} from one unit, that is we go from x_{0j} to $x_{0j} + 1$, keeping constant the values of the rest of the variables. We consider only in the notation x_{0j}, as the rest of the explanatory variables are fixed. This analysis applies as well to the coefficient of a dummy variable that moves from the value zero to one. The change in the odds ratio will be

$$O(x_{0j} + 1) = O(x_{0j}) \exp(\beta_j)$$

where $O(x_{0j} + 1)$ denotes the odds ratio when the variable x_{0j} increases in one unit. Therefore, if $\exp(\beta_j) > 1$, then $\Pr(y = 1|x_{0j} + 1) > \Pr(y = 1|x_{0j})$, so that we conclude that the sign of the coefficient indicates if increasing the value of this variable in one unit has a positive or negative effect on the probability that the response is equal to one. The increase in probability depends on $\Pr(y = 1|x_{0j}) = \Pr(y = 1|\mathbf{x}_0)$. Assuming that the starting point is $\Pr(y = 1|x_{0j}) = 0.5$, we have

$$\Pr(y = 1|x_{0j} + 1) = \frac{\exp(\beta_j)}{1 + \exp(\beta_j)} = PC(x_j), \tag{11}$$

and we will call $PC(x_j)$ the probability change when the variable x_j increase in one unit with respect to a situation in which this probability was 0.5. In summary: (i) $PC(x_j) > 0.5$ means that if x_j increases, then $\Pr(y = 1|\mathbf{x}_0)$ also increases; (ii) $PC(x_j)$ 0.5 means that x_j does not have influence on $\Pr(y = 1|\mathbf{x}_0)$; and (iii) $PC(x_j) < 0.5$ means that if x_j increases, then $\Pr(y = 1|\mathbf{x}_0)$ decreases.

To compute estimates of the default probabilities and the probability changes in (11), we need to estimate the model parameters β and replace their values in the formulas.

Table 1 Percentages of wrong classifications for June 2015 and December 2015

Group	June, 2015		December, 2015	
	No default	Default	No default	Default
Freelancers and very strong	0.04%	1.64%	0.16%	3.29%
Freelancers and strong	0.02%	1.77%	0.10%	4.22%
Freelancers and weak	0.07%	1.34%	0.17%	2.46%
Freelancers and very weak	0.19%	0.70%	0.58%	1.07%
Companies and very strong	0.04%	1.22%	0.13%	2.35%
Companies and strong	0.02%	1.21%	0.10%	2.93%
Companies and weak	0.06%	0.74%	0.15%	1.28%
Companies and very weak	0.19%	0.63%	0.39%	0.81%
Individuals and very strong	0.02%	2.58%	0.09%	4.62%
Individuals and strong	0.01%	2.31%	0.05%	5.65%
Individuals and weak	0.02%	1.90%	0.07%	3.64%
Individuals and very weak	0.07%	0.86%	0.23%	0.97%

However, an initial analysis showed that a large number of the 80 explanatory variables considered for each of the 24 logistic regressions, had little predictive power. Thus, we considered two alternative ways to tackle the high-dimensionality problem. First, we used step-AIC backward deletion (see Hastie and Pregibon 1992), that discard variables with low prediction power using the Akaike information criterion (AIC). Second, we used Lasso logistic regression that, as explained in Sect. 2.6 maximizes a penalized log-likelihood function to shrink non-important parameters toward 0. The two methods led to very similar results in all the groups considered and here we summarize the results. First, Table 1 shows the proportion of wrong classifications with the 12 models obtained for each of the two periods, June, 2015, and December, 2015. Usually the errors are small so that the models work well. The largest errors appear with customers with strong linkage with BS, where the default is usually due to a very minor debts, such as the non-payment of a receipt due to neglect or forgetfulness. Then, customers with very good economic conditions can appear promptly as a default customer, which makes the classification of these persons very hard. The largest error is 5.65%, that corresponds to the default in June 2015, and corresponds to individuals with strong relation with the bank.

Second, Table 2 shows the two most important variables according to the statistic (11) for each of the 24 models considered. As it can be seen, the most important variable is being in default in the previous period (denoted by "Previous default" in the table). In all the cases considered, the value of the importance measure in (11) is equal to 1. Therefore, as one may expect, the fact of being in default 6 months ago appears to explain very well whether the currently defaults' customer status. Additionally, in most of the models (23 out of 24), the second most important variable is one of the new network variables considered related: having neighbors in the network that are default customers. In most of the cases (16) , the network variable is the proportion of neighbors or neighbors of neighbors (denoted by "Related with default" in the table)

Table 2 Two most important variables for all the models considered in terms of the customer and link types. Values of the probability of change, PC, are given in parentheses

Period	Link type	Customer type		
		Freelancers	Companies	Individuals
June, 2015	Very strong	Previous default (1)	Previous default (1)	Previous default (1)
		Related with default (.9925)	Related with default (.9814)	Increase of related default (.9798)
	Strong	Previous default (1)	Previous default (1)	Previous default (1)
		Related with default (.9537)	Increase of related default (.9146)	Related with default (.9475)
	Weak	Previous default (1)	Previous default (1)	Previous default (1)
		Related with default (.9791)	Related with default (.9819)	Increase of related default (.8849)
	Very weak	Previous default (1)	Previous default (1)	Previous default (1)
		Related with default (.9846)	Related with default (.9759)	Stop paying receipts with BS (.1341)
December, 2015	Very strong	Previous default (1)	Previous default (1)	Previous default (1)
		Related with default (.9732)	Related with default (.9723)	Related with default (.9799)
	Strong	Previous default (1)	Previous default (1)	Previous default (1)
		Related with default (.9589)	Related with default (.9614)	Related with default (.9492)
	Weak	Previous default (1)	Previous default (1)	Previous default (1)
		Increase of related default (.8486)	Increase of related default (.9419)	Related with default (.9627)
	Very weak	Previous default (1)	Previous default (1)	Previous default (1)
		Increase of related default (.7786)	Increase of related default (.9327)	Related with default (.9276)

that are default customers, while in the remaining 7 cases, the second most important variable is a variable that measures the increment in the proportion of neighbors or neighbors of neighbors (denoted by "Increase of related default" in the table) with respect to the previous period. Consequently, the proportion of neighbors in default is a fundamental factor to explain the customers' default status.

We conclude that the introduction of network variables in a statistical model can increase its power to provide a good representation of the data.

4.2 Monitoring customers loyalty

4.2.1 The problem and the data

A large food supermarket company (DIA) was interested in identifying clients that have a moderate or large probability of stop buying in their shops. Having this information, the company can use marketing strategies to retain these clients. Also, understanding their reason to leave will be helpful to develop strategies to increase the satisfaction and loyalty of their customers. Thus, the objective of the study was to provide to the company evidence of changes in the purchase behavior of the clients so that corrective actions could be taken. Our approach was to identify when a customer has a change in his/her pattern of purchases and build a model to estimate how this change modifies its probability of attrition or loyalty to the company. Identifying changes in pattern behavior is similar to the problem of statistical quality control, where we want to identify changes in a system in order to introduce the due adjustments to keep the system in a stable state but do not want to apply unnecessary adjustments when there is no evidence of change. Therefore, we proposed to combine dynamic variables, obtained from the analysis of the time series data of purchases, with the cross-section data of the characteristics of the clients, to build a predictive model to estimate the probability of a next purchase.

The available data for each customer are the amount spent each month in one of the supermarkets of the company in Spain in the period January 2014 to March 2016 ($M = 27$ months) by clients that use a fidelity card to obtain discounts for their food purchases. Thus, we also know some personal characteristics of these clients, such as sex, age, number of persons in the household, discounted received, and type of payment (credit card or cash). We say that a client is active in a given month if the amount spent in this month is greater than zero. The initial data base provided for the company includes 15, 9 million customers and after cleaning this initial data base by deleting obvious outliers, and clients with no activity in the period studied, we end up with about $N = 8, 3$ millions of customers that have at least a purchase in this period. This will be the number of time series to be analyzed.

4.2.2 Splitting heterogeneous data in groups

We assume that the probability that a client is active in a given month depends on his/her previous history of buying in the supermarket, summarized in a vector of variables, H_i, and of his/her personal characteristics, given by a vector of variables, C_i. The

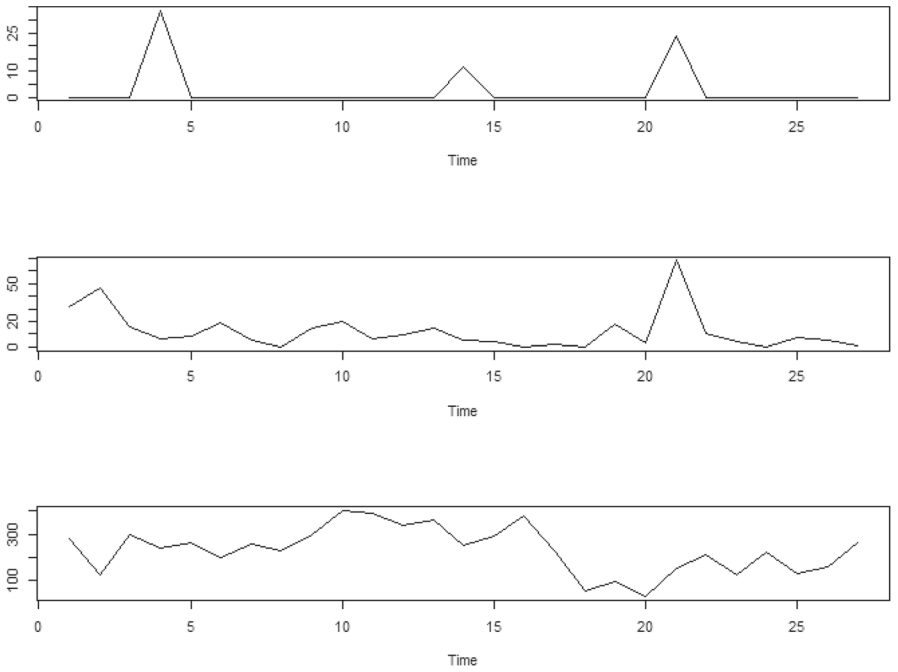

Fig. 5 Three time series of purchases of occasional (1st panel), frequent (2nd panel) and loyal (3rd panel) clients

variables, H_i, will be obtained from the time series of purchases by summarizing the dynamic features that can affect the probability of future buying. Figure 5 shows three time series of purchases that are representative of three typical patterns of customer behavior. The one in the first panel corresponds to a client that is only active in a few months, and the purchases amount is in general low. It can be seen in the time series plot that there are only three active months in the period and that the amount expended goes from zero to 25 euros/month. We will call *occasional* clients (O) persons that broadly buy less than half of the months in the studied period (a more precise definition will be done later). The second group of clients corresponds to those that are active in most of the months, and have moderate purchase expenses. The client in the second panel of Fig. 5 only misses four months, that is buys 85% of the time, and the expenses go from zero to 50 euros/month. These will be called *frequent* clients (F). The series in the third panel of Fig. 5 corresponds to a client that is always active, and the purchase amount goes from 30 euros/month to 402. The clients that are active all the months observed are called *loyal* clients (A).

As the frequency of buying seems to be a key variable in the analysis, we assume that the ith client, $(i = 1, \ldots, N)$, in each month, m, $(m = 1, \ldots, M)$ has a probability p_{im} of being active this month. We call $p_i = \sum_{i=1}^{M} p_{im}/M$ to the average probability of being active in the observed period. This probability is estimated as the proportion of active months, p_i, in the time series of purchases. Figure 6 shows the distribution of these estimated probabilities.

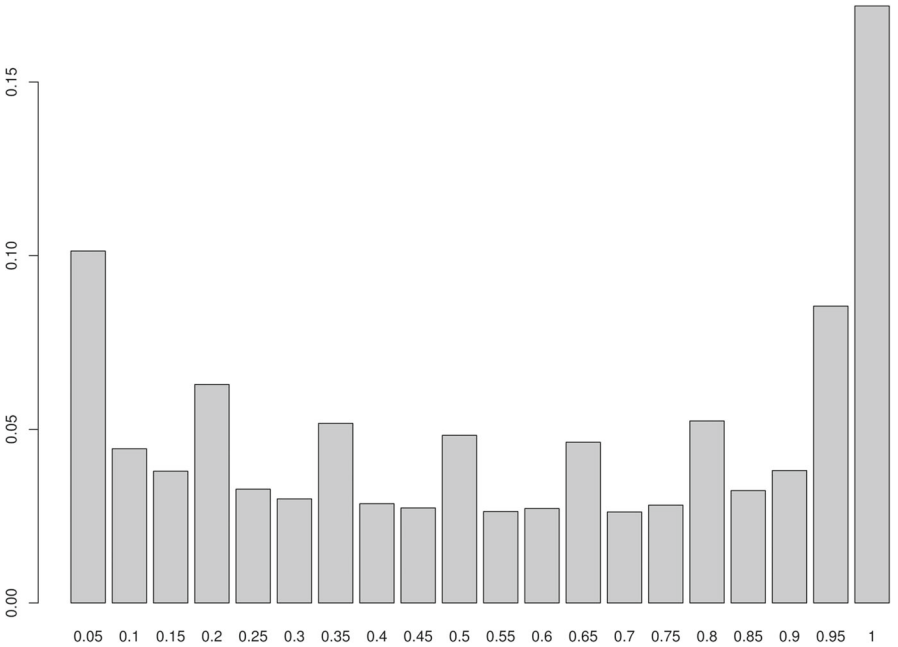

Fig. 6 Histogram of the proportion of months that the customers have been active

This histogram is an interesting example of false features in the data that can appear when using the default option in many computer programs. Note the relative increase in frequency of purchases in intervals with centers separated by 0.15. As the possible values of p_i are $k/27 = 0.037k$, for $k = 1, \ldots, 27$, and the histogram has 20 classes with width 0.05, larger than the increase in the values of p_i, 0.037, most classes in the histogram will include one of the possible values of p_i, but seven of then must include the sum of frequencies of two values. These happen in intervals 1, 4, 7, 10, 13, 16 and 19. Apart from this spurious effect, the general form of the histogram suggests a mixture of three clusters or populations. First, clients that are active (A) every month. They are concentrated in the interval for $p = 1$ and represent the 17.2% of the sample. Second, clients that are frequently active (F), and they can be defined as being active at least 60% of the times, that is in agreement with making purchases with more probability than the median of the data, ($p_{\text{Med}} = 0.59$). They represent the 31.4% of the sample. Third, occasional clients (O), that are active less than the 60% of the period considered and represent the 51.4% of the sample.

The data also show that the probability of being active depends strongly on a run of inactive months. Figure 7 illustrates this dependency and suggests the importance of the length of a period without buying in determining the probability of a purchase next month. As expected, this probability of being active after a period of inactivity is different for the frequent clients than for the occasional ones, as shown in Fig. 8. Also, the distribution of the purchase amount spend in food every month is different for the three types of clients. For the A group, the average is 104.20 euros/month,

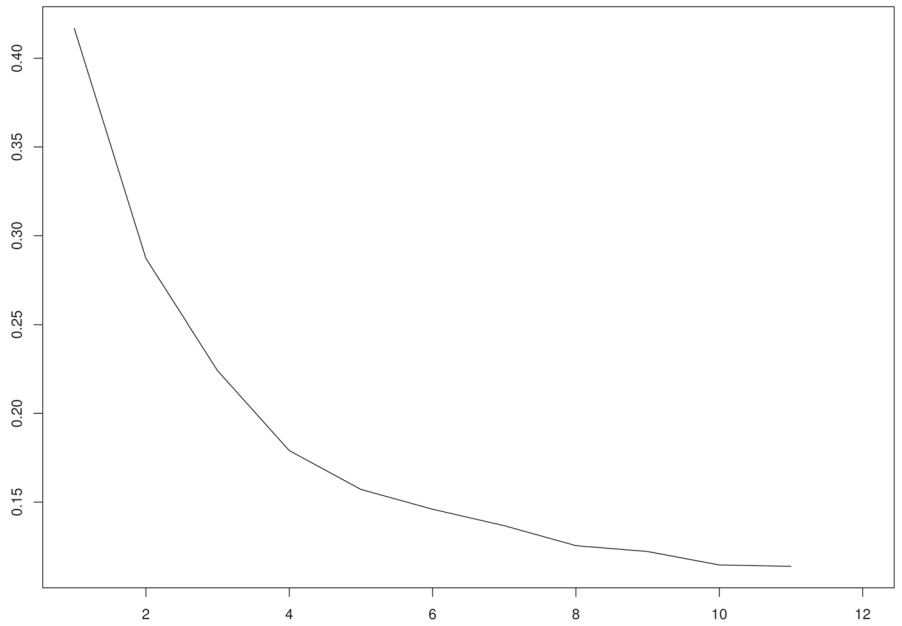

Fig. 7 Probability of being active next month (y-axis) after x months of inactivity

for the F group is 54.84 euros/month, and for the O group is 31.56 euros/month. The distribution is log normal, as shown in Fig. 9.

4.2.3 Summarizing the time series in dynamic variables

Two indicators in the purchases of each client are considered to forecast future activity: (1) a significant increase or decrease in the amount spent in the supermarket; and (2) the number of months without activity. We will say that a client has an inactive run of r months when he/she has not made any purchase in r consecutive months. We first describe how to identify a level shift in the purchases of a client and then how to summarize this information in a set of variables. Second, we analyze the inactive runs and propose several variables to describe them.

We want to identify a level shift in a time series. Let $x_{i,t}$ be the purchase amount of the ith client ($i = 1, \ldots, N$) , and $t = 1, \ldots, T$. As the series are expected to be seasonal, and this is confirmed by a peak in the autocorrelation at lag 12 in the series of loyal clients, we apply a multiplicative seasonally adjustment by computing the average purchase amount on month m for all the clients, $\overline{x}_m$, the total average purchase amount, $\overline{\overline{x}}$, and estimate the seasonal coefficients by the ratio of these averages, $S_m = \overline{x}_m/\overline{\overline{x}}$. These coefficients indicate a clear seasonal effect in August and smaller in February. The seasonally adjusted time series is $z_{i,t} = x_{i,t}/S_m$. We analyze the series in logs, the variability of the purchases depends on the average level, and call $y_{i,t} = \ln\left(z_{i,t} + 1\right)$, where we add one to avoid the problem with zero purchases. In order to find level shifts in these series we assume an $AR\,(1)$ model $y_{i,t} = \mu_i + \phi y_{i,t-1} + u_{i,t}$

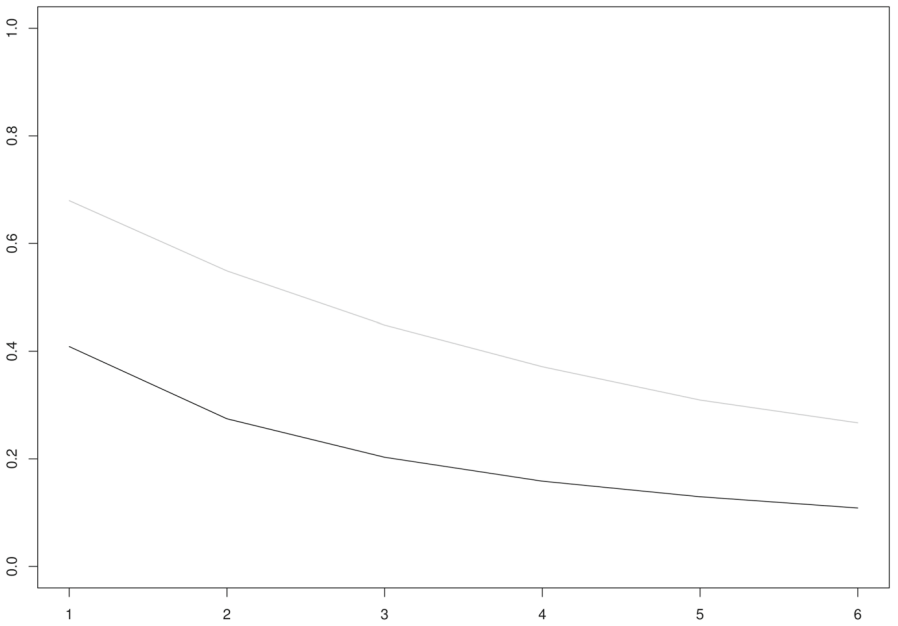

Fig. 8 Probability of being active next month after some months of inactivity for frequent clients (higher curve) and occasional clients (lower curve)

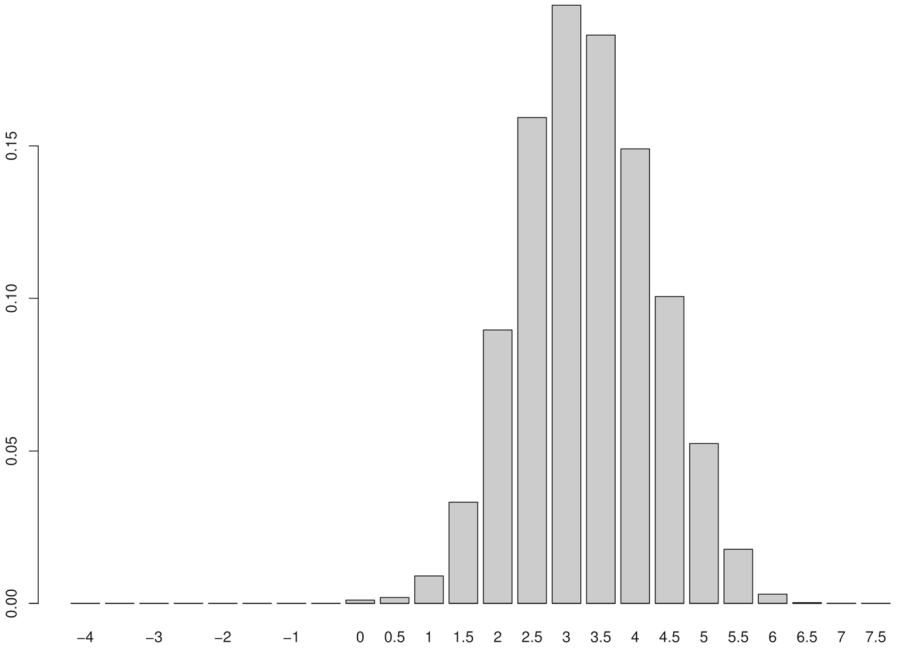

Fig. 9 Distribution of the log of the purchase amount

that (i) is consistent with the autocorrelation observed in most of the time series; (ii) allows a linear trend in the time series when $\phi = 1$. We apply to the series $y_{i,t}$ the algorithm for level shift detection explained in Peña et al. (2001, p. 156) with the following modifications. A window of at least L observations is required to start the search by comparing the mean of the residuals after the $AR(1)$ fit in this window to the mean of the next L observations. If a significant change with a t test is found, we kept the time of the level shift and move to the next observation to continue checking. If a level shift is not found, we increase by one the length of first window and continue checking.

As a result of this analysis, we define for each time $t = L + 1, \ldots, T$ a set of six dynamic variables for each client, M_{it}: (1) a dummy variable to indicate whether an increase in the level of the purchase amount has occurred before this time; (2) the number of identified increasing level shifts before this time; (3) the relative amount of the last increasing level shift before this time; (4) a dummy variable to indicate whether a decrease in the level of the purchase amount has occurred before this time; (5) the number of identified decreasing level shifts before this time; and (6) the relative amount of the last decreasing level shift. Note that these variables depend on the time t because they describe the history of the level shifts before this point.

We also analyze the number and length of the inactive runs for each client. For each inactive run, and a client may have several, we build a set of dynamic variables, A_{it}, depending on the time t in which the inactive run starts. These variables are: (1) a dummy variable to indicate whether there exist runs of no activity before the present one; (2) the proportion of months with activity before this time; (3) the length of the previous run; and (4) the number of inactive runs before this time. In the next section, we will see how to incorporate these variables to forecast future buying behavior.

4.2.4 Estimating probabilities of attrition for each client

Given the large set of clients to be considered, more than eight millions, and the need of a fast response of the company when a change is observed, we want to monitor every month only the clients that have shown some change in his purchase behavior with some probability p_0. The value of p_0 must be fixed taking into account the cost for the company of the two possible errors. In this case, we have selected $p_0 = 0.75$. Then, we assume that the next observation, $y_{i,t+1}$, shows evidence of change if any of the two following situations occurs: (1) $y_{i,t+1} = 0$ and the probability for this client of being inactive is smaller than p_0; and (2) $\left(y_{i,t+1} - \overline{y}_i\right)/s_i < -0.68$, that is the .25th percentile of the standard normal. For these clients with evidence of change, we will compute the probability of leaving the system.

We will consider first the situation when a significant new inactive observation arrives. Let $q_r^i(h)$ the probability that the ith client with a history of purchases summarized in the variables M_{it} and A_{it}, and personal variables C_i, remains inactive for an additional period of length h after observing an inactivity run of size r. We will estimate different models for different values of both parameters, r and h. The values chosen are $r = 1, 2, 3$ and $h = 1, \ldots, 6$, respectively. It is considered that, as shown in Fig. 7, after a run of nine $(r + h)$ inactive months the probability of buying is very small, below .1. Suppose we want to estimate $q_1^i(1)$. As clients in group A

by definition do not have inactive months, when this inactivity occurs for a client in this group it will be automatically classified in group F. Then we analyze all the runs of size one that happen in the period $(2 \leq t \leq T-1)$ and the runs of size two starting in t for $(2 \leq t \leq T-2)$ in clients in groups F and O. Clients in group F may have one or more runs of these length and clients of group O will probably have several. Then, we define a response variable for each run that will be 1 for runs of length two and 0 for runs of length one. In other words, if the run is of size one this implies that after one inactive month the client makes a purchase next month and becomes active, and the response is zero as he/she did not continue inactive. On the other hand, for runs of size two after observing a run of size one, the client continues inactive next month and the response after a run of size one is one. From this analysis, we conclude that a first estimator of the average value of $q_1^i(1)$ in the sample will be # (inactive runs of length 2) /# (inactive runs of length 1). The probabilities $q_1^i(1)$ are estimated with the logistic model

$$\log \frac{q_1^i(1)}{1-q_1^i(1)} = \beta_1' M_{it} + \beta_2' A_{it} + \beta_3' C_{it}$$

which is estimated in the data set formed by inactive runs of length one and two that has as response variable 0 or 1, as defined before, and as explanatory variables the set (M_{it}, A_{it}) corresponding at the time the run starts, and the C_i variables that depend on the client. A similar process is carried out to estimate $q_r^i(h)$. Then we consider inactive runs of size $r+h$ and $r+h-1$, estimate the average value of this probability by the ratio

$$\widetilde{q}_r(h) = \frac{\#(\text{inactive run of length } r+h)}{\#(\text{inactive runs of length } r+h-1)} \tag{12}$$

and the probabilities $q_r^i(h)$ are estimated with model (12) using the set of data build from inactive runs of size $r+h$ and $r+h-1$.

4.2.5 Results

A total of 72 logistic models were estimated by ML and Lasso, half of them correspond to frequent clients and the other half to occasional clients. The previous run length was from one to six, and the future periods without buying were also from one to six. Table 3 presents the precision of some of these models for frequent clients where h is the length of the predicted future inactivity run and r is the run of the observed inactive period. It is shown that the precision of these models decreases with r, is easier to forecast with one month of inactivity than with 3 months, and increase with h, it is easier to forecast clients that are going to have several months of inactivity, that is associated to a change of behavior in a frequent client, that to forecast next future month being inactive, that correspond to a more random behavior.

From the fitted models, we conclude that the probability of buying increases with (1) the average amount of the purchases; (2) the variability of the amount of purchases

Table 3 Precision of the fitted models for frequent clients

r/h	1	2	3	4	5	6
1	0.7	0.86	0.93	0.96	0.98	0.99
2	0.66	0.8	0.88	0.92	0.95	0.97
3	0.67	0.77	0.83	0.89	0.94	0.97

The values are for different future inactive periods (h) as a function of the length of the observed inactive run (r)

before the observed run; (3) a significant increase in the amount of purchases; and (4) using digital coupons. On the other hand, the probability of being inactive increases with (1) a significant decrease in the amount of purchases; (2) the amount of return of purchases; and (3) the use of financing of the purchases instead of paying by cash or credit card.

5 Conclusions

In this article, we have revised some of the changes that the Big Data revolution has produced in the analysis of data and in the role of statistics. The automatic generation of large amounts of data will increase in the future with the Internet of Things (IoT) and the decrease in the cost of sending and storing information. Images and videos will play a more central role as data information and statistics and operation research will be blended with machine learning and artificial intelligence to create prediction methods useful to analyze new types of information. Thus, it is important to create spaces to facilitate this interchange of ideas, as degrees on data science and research institutes of data science, Big Data or data learning, where people with different backgrounds, as applied mathematics, computer science, engineering, machine learning and statistics, work together.

The Big Data area is here to stay, and it will speed up learning in all fields of science. It is important that universities and research institutes promote joint appointments to facilitate interdisciplinary collaboration and stimulate the needed cross-fertilization among different fields. The experience of a century of data analysis has shown that procedures that have been designed for a specific problem in one field of application, as design of experiments in agronomy, censored data estimation in medicine or the Kalman filter in engineering, have found general applications in other areas. Thus, it is important that data science researchers have joint appointments in applied fields, but they must also work together in solving methodological problems that can be useful in many other fields of science.

Acknowledgements The invitation to write this article came from the editor Jesús López-Fidalgo and we are very grateful to him for his encouragement. The applications presented in this paper were carried out with Federico Liberatore, Lara Quijano-Sánchez and Carlo Sguera, post-docs at the UC3M-BS Institute of Financial Big Data. Iván Blanco and Jose Luis Torrecilla, also post-docs in the Institute, have also contributed with useful discussions. The ideas in this article have been clarified with the comments of Andrés Alonso, Anibal Figueiras, Rosa Lillo, Juan Romo and Rubén Zamar. To all them, our gratitude.

References

Aghabozorgi S, Shirkhorshidi AS, Wah TY (2015) Time-series clustering—a decade review. Inform Syst 53:16–38

Akaike H (1973) Information theory and an extension of the maximum likelihood method. In: Petrov N, Caski F (eds) Proceeding of the 2nd symposium on information theory. Academiai Kiado, Budapest, pp 267–281

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19:716–723

Alonso A, Peña D (2018) Clustering time series by linear dependency. Stat Comput. https://doi.org/10.1007/s11222-018-9830-6

Ando T, Bai J (2017) Clustering huge number of financial time series: a panel data approach with high-dimensional predictors and factor structures. J Am Stat Assoc 112(519):1182–1198

Arlot S, Celisse A (2010) A survey of cross-validation procedures for model selection. Stat Surv 4:40–79

Arribas-Gil A, Romo J (2014) Shape outlier detection and visualization for functional data: the outliergram. Biostatistics 15(4):603–619

Asimov D (1985) The grand tour: a tool for viewing multidimensional data. SIAM J Sci Stat Comp 6:128–143

Bai J, Ng S (2002) Determining the number of factors in approximate factor models. Econometrica 70(1):191–221

Bailey TC, Sapatinas T, Powell KJ, Krzanowski WJ (1998) Signal detection in underwater sound using wavelets. J Am Stat Assoc 93:73–83

Banfield JD, Raftery AE (1993) Model-based Gaussian and non-Gaussian clustering. Biometrics 49:803–821

Barabási AL (2016) Network Science. Cambridge University Press, Cambridge

Barber RF, Candès EJ (2015) Controlling the false discovery rate via knockoffs. Ann Stat 43(5):2055–2085

Basu S, Michailidis G (2015) Regularized estimation in sparse high-dimensional time series models. Ann Stat 43:1535–1567

Benito M, García-Portugués E, Marron JS, Peña D (2017) Distance-weighted discrimination of face images for gender classification. Stat 6(1):231–240

Benjamini Y (2010) Discovering the false discovery rate. J R Stat Soc B 72(4):405–416

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57(1):289–300

Bergmeir C, Benítez JM (2012) On the use of cross-validation for time series predictor evaluation. Inf Sci 191:192–213

Bertini E, Tatu A, Keim D (2011) Quality metrics in high-dimensional data visualization: an overview and systematization. IEEE Trans Vis Comput Graph 17:2203–2212

Besag J (1986) On the statistical analysis of dirty pictures. J R Stat Soc B 48(3):259–302

Bickel PJ, Levina E (2008) Regularized estimation of large covariance matrices. Ann Stat 36(1):199–227

Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp. https://doi.org/10.1088/1742-5468/2008/10/P10008

Bouveyron C, Brunet-Saumard C (2014) Model-based clustering of high-dimensional data: a review. Comput Stat Data Anal 71:52–78

Box GEP, Tiao GC (1968) A bayesian approach to some outlier problems. Biometrika 55(1):119–129

Breiman L (2001) Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat Sci 16:199–231

Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. Chapman and Hall/CRC, New York

Brockwell SE, Gordon IR (2001) A comparison of statistical methods for meta-analysis. Stat Med 20:825–840

Bühlmann P, van de Geer S (2011) Statistics for high-dimensional data: methods, theory and applications. Springer, Berlin, Heidelberg

Bühlmann P, van de Geer S (2018) Statistics for big data: a perspective. Stat Prob Lett 136:37–41

Bühlmann P, Drineas P, Kane M, van der Laan M (2016) Handbook of big data. Chapman and Hall/CRC, Boca Raton

Cai TT (2017) Global testing and large-scale multiple testing for high-dimensional covariance structures. Annu Rev Stat Appl 4:423–446

Cai TT, Liu W (2011) Adaptive thresholding for sparse covariance matrix estimation. J Am Stat Assoc 106:672–684
Cai TT, Liu W (2016) Large-scale multiple testing of correlations. J Am Stat Assoc 111:229–240
Cai TT, Zhuo HH (2012) Optimal rates of convergence for sparse covariance matrix estimation. Ann Stat 40(5):2389–2420
Cai TT, Liu W, Luo X (2011) A constrained ℓ_1 minimization approach to sparse precision matrix estimation. J Am Stat Assoc 106:594–607
Caiado J, Maharaj EA, D'urso P (2015) Time series clustering. In: Handbook of cluster analysis, CRC Press, pp 241–264
Cairo A (2016) The truthful art: data, charts, and maps for communication. New Riders
Candès E, Tao T (2006) Near-optimal signal recovery from random projections: universal encoding strategies. IEEE Trans Inf Theory 52:5406–5425
Candès E, Romberg JK, Tao T (2006) Stable signal recovery from incomplete and inaccurate measurements. Commun Pure Appl Math 52:1207–1223
Candès E, Li X, Ma Y, Wright J (2011) Robust principal component analysis? J ACM 58(3):11
Candès EJ, Fan Y, Janson L, Lv J (2016) Panning for gold: model-free knockoffs for high-dimensional controlled variable selection. Technical report, May 2016, Department of Statistics, Stanford University
Cao R (2017) Ingenuas reflexiones de un estadístico en la era del big data. Bol de Estad e Investig Oper 33(3):295–321
Carmichael I, Marron JS (2018) Data science vs. statistics: two cultures? Jpn J Stat Data Sci 1(1):117–138
Cerioli A, Farcomeni A, Riani M (2013) Robust distances for outlier-free goodness-of-fit testing. Comput Stat Data Anal 65:29–45
Chen CP, Zhang CY (2014) Data-intensive applications, challenges, techniques and technologies: a survey on big data. Inform Sci 275:314–347
Chen H, De P, Hu YJ, Hwang BH (2014) Wisdom of crowds: the value of stock opinions transmitted through social media. Rev Financ Stud 27(5):1367–1403
Chen J, Chen Z (2008) Extended Bayesian information criteria for model selection with large model spaces. Biometrika 95(3):759–771
Chernozhukov V, Galichon A, Hallin M, Henry M (2017) Monge–Kantorovich depth, quantiles, ranks and signs. Ann Stat 45(1):223–256
Cook RD (2018) An introduction to envelopes: dimension reduction for efficient estimation in multivariate statistics. Wiley, New York
Cook D, Buja A, Cabrera J, Hurley C (1995) Grand tour and projection pursuit. J Comput Graph Stat 4:155–172
Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
Cover TM, Hart PE (1967) Nearest neighbour pattern classification. IEEE Trans Inform Theory 13:21–27
Cuesta-Albertos JA, Gordaliza A, Matrán C (1997) Trimmed k-means: an attempt to robustify quantizers. Ann Stat 25(2):553–576
Cuevas A (2014) A partial overview of the theory of statistics with functional data. J Stat Plan Inference 147:1–23
Domingos P, Pazzani M (1997) On the optimality of the simple Bayesian classifier under zero-one loss. Mach Learn 29:103–130
Donoho D (2006a) Compressed sensing. IEEE Trans Inf Theory 52:1289–1306
Donoho D (2006b) For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution. Commun Pure Appl Math 59:797–829
Donoho D (2017) 50 years of data science. J Comput Graph Stat 26(4):745–766
Dryden IL, Hodge DJ (2018) Journeys in big data statistics. Stat Prob Lett 136:121–125
Efron B, Hastie T (2016) Computer age statistical inference. Cambridge University Press, Cambridge
Evergreen SDH (2016) Effective data visualization: the right chart for the right data. SAGE Publications
Faith J, Mintram R, Angelova M (2006) Targeted projection pursuit for visualizing gene expression data classifications. Bioinformatics 22:2667–2673
Fan J, Han F, Liu H (2014) Challenges of big data analysis. Natl Sci Rev 1(2):293–314
Forni M, Hallin M, Lippi M, Reichlin L (2005) The generalized dynamic factor model: one-sided estimation and forecasting. J Am Stat Assoc 100:830–840
Fraiman R, Justel A, Svarc M (2008) Selection of variables for cluster analysis and classification rules. J Am Stat Assoc 103:1294–1303

Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9(3):432–441
Frühwirth-Schnatter S (2006) Finite mixture and Markov switching models. Springer, New York
Galeano P, Peña D (2019) Outlier detection in high-dimensional time series (**Unpublished manuscript**)
Galeano P, Peña D, Tsay RS (2006) Outlier detection in multivariate time series by projection pursuit. J Am Stat Assoc 101:654–669
Galimberti G, Manisi A, Soffritti G (2017) Modelling the role of variables in model-based cluster analysis. Stat Comput 28(1):1–25
Gandomi A, Haider M (2015) Beyond the hype: big data concepts, methods, and analytics. Int J of Inf Manage 35(2):137–144
García-Ferrer A, Highfield RA, Palm F, Zellner A (1987) Macroeconomic forecasting using pooled international data. J Bus Econ Stat 5:53–67
Geisser S (1975) The predictive sample reuse method with applications. J Am Stat Assoc 70:320–328
Genton MG (2001) Classes of kernels for machine learning: a statistics perspective. J Mach Learn Res 2:299–312
Genton MG, Johnson C, Potter K, Stenchikov G, Sun Y (2014) Surface boxplots. Stat 3(1):1–11
Genton MG, Castruccio S, Crippa P, Dutta S, Huser R, Sun Y, Vettori S (2015) Visuanimation in statistics. Stat 4(1):81–96
Giannone D, Reichlin L, Small D (2008) Nowcasting: the real-time informational content of macroeconomic data. J Monet Econ 55:665–676
Gómez V, Maravall A (1996) Programas tramo and seats. Documento de Trabajo, Banco de España SGAPE-97001
Guhaniyogi R, Dunson DB (2015) Bayesian compressed regression. J Am Stat Assoc 110:1500–1514
Hall P, Marron JS, Neeman A (2005) Geometric representation of high dimension, low sample size data. J R Stat Soc B 67(3):427–444
Härdle WK, Lu HHS, Shen X (2018) Handbook of big data analytics. Springer
Hastie T, Pregibon D (1992) Generalized linear models. In: Chambers JM, Hastie TJ (eds) Statistical models in S, Chap 6. Wadsworth & Brooks/Cole
Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. Springer, New York
Hastie T, Tibshirani R, Wainwright M (2015) Statistical learning with sparsity: the lasso and generalizations. Chapman and Hall/CRC, Boca Raton
Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12:55–67
Hornik K (1991) Approximation capabilities of multilayer feedforward networks. Neural Netw 4:251–257
Huber PJ (1964) Robust estimation of a location parameter. Ann Math Stat 35(1):73–101
Hyvärinen A, Oja E (2000) Independent component analysis: algorithms and applications. Neural Netw 13:411–430
Irizarry RA (2001) Local harmonic estimation in musical sound signals. J Am Stat Assoc 96:357–367
Jain AK (1989) Fundamentals of digital image processing. Prentice Hall, Englewood Cliffs, NJ
James W, Stein C (1961) Estimation with quadratic loss. In: Proceedings of 4th Berkeley symposium on mathematical statistics and probability, vol I, University of California Press, pp 361–379
Johnstone IM, Titterington DM (2009) Statistical challenges of high-dimensional data. Philos Trans R Soc A 367:4237–4253
Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 53:457–481
Kaufman L, Rousseeuw PJ (1990) Finding groups in data: an introduction to cluster analysis. Wiley, New York
Kokoszka P, Reimherr M (2017) Introduction to functional data analysis. Chapman and Hall/CRC, Boca Raton
Kolaczyk ED (2009) Statistical analysis of network data. Springer, New York
Kriegel HP, Kröger P, Zimek A (2009) Clustering high-dimensional data: a survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM Trans Knowl Discov Data 3(1):1
Lam XY, Marron JS, Sun D, Toh KC (2018) Fast algorithms for large-scale generalized distance weighted discrimination. J Comput Graph Stat 27(2):368–379
Lauritzen SL (1996) Graphical Models. Oxford University Press Inc., New York
LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444

Liu W (2013) Gaussian graphical model estimation with false discovery rate control. Ann Stat 41(6):2948–2978
López-Pintado S, Romo J (2009) On the concept of depth for functional data. J Am Stat Assoc 104:718–734
Lu X, Marron JS, Haaland P (2014) Object-oriented data analysis of cell images. J Am Stat Assoc 109:548–559
MacQueen J (1967) Some methods for classification and analysis of multivariate observations. Proceedings of the 5th Berkeley symposium on mathematical statistics and probability vol 1, pp 281–297
Majumdar A (2009) Image compression by sparse PCA coding in curvelet domain. Signal Image Video Process 3:27–34
Maronna RA, Martin RD, Yohai V, Salibián-Barrera M (2019) Robust statistics: theory and methods (with R), 2nd edn. Wiley, Hoboken, NJ
Meinshausen N, Bühlmann P (2006) High dimensional graphs and variable selection with the lasso. Ann Stat 34(3):1436–1462
Mosteller F, Wallace DL (1963) Inference in an authorship problem: a comparative study of discrimination methods applied to the authorship of the disputed federalist papers. J Am Stat Assoc 58:275–309
Munzner T (2014) Visualization analysis and design. Chapman and Hall/CRC, Boca Raton
Norets A (2010) Approximation of conditional densities by smooth mixtures of regressions. Ann Stat 38(3):1733–1766
de Oliveira MF, Levkowitz H (2003) From visual data exploration to visual data mining: a survey. IEEE Trans Vis Comput Graph 9:378–394
Pan W, Shen X (2007) Penalized model-based clustering with application to variable selection. J Mach Learn Res 8:1145–1164
Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr 2:1–135
Paradis L, Han Q (2007) A survey of fault management in wireless sensor networks. J Netw Syst Manag 15:171–190
Peña D (2014) Big data and statistics: trend or change. Bol de Estad e Investig Oper 30:313–324
Peña D, Box GEP (1987) Identifying a simplifying structure in time series. J Am Stat Assoc 82:836–843
Peña D, Poncela P (2004) Forecasting with nonstationary dynamic factor models. J Econom 119(2):291–321
Peña D, Prieto FJ (2001a) Cluster identification using projections. J Am Stat Assoc 96:1433–1445
Peña D, Prieto FJ (2001b) Robust covariance matrix estimation and multivariate outlier detection. Technometrics 43:286–310
Peña D, Sánchez I (2005) Multifold predictive validation in armax time series models. J Am Stat Assoc 100:135–146
Peña D, Tiao GC, Tsay RS (2001) A course in time series analysis. Wiley, Hoboken, NJ
Peña D, Viladomat J, Zamar R (2012) Nearest-neighbors medians clustering. Stat Anal Data Min 5(4):349–362
Peña D, Smucler E, Yohai VJ (2019a) Forecasting multiple time series with one-sided dynamic principal components. J Am Stat Assoc. https://doi.org/10.1080/01621459.2018.1520117
Peña D, Tsay RS, Zamar R (2019b) Empirical dynamic quantiles for visualization of high-dimensional time series. Technometrics. https://doi.org/10.1080/00401706.2019.1575285
Pigoli D, Hadjipantelis PZ, Coleman JS, Aston JAD (2018) The statistical analysis of acoustic phonetic data: exploring differences between spoken romance languages (with discussion). J R Stat Soc C 67:1–27
Quijano-Sánchez L, Liberatore F (2017) The big chase: a decision support system for client acquisition applied to financial networks. Decis Support Syst 98:49–58
Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77:257–286
Radke RJ, Andra S, Al-Kofahi O, Roysam B (2005) Image change detection algorithms: a systematic survey. IEEE Trans Image Process 14:294–307
Raftery AE, Dean N (2006) Variable selection for model-based clustering. J Am Stat Assoc 101:168–178
Ramsay JO, Silverman BW (2005) Functional data analysis, 2nd edn. Springer, New York
Ren Z, Sun T, Zhang CH, Zhou HH (2015) Asymptotic normality and optimalities in estimation of large gaussian graphical model. Ann Stat 43(3):991–1026
Riani M, Atkinson AC, Cerioli A (2009) Finding an unknown number of multivariate outliers. J R Stat Soc B 71(2):447–466
Riani M, Atkinson AC, Cerioli A (2012) Problems and challenges in the analysis of complex data: static and dynamic approaches. In: di Ciaccio A, Coli M, Angulo JM (eds) Advanced statistical methods for the analysis of large data-sets. Springer, Berlin, Heidelberg, pp 145–157

Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev 65(6):386–408
Rousseeuw P, van den Bossche W (2018) Detecting deviating data cells. Technometrics 60(2):135–145
Ryan TP, Woodall WH (2005) The most-cited statistical papers. J Appl Stat 32(5):461–474
Samuel AL (1959) Some studies in machine learning using the game of checkers. IBM J Res Dev 3:210–229
Schölkopf B, Smola A, Müller KR (1997) Kernel principal component analysis. In: Gerstner W, Germond A, Hasler M, Nicoud JD (eds) Artificial Neural Networks ICANN'97, vol 1327. Lecture Notes in Computer Science, pp 583–588
Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464
Sesia M, Sabatti C, Candès EJ (2018) Gene hunting with knockoffs for hidden Markov models. Biometrika. https://doi.org/10.1093/biomet/asy033
Shao J (1993) Linear model selection by cross-validation. J Am Stat Assoc 88:486–494
Shen H, Huang JZ (2008) Sparse principal component analysis via regularized low rank matrix approximation. J Multivariate Anal 99(6):1015–1034
Shi JQ, Choi R (2011) Gaussian process regression analysis for functional data. CRC Press, Boca Raton
Small C (1990) A survey of multidimensional medians. Int Stat Rev 58:263–277
Stock JH, Watson MW (2002) Forecasting using principal components from a large number of predictors. J Am Stat Assoc 97:1167–1179
Stone M (1974) Cross-validatory choice and assessment of statistical predictions. J R Stat Soc B 36(2):111–147
Stone M (1977) An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. J R Stat Soc B 39(1):44–47
Sun Y, Genton MG (2011) Functional boxplots. J Comput Graph Stat 20(2):316–334
Tausczik YR, Pennebaker JW (2010) The psychological meaning of words: Liwc and computerized text analysis methods. J Lang Soc Psychol 29:24–54
Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc B 12:267–288
Tong H (2012) Threshold models in non-linear time series analysis. Springer, New York
Tong H, Lim KS (1980) Threshold autoregression, limit cycles and cyclical data (with discussion). J R Stat Soc B 42(3):245–292
Torrecilla JL, Romo J (2018) Data learning from big data. Stat Prob Lett 136:15–19
Tsay RS, Chen R (2018) Nonlinear time series analysis. Wiley, Hoboken, NJ
Tukey JW (1970) Exploratory data analysis. Addison-Wesley Pub, Co, Reading, MA
Tzeng JY, Byerley W, Devlin B, Roeder K, Wasserman L (2003) Outlier detection and false discovery rates for whole-genome DNA matching. J Am Stat Assoc 98:236–246
Vidal R (2011) Subspace clustering. IEEE Signal Proc Mag 28:52–68
Wang S, Zhu J (2008) Variable selection for model-based high-dimensional clustering and its application to microarray data. Biometrics 64:440–448
Wei F, Tian W (2018) Heterogeneous connection effects. Stat Prob Lett 133:9–14
Witten DM, Tibshirani R (2010) A framework for feature selection in clustering. J Am Stat Assoc 105:713–726
Witten DM, Tibshirani R, Hastie T (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10(3):515–534
Xia Y, Cai T, Cai TT (2016) Testing differential networks with applications to detecting gene-by-gene interactions. Biometrika 102:247–266
Yang Y (2005) Can the strengths of aic and bic be shared? A conflict between model identification and regression estimation. Biometrika 92:937–950
Zhang P (1993) Model selection via multifold cross validation. Ann Stat 21(1):299–313
Zhao SD, Cai TT, Li H (2014) Direct estimation of differential networks. Biometrika 101:253–268
Zhou Z, Wu WB (2009) Local linear quantile estimation for nonstationary time series. Ann Stat 37:2696–2729
Zhu X, Pan R, Li G, Liu Y, Wang H (2017) Network vector autoregression. Ann Stat 45(3):1096–1123