

Problems and Challenges in the Analysis of Complex Data: Static and Dynamic Approaches

Marco Riani¹

Dept. of Economics, University of Parma, Italy, e-mail: mriani@unipr.it

Abstract: This paper summarizes results in the use of the Forward Search in the analysis of corrupted datasets, or those with mixtures of populations. Methods developed for regression and clustering are described. However, the emphasis is on new challenges that arise in the analysis of large, complex datasets. These are motivated through an example in fraud detection. Much of this research comes from joint work with Anthony Atkinson, Andrea Cerioli, Domenico Perrotta and Francesca Torti.

Keywords: Robust estimates, forward search, mixtures, complex data

1. Introduction

Data are an overwhelming feature of modern life. As the amount of data increases so do the challenges facing the statistician in trying to extract information from ever larger data sets. One of the arguments of the talk, of which this is a summary, is that larger data sets are also more complex and require flexible multiple analyses in order to reveal their structure. Only then can information be efficiently extracted.

The analysis of large data sets may be complicated by the high dimensionality of responses, large numbers of observations and complexity of the choices to be made among explanatory variables. Although appreciable, these challenges to the statistician are not different in kind from those faced in the analysis of smaller data sets. In my talk I will focus on problems that become severe in large complex data sets, such as our inability to find a single model for all the data.

The simplest situation is that of a single model with possibly many outliers. In the presence of a core population and some isolated or clustered outliers, traditional robust methods (Maronna *et al.*, 2006) can be successfully used to find proper models. However, when there are several populations and different sources of heterogeneity, traditional robust methods fail to recover the real structure of the data and more sophisticated procedures, such as those derived from the Forward Search (FS) (Atkinson and Riani, 2000; Atkinson, Riani and Cerioli, 2004) are required. Some examples are presented in the next section. Section 3 introduces an example of complex data in a situation where automatic procedures need to be developed.

2. Some Difficulties in Data Analysis

The Presence of Outliers. The presence of atypical observations may strongly and wrongly influence the output of statistical analyses. When the number of observations

¹ Address of correspondence: Section of Statistics and Computing, Faculty of Economics, Via Kennedy 6, 43100 Parma, Italy

is large it is likely that there will be several atypical observations which mask one another. They will not be revealed by a single static analysis, although the dynamic analysis of many subsets of data through the FS will reveal such structure. However, the outliers should not be seen only as bad observations that estimation procedures must avoid; they may themselves contain valuable information. The discovery of the hole in the ozone layer is one example. In drug development, the existence of a subset of individuals with an adverse reaction to the drug might be one target of the analysis.

Calibration of Test Procedures. Particular as the sample size grows, it is necessary to calibrate tests of the outlyingness of individual observations. The repeated application of statistical tests makes it necessary to calibrate for simultaneity. Use of procedures that are correctly calibrated to provide tests of the desired size will keep false alarms under control (Riani *et al.* 2009).

Subpopulations. Large datasets often contain hidden groups, which are not revealed by application of single population methods, even in their robustified forms. For multivariate data there are well established techniques of cluster analysis, which may work well for normal populations. However, automatic methods such as MCLUST (Fraley and Raftery, 1999) for establishing cluster membership often indicate too many clusters. Standard clustering procedures are also badly affected by the presence of outliers. Procedures based on the forward search have been shown to work well in identifying clusters and establishing cluster membership, even in the presence of outliers, but are far from automatic, requiring appreciable input from the statistician.

3. An Example of Large Complex Corrupted Data

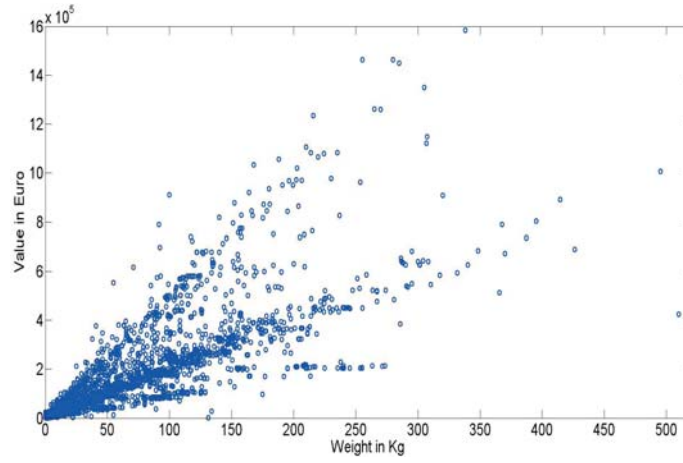
As an illustration of the problems involved with the analysis of complex data, consider the example given in Figure 1 referred to the quantity (x) and the value (y) of 4719 import declarations of a specific technological product². This is an example of one of the thousands of datasets provided by the “Office Européen de Lutte Anti-Fraude” (OLAF) or by its partners in the Member States. The purpose is to find atypical transactions, which might correspond to cases of potential fraud (e.g. the evasion of import duties) or to potential money laundering activities.

The observations appear roughly distributed along three main different lines departing from the origin of the coordinate axes. However, there seem also to be horizontal strips of concentrated data. It is certainly not clear how many groups are present in the data. Nonetheless, it is obvious that traditional models which assume one single regression population will fail in revealing the real structure as will their robust counterparts.

The general structure is of a mixture of linear models heavily contaminated by observations that do not follow the general pattern (Riani *et al.*, 2008). Outliers may be isolated, originating from recording errors during the data collection process, or they may be clustered, when they represent some systematic behaviour. In the context of international trade data, useful information for fraud detection purposes may come from outliers that cluster themselves around an “anomalous” regression line. Also inliers that do not follow any specific regression structure and are intermediate among different lines may be of interest, especially if they represent economically important transactions, that is those

²More details about this dataset can be found in the paper by Perrotta and Torti given in this volume.

Figure 1: *An example of international trade data.*



with large values of either variable. Of course, in the absence of fraud, large values of both variables occur together. The context of anti-fraud analysis nicely puts in context the need for calibration of the test procedures mentioned above. A procedure with too large a size will result in excessive indications of fraudulent activities. If these indications are followed up to the point of prosecutions which fail, encouragement will be given to fraudsters while law enforcement agencies will become discouraged. In general, complex data have an intrinsic multidimensional nature and the two dimensional projections that we can plot are not able to reveal their real structure. In this example, the information about the price and quantity of transactions is generally coupled with additional information from additional variables and it is not clear what is the functional relationship among the different indicators.

Use of the “multiple random starts forward search” (Atkinson and Riani, 2007) enables us to dissect these data into components and outliers. However, the clustering of regression lines is again a procedure that involves considerable statistical intervention. The context of anti-fraud indicates several important directions for statistical development.

4. Forward Directions for the Forward Search

Automatic Classification Procedures. While for a small number of datasets it is possible to envisage human intervention for each dataset, including the use of exploratory data analysis tools, in the presence of a huge amount of data only automatic procedures are feasible. These developments are required both for the clustering of multivariate data mentioned in §2 and for the mixtures of regression lines of §3.

Interactive Dynamic Plots. The procedures of the forward search provide a graphics rich environment for the analysis of data. However, once a unified statistical methodology is established for data analysis, the challenge is to provide powerful interactive dynamic plots incorporated in user friendly graphical interfaces. The hope is to extend the classical tools of exploratory data analysis, first of all by means of dynamic linking and interaction among different plots, but also through simple representations of statistical information

which may be easily interpreted by the end-users of the statistical analyses (in the spirit of Tufte 2001, and Spence, 2001).

Timeliness and On-Line Systems. Typical analysis of the presence of potential outliers is done completely offline. It is more realistic to consider the challenge of on-line operations, where the data are analysed as soon as they become available and the models are updated just on the basis of the new data rather than re-estimated using the whole extended dataset. The availability of on-line analysis tools also responds to the need for timeliness. If a fraud is being committed it needs to be detected and prevented as quickly as possible. An example includes trade data in which two lines can be detected. Both have slopes lower than that for the bulk of the data. However, the lower of the two slopes is more recent. A second challenge in on-line analysis is to disseminate the results in a form that is again understandable by the final users. The importance of timeliness and on-line systems accentuates the need for research into the theoretical and practical aspects of dynamic updating methods.

Automatic Model Selection Procedures. For simple regression models with several explanatory variables the FS provides a robust form of the C_p statistic for selection of regression models (Atkinson and Riani, 2008). The resultant plot can be used for selection of a good model, but automatic procedures need developing.

However, for high-dimensional data, the phrase “model selection” refers not only to the choice of the functional form between the explanatory variables and the response, but also to the choice of the most important predictors.

References

- Atkinson A.C. and Riani M. (2000) *Robust Diagnostic Regression Analysis*, Springer–Verlag, New York.
- Atkinson A.C. and Riani M. (2007) Exploratory tools for clustering multivariate data, *Computational Statistics and Data Analysis*, 52, 272–285, doi:10.1016/j.csda.2006.12.034.
- Atkinson A.C. and Riani M. (2008) A robust and diagnostic information criterion for selecting regression models, *Journal of the Japanese Statistical Society*, 38, 3–14.
- Atkinson A.C., Riani M. and Cerioli A. (2004) *Exploring Multivariate Data with the Forward Search*, Springer–Verlag, New York.
- Fraley C. and Raftery A.E. (1999) Mclust: software for model-based cluster analysis, *Journal of Classification*, 16, 297–306.
- Maronna R.A., Martin R.D. and Yohai V.J. (2006) *Robust Statistics: Theory and Methods*, Wiley, Chichester.
- Riani M., Atkinson A.C. and Cerioli A. (2009) Finding an unknown number of multivariate outliers, *Journal of the Royal Statistical Society, Series B*, 71, 201–221.
- Riani M., Cerioli A., Atkinson A., Perrotta D. and Torti F. (2008) Fitting mixtures of regression lines with the forward search, in: *Mining Massive Data Sets for Security*, Fogelman-Soulie F., Perrotta D., Piskorski J. and Steinberger R., eds., IOS Press, Amsterdam (Netherlands), 271–286.
- Spence R. (2001) *Information Visualization*, Addison Wesley, California.
- Tufte E.R. (2001) *The Visual Display of Quantitative Information*, CT:Graphics Press, Cheshire.