

Doctoral Dissertation

Advances in the Forward Search:
methodological and applied contributions

Francesca Torti

Universita' degli Studi di Milano-Bicocca

Facolta' di Statistica

04 December 2009

Supervisor: Professor Marco Riani, Universita' degli Studi di Parma

Jury: Professor Pier Alda Ferrari, Universita' degli Studi di Milano
Professor Annibale Biggeri, Universita' degli Studi di Firenze
Professor Enrico Gori, Universita' degli Studi di Udine

I thank *Marco Riani* for his constant guide and *Domenico Perrotta* without whom I would have never started and finished my Ph.D studies.

I also thank *Spyros Arsenis* for having kindly supported my studies during my stay at the Joint Research Centre, *Andrea Cerioli* and *Sergio Zani* for their feedback on difficult topics and constant encouragement.

Contents

Contents	iii
List of Figures	v
1 Introduction	1
2 Challenges in the Analysis of Complex Data	7
2.1 Introduction	7
2.2 An example of complex corrupted data	7
2.3 Key problems addressed in the thesis	10
2.4 Next challenges	15
3 The Forward Search	19
3.1 Introduction	19
3.2 The data-driven flexible trimming approach	20
3.3 Key phases in the Forward Search history	22
3.4 Searching outliers backward and forward	25
3.5 The inferential approach to the forward detection of outliers	26
The forward progression	26
Monitoring the search	27
Testing for outliers	30
Using the envelopes	31
3.6 On the nominal size of the Forward Search	34
3.7 On the simultaneity of the Forward Search tests	37
4 Statistical Performances of the Forward Search in Regres-	
sion	39
4.1 Introduction	39
4.2 Reference robust regression techniques	40
4.3 A backward iterative outlier method	42
4.4 Benchmark setup for the statistical size	43

4.5	Results on the statistical size	45
4.6	Results on the statistical size for small samples	47
4.7	Benchmark set-up for the statistical power	49
4.8	Results on the statistical power	50
4.9	Results on the statistical power for small samples	54
4.10	The statistical power of the backward method	58
4.11	Considerations on computational efficiency	58
4.12	Conclusions	60
5	Extending the Forward Search to Linear Mixtures: a Proposal	63
5.1	Introduction	63
5.2	Tentative allocation of the groups	65
	Basic idea: repeated group identification and trimming . . .	65
	Adaptation through re-weighting the Forward Search	68
	Adaptation through R^2 control	70
5.3	Confirmation of groups and outliers	72
5.4	Experimenting different mixture solutions on a complex dataset	75
5.5	Testing the new Forward Search solution	78
	Tentative components in original-scale data	78
	Tentative components in the logarithmic-scale data	81
	Final components in the log-transformed data	82
	Tentative and final components with the R^2 -relaxed strategy	83
	Components driven by the expert opinion	83
5.6	Testing the Clusterwise Linear Regression solution	85
5.7	Testing the Trimmed Likelihood Estimator solution	87
5.8	Testing the Fixed Point Clustering solution	87
5.9	Towards a rigorous assessment of mixture estimation methods	89
5.10	Discussion and back to the perfect fit problem	91
6	Exploring and interacting with forward plots	97
6.1	Introduction	97
6.2	From static to dynamic linked robust forward plots	98
7	Conclusions and Future Perspectives	103
	Appendix to Chapter 3	107
	Algorithmic Details of the Forward Search Inference Approach	109
A	Introduction and formalism	109
B	Stage 1a: signal detection	110

C	Stage 1b: signal validation	112
D	Stage 2: superimposition of the envelopes	113
E	Bonferroni signal	114
Appendix to Chapter 6		115
Forward Search Data Analysis Matlab toolbox		117
A	Introduction	117
B	Robust analysis of complex data sets	117
B.1	Robust regression techniques	118
B.2	Transformations	119
B.3	Variable selection	121
C	Data visualization functions	122
Bibliography		131

List of Figures

2.1	Technological product scatterplot	8
2.2	Price outliers in a trade dataset	10
2.3	Outliers and leverage points	12
2.4	Studentised deletion residuals plot of Perfect Fit data	14
2.5	Spatial and time correlation in Fishery data	16
3.1	Perfect Fit data forward plots	28
3.2	Forward plot of minimum deletion residuals with envelopes . . .	32
3.3	Forward plot of minimum deletion residuals with envelopes for different sample sizes	33
3.4	The Bonferroni bounds of the minimum deletion residual	38
4.1	Table of the empirical size for FS	45
4.2	Table of the empirical size for LTS and LMS	46
4.3	Empirical size	46
4.4	Table of the empirical size for small sample sizes	48
4.5	Empirical size for small sample sizes	49
4.6	Table of the measures of power for $n = 500$ and $p = 6$	51

4.7	Measures of power for $n = 500$ and $p = 6$	52
4.8	Logit of measures of power for $n = 500$ and $p = 6$	53
4.9	Table of measures of power for $n = 50$ and $p = 2$	55
4.10	Measures of power for $n = 50$ and $p = 2$	56
4.11	Logit of measures of power for $n = 50$ and $p = 2$	57
4.12	Average power of the backward method	59
4.13	Average elapsed time in seconds for all methods	60
5.1	FS tentative mixture components in Fishery data	67
5.2	FS tentative mixture components in Fishery data (separate plots)	67
5.3	FS re-weighted tentative mixture components in Fishery data .	69
5.4	FS R^2 -relaxed tentative mixture components in Fishery data . .	72
5.5	FS confirmed mixture components in Fishery data	73
5.6	FS confirmed mixture components in Fishery data (separate plots)	74
5.7	Scatterplot of technological product data (Tech data)	75
5.8	FS tentative mixture components in the Tech data	79
5.9	FS R^2 -relaxed tentative mixture components in Tech data . . .	81
5.10	FS tentative mixture components in the log transformed Tech data	82
5.11	FS confirmed mixture components in the log transformed Tech data	83
5.12	FS R^2 -relaxed mixture components in Tech data	84
5.13	FS mixture components driven by the expert opinion in Tech data	85
5.14	Clusterwise Linear Regression on Tech data	86
5.15	Clusterwise Linear Regression on log-transformed Tech data . .	86
5.16	Trimmed Likelihood Estimator on Tech data	88
5.17	Trimmed Likelihood estimator on log-transformed Tech data . .	88
5.18	Fixed Point Clustering on Tech data	90
5.19	Fixed Point Clustering on log-transformed Tech data	90
5.20	Perfect fit dataset. Tentative components detected by the FS . .	93
5.21	Perfect fit dataset. Tentative components detected by the R^2 - relaxed mixture algorithm	93
5.22	Perfect fit dataset. Tentative components detected by FS-reweighed	94
5.23	Perfect fit dataset. Outliers detected by FS and by LTS	95
5.24	Perfect fit dataset. Outliers detected by the Backward algorithm	95
6.1	Fan plot of Fishery Trade Data	99
6.2	Fan plot of Fishery Trade Data (log transformation)	99
6.3	Table and scatterplot of Fishery Trade Data	101
6.4	Scaled residuals and minimum deletion residuals plots of Fishery Trade Data	102

A	FSDA Help	118
B	LTS function	119
C	FSR function	120
D	FSReda function	121
E	FSRbsb function	122
F	FSRmdr function	123
G	Score function	124
H	FSRfan function	125
I	FSRaddt function	126
J	resplot function	127
K	minplot function	128
L	yXplot function	129

Introduction

This thesis is about the Forward Search, an approach to robust data analysis that I have addressed in the regression context and for problems of general interest.

In statistics the basic concepts of the Forward Search can be dated back to 1992 (Hadi, [Had92]). The same key ideas were also introduced in the field of image analysis at the beginning of the '80s with a method known as Random Sample Consensus (RANSAC [BF81]). The first two comprehensive treatments in the regression and multivariate contexts were published respectively in 2000 and in 2004 in the books of Atkinson and Riani [AR00] and Riani, Atkinson and Cerioli [ARC04]. In the last few years the bibliography on the subject and the number of authors that study or apply the method are expanding.

The Forward Search was proposed by the authors as a general method for detecting unidentified subsets and masked outliers in complex data and for determining their effect on models fitted to the data. The peculiarity that characterises the Forward Search from other robust methods is that the amount of data used to produce robust estimates, i.e. estimates not distorted by abnormal data, is not fixed in advance and is chosen conditionally on the data. This is obtained through a basic idea: to monitor how the fitted model changes whenever a new observation is added to a subset of the data. A sequence of subsets of increasing size is built by moving forward from a basic subset and by choosing observations in the next subset with ordering criteria that in general are very simple to implement and quick to execute.

I have addressed the Forward Search, its application potential and distinctive features with respect to other robust and non-robust methods during the last years of my service in the European Commission's Joint Re-

search Centre, where I worked on anti-fraud problems related to European trade.

The protection of the budget of the European Community against fraud is an obligation for the Commission and the Member States. At this goal, the European Anti-fraud Office and the European Commission’s Joint Research Centre routinely collect data sets including millions of trade flows grouped in a large number of small to moderate size samples. These data samples are then analysed with the purpose of detecting anomalies of various kinds (e.g. recording errors), specific market price dynamics (e.g. discounts in trading big quantities of product) and cases of unfair competition or fraud. In general these behaviours can be re-conducted to few statistical patterns, in particular outliers and mixtures of linear regression components.

In analysing the European trade data I observed several levels of complexity. In the following paragraphs I introduce some of them and I elaborate on how in the thesis the Forward Search relates to these issues.

First of all in anti-fraud, as in many other applications, it is necessary to quantify how aberrant is an observation with respect to the rest of the data, i.e. to summarize its “outlyingness” through a single measure such as a p -value from a test statistic. This guarantees routine application of the methods and comparability of outlier signals detected in different datasets. In this process hundreds of diagnostic tests must be computed within thousands of samples and a very severe problem of multiplicity occurs. *I will give special attention on how the Forward Search copes with the multiplicity problem in chapter 3*, when I will present the Forward Search in regression with focus on the main inferential results derived in the last years. The chapter is complemented by the Appendix of the dissertation, which describes the steps of the Forward Search outlier detection algorithm *with a level of detail that, for regression, was not addressed so far in the literature*. Note that the algorithm described in the Appendix was ported to regression strictly following the multivariate version described by [RAC09], with surprisingly good results. However, in the dissertation *I argue that the detection algorithm for regression outliers has to be slightly modified to address more satisfactorily two specific patterns that I describe in Chapter 2*, namely the almost “perfect fit” and the need to merge homogeneous groups which are “too close”. I propose a possible modification of the outlier detection algorithm and a natural approach to use it in Chapter 5.

Secondly, a basic requirement in analysing trade data for anti-fraud purposes is that the statistical method must be *reliable*. This means that the number of false positives produced by the method must be limited, for not providing the investigators with an intractable number of cases

to inspect, but at the same time the ability of detecting truly anomalous transactions must be high for minimising the risk of omitting potential frauds. In other words, in anti-fraud it is crucial to adopt techniques that can combine high power for outlier detection with (small) size values that guarantee consistency when the data are not contaminated. Satisfactory performance on these grounds translates into high rewards when frauds are correctly detected, and reduced efforts by the anti-fraud staff if only few non-anomalous transactions are examined. This strong requirement is not common to many other application contexts and, in fact, in general robust statistics gives much more stress on the power of a detection method than on its size. In the multivariate context the actual size and power of the Forward Search was studied by [RAC09], but no equivalent extensive evaluation exists for regression in the literature. *I filled this gap in chapter 4 of the thesis, with a rigorous assessment of the actual size and different power measures achieved by the Forward Search and by other robust and non robust outlier test methods in regression.*

Thirdly, trade analysis typically involves a very large number of datasets and a diagnostic method, to be applicable in this context, should run *automatically* on millions of cases and should be sufficiently flexible to detect different kinds of patterns at the same time, in particular isolated or groups of outliers and mixtures of linear components. *In chapter 5 I will propose an extension of the Forward Search algorithm for identifying at the same time mixture of linear components and outliers.* The approach is articulated in a detection step in which the components are tentatively identified and in a confirmation step in which the classification of each observation to a tentative group is validated. The new approach will be demonstrated on a specific international trade dataset containing different homogeneous sub-populations and outliers. I will discuss the features of the approach in relation with other three known methods of linear mixture estimation [DC88], [Hen02], [NFDN07]. Of course it is not possible to draw general conclusions on any clustering method on the basis of specific data in a specific application context. Nonetheless, this thesis shows the very good ability of the Forward Search to treat complex data of this kind.

Another important problem in anti-fraud has to do with the ability to effectively present and communicate statistical results to the anti-fraud subject matter experts, who need to go beyond the work of the statisticians and investigate on the nature of the outliers or other relevant detected patterns. Moreover, one of the problems of the Forward Search has always been the lack of an automatic link among the great variety of forward plots, in which the evolution of many statistics are monitored as a function of the subset size. These two apparently different problems have in common the

need to extract information from the forward plots or any other graphical or tabular representation of statistical data. *I will dedicate a chapter of the thesis to a set of new robust graphical tools that I have contributed to develop in view of providing the Forward Search with interactive and flexible data exploration instruments.* The Forward Search, when introduced, was initially used more as an exploratory data analysis tool; then for several years it has progressed with considerable achievements as an inferential tool, but for many researchers its major appeal is still in the powerful exploratory data analysis capacity. I hope that the new robust graphical and inferential tools, which absorbed a considerable amount of energies in the last eighteen months, will enable a wider adoption of the Forward Search in the statistical community and a full appreciation of its potential in the analysis of complex data.

The thesis is divided in 5 chapters.

- Chapter 2 summarises the main challenges that I have encountered in analysing complex data, most of which are addressed in the dissertation.
- Chapter 3 is a systematisation of the main literature results on the inferential properties of the Forward Search in the regression context.

As discussed above, the chapter is complemented by the first Appendix of the dissertation. The appendix provides algorithmic details that are necessary for readers who are willing to reproduce in their own programming environment the results described in the dissertation. Otherwise, such details can be skipped without losing the main objectives of the thesis.

- Chapter 4 presents the benchmark used to assess the actual size and several measures of power for the Forward Search, the Least Trimmed Squares and the Least Median of Squares of Peter Rousseeuw and for a non robust backward method based on standard regression diagnostics.
- Chapter 5 proposes a new Forward Search based approach for the identification of mixtures of linear components and outliers.
- Chapter 6 presents new interactive tools which implement the idea of connecting dynamically different robust plot as well as traditional plots and data tables, for the purpose of retrieving information which would be difficult or impossible to infer otherwise.

The Chapter is complemented by the second Appendix of the dissertation, which provides more practical information on the use and main features of the tools implemented.

- Conclusions and discussions of future research topics are contained, as usual, in the last Chapter.

Challenges in the Analysis of Complex Data

2.1 Introduction

During the last years of research activity I encountered a number of problems and challenges in the analyses of complex data. Some of them have been addressed extensively in this thesis (Section 2.3), others are problems of more practical or operational nature that I encountered while working at the Joint Research Centre of the European Commission and others are problems that I plan to address in the next future (Section 2.4).

With the term “complex” I do not necessarily refer to high dimensional data but to situations which are difficult to analyse for various reasons such as the lack of linearity in the data and the presence of atypical observations. The next Section presents an example of such complex data.

2.2 An example of complex corrupted data

We start with a complex dataset, which is shown in Figure 2.1, to which I dedicated special attention during my working experience at the Joint Research Centre (JRC). We will often refer to it throughout the dissertation.

The data represent the quantity (x axis) and the value (y axis) of 4719 import declarations of a specific technological product into a Member State of the European Union. This example is one of the many datasets provided by the “Office Européen de Lutte Anti-Fraude” or by its partners in the Member States to the statisticians of the Ispra site of the Joint Research Centre of the European Commission with the purpose of finding atypical

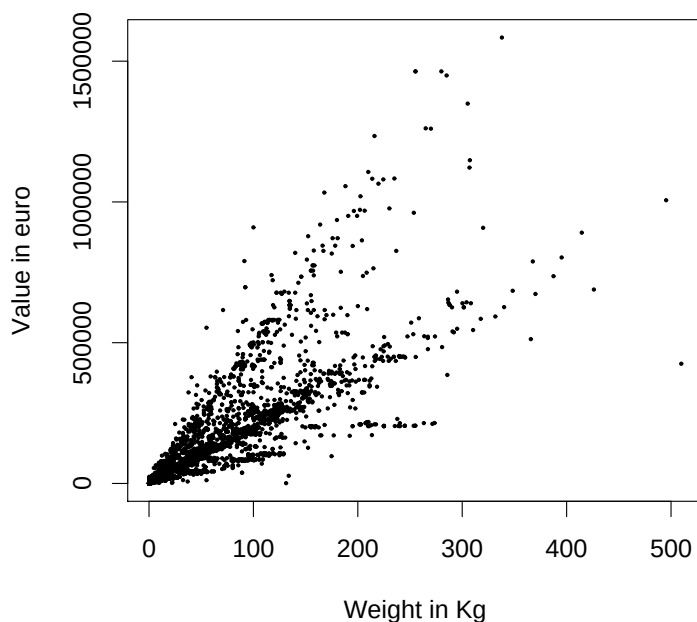


Figure 2.1: An example of trade dataset: quantities (in kilograms) and values (in thousands of Euro) of 4719 import transactions of a technological product into a Member State of the European Union, in a period of one year.

transactions that might correspond to potential frauds (e.g. import duties evasion) or money laundering cases. Depending on the context there can be few thousands to some hundred thousands of such datasets to scan in the hope of finding, among the many signals detected, potential fraud cases.

At first sight the observations in Figure 2.1 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. However, traditional models which assume one single regression population will fail in revealing the real structure.

A further complication consists in the fact that, as with the majority of similar trade datasets, the 4719 observations are quite concentrated towards the origin of the coordinate axes and therefore the three informal groups in the figure overlap considerably in this highly dense area. The degree of overlapping depends on the economic features of the markets where the flows originate.

Given that the observations in Figure 2.1 have been collected over dif-

ferent periods of time, there can be an effect due to serial correlation of the transactions. In fact, as will be presented in chapter 5, different methods of analysis can produce very different mixtures and clusters.

In addition, this mixture of linear models can be heavily contaminated by observations that do not follow the general pattern. Outliers may be isolated, originating for example 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 (or hyperplane).

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 with large values of either variable. Of course, in the absence of fraud, large values of both variables occur together.

In the presence of a core population and some isolated or clustered outliers, traditional robust methods [MMY06] can be successfully used to find proper models. However, when there are several populations and different sources of heterogeneity, even traditional robust methods fail to recover the real structure of the data.

We conclude the description of this complex dataset by remarking that, 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. For example, in Figure 2.1 the information about the value and quantity of the transactions can be coupled with additional information from other variables which are present in the dataset. Unfortunately it is not clear if there is some relationship among the different variables and what is the relevance of such indicators for anti-fraud purposes. Therefore, in the dissertation we limit our analyses to the main two numeric variables: the traded quantity and value.

The thesis will address other examples of trade datasets. For all, the trade values will be regressed over the trade quantities, i.e. quantity and value will be represented in the x and y axes respectively. In doing so, the hypothesis that the intercept of the regression line is not equal to zero is in general rejected if tested on homogeneous groups of trade flows. Thus, the slope of the regression line will be naturally interpreted as an estimate for the trade price of that group of flows.

2.3 Key problems addressed in the thesis

This Section describes the main problems and challenges that I have encountered in analysing complex data and that I have tried to face in the dissertation.

The presence of outliers

It is well known that the presence of atypical observations may wrongly influence the output of statistical analyses, especially if it is based on the classical least square approach. Figure 2.2 exemplifies this fact on a trade dataset by showing how the import price estimate of a certain product (i.e. the slope of the regression line) changes as long as outliers are removed from the dataset, from the initial 7.4 to the final 12.3 Euro/Kg.

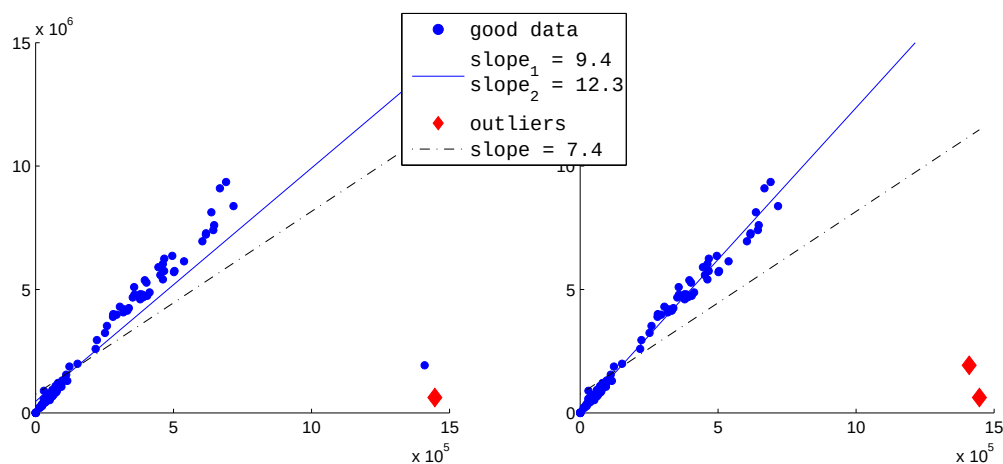


Figure 2.2: Price outliers in a trade dataset for a specific product. The estimate of the import price (i.e. the slope of the regression line) changes each time an outlier is removed. The difference between the initial (dashed grey line) and final (continuous blue line) estimates is considerable and may be relevant for the operational purposes of the data owners.

In general, especially when the number of observations is large, it is likely to find one or more atypical observations that affect the fitted model so strongly to mask other deviating observations, which remain undetected. This is the so called *masking effect*. For example, in the context of international trade data, the fact that the import duty rates are usually proportional to the value of the imported product, may induce fraudsters to considerable under-declaration of the value, producing groups of low price

flows located in the same area. On the other hand, export prices that are too high may also reflect money laundering or an attempt to benefit from export subsidies. Again, groups of high price flows may occur together in the same area.

Obviously not all abnormal prices are associated with fraudulent transactions. For example, it is possible to find errors in recording, although their frequency and distribution are not known. The outliers of Figure 2.2 are so abnormal in price that may indeed be due to errors in the registration of the value or even of the product code. The first scenario would imply a rectification of the wrong values, while the second would require moving the two observations, that are outlying here, to a different dataset.

It is not the purpose of the statistician to obtain definitive conclusions about fraud, since the analysis requires additional subject matter information and further controls on the suspect transactions. However, the use of appropriate statistical methods can play a crucial role in focusing attention on individual cases among tens of thousands of similar situations, thereby allowing timely activation of the investigations necessary to the detection of fraud.

Therefore, in this thesis the outliers are not seen as bad observations that estimation procedures must avoid, since they may themselves contain valuable information and can be the ultimate target of the analysis.

The presence of high leverage points

A well known concept in regression diagnostics is the *leverage* of an observation. Informally, *leverage points* can be seen as observations that are outlying in the explanatory variables (see e.g. [RL87] or [BKW80]).

In the context of international trade data this phenomenon occurs in the presence of transactions with much higher volumes than the majority of the data. Figure 2.3 shows the scatter plot of a trade dataset together with the plot of the scaled residuals obtained by fitting robustly the dataset. Observation 16 in x-space is far from the other observations and is therefore a leverage point. It is also a strong outlier, as we can see from the plot of the standardised residuals, which are defined as

$$r_i = \frac{e_i}{s\sqrt{1-h_i}}, \quad (2.1)$$

where e_i is the least square residual of observation i , s^2 the residual mean square estimator and h_i the diagonal elements of the Hat matrix (see Section 3.5). Therefore, this “bad” leverage point would influence a lot a standard

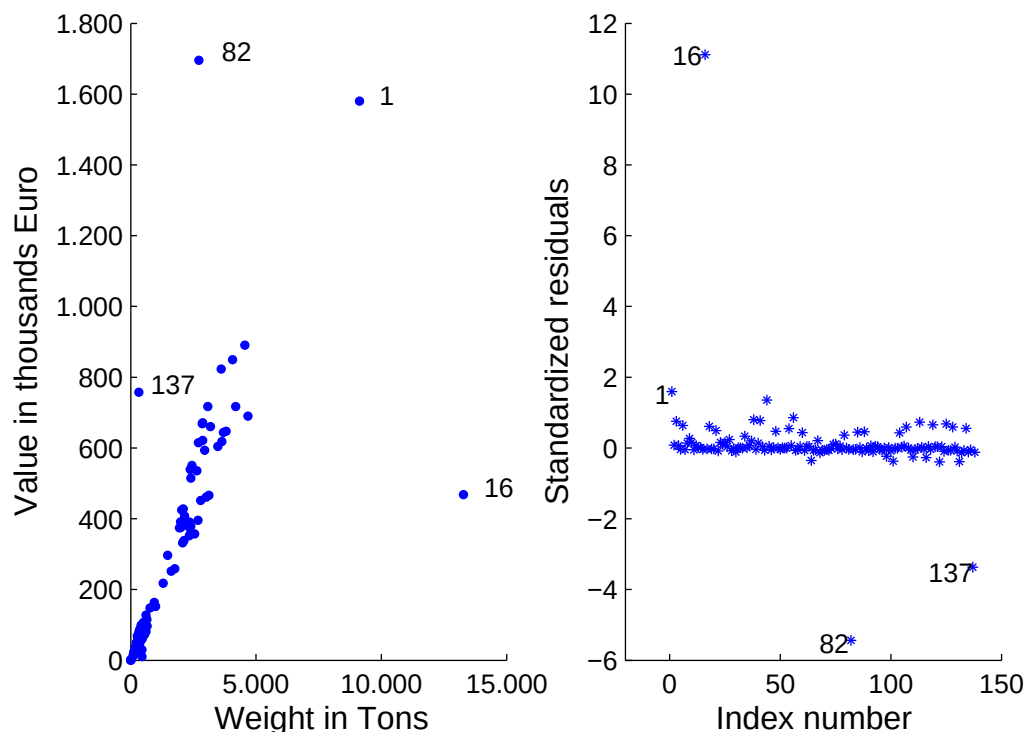


Figure 2.3: A trade dataset with three clear outliers (observations 16, 82 and 137) and two leverage points (observations 1 and 16).

least square fit. Observation 1 is also far in x -space from the other observations, but is rather in line with the other observations in the xy -space. In fact its standardised residual is close to that of the other observations. This is a leverage point which would make the linear correlation coefficient increase but would not influence the least square fit. Finally the observations 82 and 137 are outliers with very low leverage.

As is well known, the presence of outlying leverage points can lead to the identification of good data points as outliers, the so called *swamping* phenomenon.

The presence of sub-populations or hidden groups

The trade data example of Figure 2.1 is one of the many cases where we found more than one population in the data. In the statistical literature there is a big separation between methods which assume one population and those of cluster analysis which force all observations to fall into a certain number of groups, often ignoring outliers. More specifically, traditional

robust methods fail when the data contain two or more populations. These facts motivate the research described in chapter 5, where we propose an extension of the Forward Search for identifying sub-populations in the data and model them as a mixture of linear models.

The presence of high-density areas

We found that in international trade there are often portions of the data space which have variability very different from that of the remaining part of the data. In the literature of robust statistics, the groups of observations which show a high degree of concentration are called *inliers*. Just as outliers have a big effect on the results of statistical analyses, so also do inliers.

In particular, as shown in Figure 2.1, trade data can have a large number of transactions characterized by low quantities. These small amounts often arise from common data recording practices. For example, the import declaration of a good may be split between different subcategories and to each the same nominal small amount (quantity) is assigned. In other words, in highly complex trade data it is not difficult to encounter observations affected by measurement errors concentrated in particular areas.

Typically an area of high concentration causes the rejection of the normality hypothesis, central to much non-robust inference. Besides, in the presence of multiple groups it is not clear whether it is necessary to transform the data robustly, to trim this highly dense area or to treat these observations in some different way.

The extension of the Forward Search for the identification of mixture components that will be proposed in Chapter 5 has been conceived to be robust to the presence of high dense areas.

Perfect fit

This phenomenon happens when the observations in the dataset follow a straight line (or hyperplane) with a very small random fluctuations around it. The small variance of the errors can lead to deletion residuals (see equation 3.6) which are large and so to declare as anomalous transactions which in absolute terms lie very close to the regression line (or hyperplane). An example of this situation is represented by another trade dataset, that is plotted in Figure 2.4: the points in the scatterplot appear almost perfectly aligned, but the Studentised residuals for some units are high and lie even outside the 99.9% envelopes.

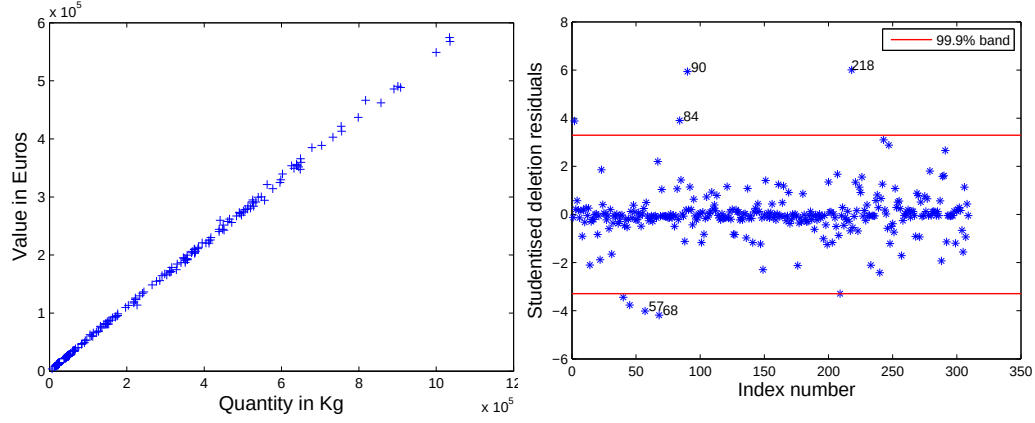


Figure 2.4: On the left a trade dataset with “perfect fit”. On the right the Studentised deletion residuals plot for an ordinary least square fit. The alignment is almost perfect ($R^2 = 0.99946$), but many observations still have large deletion residuals.

Calibrate for simultaneity and keeping false alarms under control

The repeated application of statistical tests makes it necessary to use simultaneous tests rather than individual tests. When no outlier is present, in *simultaneous tests* we are willing to tolerate false positives in a fraction α of the datasets, while in *individual tests* we are willing to tolerate a fraction α of false positives in each dataset. It is known that the application of traditional robust methods often produces error rates of the first kind (individual and simultaneous) much higher than nominal ([BG99], [HR05] and [CRA09]). It is therefore necessary to set up a procedure which adequately controls the size of the test ([HT87] and [Mil66]).

This is particularly important in the context of international trade data, where the number of signals given to inspectors must be limited because a high number of false alarms can lead, as they do, to unavoidable disappointment and disaffection with statistical methods and to a reduction in the number of successful prosecutions.

The problem of simultaneity in the Forward Search is discussed mainly in Section 3.6.

The need for interactive dynamic plots

In analysing trade data we have been constantly driven by the needs of the end users (the anti-fraud inspectors). A clear need is to dispose of graphics for easy validation of potential signals. The classical static plots appear to be rather limited in that sense and the following minimal scenario is envisaged.

By the use of simple mouse clicks, outliers should be highlighted automatically in all plot that might be of interest to the end-user and/or the statistician. Beside, information should be made available about which transactions are related to particular products, companies and nations of origin and destination. Particular care should be given to effective communication of the statistical information associated with an observation, an outlier, a cluster or any other pattern of interest to a non statistical audience, in the spirit of [Tuf83] and [Spe01].

In short, the challenge was to provide flexible and interactive plots incorporated in user friendly graphical interfaces, so that to extend the classical tools of exploratory data analysis. In this spirit, Chapter 6 is dedicated to the presentation of a graphical tool that makes forward plots and data tables communicate among themselves.

The need for 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. The thousands of combinations of transactions for each country of origin and country of destination require an automatic monitoring system to detect outliers which may result in potential fraud cases. At this goal the Forward Search was implemented in MATLAB with my contribution (software available in www.riani.it/MATLAB.htm) and in SAS by the Joint Research Centre. I also implemented both in SAS and Matlab a standard backward procedure for the detection of outlier based on deletion diagnostics tests (see Section 4.3), which is currently in use at the Joint Research Centre and the anti-fraud service of the EU.

2.4 Next challenges

In this Section I present problems that I have encountered but I have not yet treated. The last one is not related to international trade, but I think represents a great application challenge for the Forward Search.

The presence of serial and spatial correlation

This relates to situations where observations are taken over time and space and many multivariate time series have to be jointly analysed. This is a relevant problem in the context of international trade data, because strong signals of potentially fraudulent behaviour, money laundering and tax evasion include the discovery that a cluster of masked outliers is associated with transactions originating from a single country or company, and that the deviations seem to increase with the passing of time. Figure 2.5 illustrates a case where the spatial and time correlations are clear from the scatter plot of the weight and value of the trade flows (left panel) and the time series of the corresponding prices (the values are divided by the weights, right panel). The time series focuses on the flows of an importer which operates in a certain Member State (MS7 in the figure). I have approached this case as a problem of linear mixture estimation in Chapter 5, but it is clear that good results could come from considering time correlation.

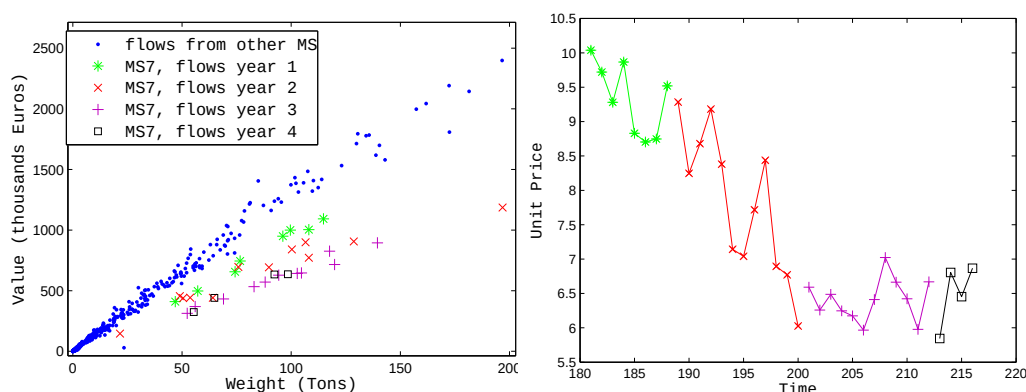


Figure 2.5: An example of spatial and time correlation in a set of import declarations of fish (Fisheries dataset). Left panel: some groups of observations that look abnormal with respect to the majority of the data are associated with declarations by a single country (MS7). Right panel: the deviation from the “fair” declarations clearly increases with the passing of time (the trend of the price is decreasing).

More in general, with many variables the simultaneous treatment of spatial and temporal correlation can be much more difficult to treat. In the presence of serial correlation, contamination in one component may be caused by an outlier in other components. Besides, multivariate outliers often cannot be detected by looking at the component series separately. In

such cases it is more appropriate to look at deviations from the bulk of the data in a multivariate framework.

Timeliness and On-Line Systems

Typically the analysis of the presence of potential outliers is done completely off-line. For example, in the international trade scenario data are made available in batches, they are processed and the results are made available to the end-users only in subsequent periods. The same data may be present in different batches and, thus, subsequent analyses may produce common outliers. In some cases it is more realistic to consider the on-line scenario, 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 new extended dataset. The availability of on-line analysis tools also resolves the need for timeliness, which could arise if the purpose is to detect potentially fraudulent transactions at the time when they are declared and recorded by the Customs officers. In other words, by “timeliness” we mean the capacity to (i) automate the on-line analysis of data at the time they are transferred to the system, (ii) complete such analysis in a time compatible with the given on-line application scenario and (iii) provide the results immediately, in a form that is understandable by the final users.

Searching in the outliers space

In many contexts there may be reasons to treat in a different way two units which have been judged as outliers even if they share similar p -values for outlyingness. In the context of the international trade declarations, for example, not only the p -value but also the raw residual, that is the estimated amount of money which is lost by the potential tax evasion, must be taken into account and the value of the transaction itself must be considered. Ranking criteria conceived to minimize the estimated loss of the potential frauds should be considered in reporting results to anti-fraud authorities. This problem is currently under study in the Joint Research Centre.

A related but different approach could be to analyse the units in the “space of the outliers”. This means to consider the transactions sorted according to a measure of severity (e.g. the p -value) in combination with other relevant variables, e.g. the expected loss, the volume or the value declared by the trader. The analysis of atypical combinations in this outlier space may lead to the detection of relevant patterns, for example outliers with high severity values associated to the same companies, products, coun-

tries of origin or destination. Therefore, in general the problem is to find appropriate graphical representations for a set of outlier signals and their associated measures of risk.

Automatic model selection procedures

For high-dimensional data, 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. At present in the literature there are several procedures which enable the researcher to perform model selection. Recently there have also been robust proposals which extend model selection to the cases where outliers are present (e.g. [SBVA08]). Certainly, what is needed is an extension of model selection procedures to data which are serially or spatially correlated when outliers are likely to be present. This need is clearly not relevant for international trade data, which is characterised by few numeric variables.

The Forward Search

3.1 Introduction

The main reason which makes, in my view, the Forward Search appealing for data analysis is that it can be seen as a new *philosophy of data analysis* and not only as a statistical outlier detection technique among others. I will propose this point of view in Section 3.2. Then in Section 3.3 I will survey the key phases of the research in the Forward Search together with the main literature results in regression and multivariate contexts.

The second part of the chapter focuses on the inferential properties of the Forward Search, which are introduced for regression in Section 3.5. I will present in particular how the Forward Search faces with problems of multiplicity and simultaneous testing. These problems arise also at application level, as in the international trade context outliers are detected among a large number of samples and finding an acceptable trade off between size and power is difficult. Considering that the issue is common to many other application scenarios it is quite surprising that the literature of robust statistic seems to avoid the subject, by focusing on achieving high power only.

In Sections 3.6 and 3.7 (that are complemented by the dissertation's Appendix) I will explain how the Forward Search achieves its default 1% nominal size and how it faces with the multiple testing issue. Then, I will propose a strategy to extend the Forward Search detection of outliers to an arbitrary confidence level. Such strategy is an original contribution of the thesis.

3.2 The data-driven flexible trimming approach

To introduce the rationale of the Forward Search approach, we consider a very basic goal, that is estimation of a parameter of interest μ , for example the mean of the population. The frequentist approach to statistical inference treats μ as an unknown constant. Furthermore it requires the specification of a family of models which are indexed by μ and uses the sample data to answer questions about μ and the model itself. If the data do not contain outliers, strong arguments can be produced for reducing the data to their mean. However, if the family of models is more sophisticated and allows the data to be contaminated from another distribution, a more complex function than the sample mean should be used to reduce the data. One example of such a function is the trimmed mean, other examples are the M-estimator of location and the sample median. To obtain robust estimates of the unknown parameter μ we need to follow one of the following strategies:

1. Use a reduced number of observations in order to exclude outliers.
2. Down-weight each observation according to its deviation from a centre.
3. Optimize an objective function which is more robust than that for classical least-squares.

A recent reference book on robust statistics that addresses such approaches and their many good properties is Maronna et al [MMY06]. The reader can find in the book pointers to many seminal contributions that we omit here. We will rather consider a series of drawbacks in relation to such strategies.

- First of all, in all strategies the simplicity of the sample mean is lost.
- Sometimes results obtained via different robust approaches are completely different and this causes dismay.
- Further disadvantages are the fact that the percentage of observations to be discarded for estimation of μ (strategy 1) needs to be fixed in advance, there is no universally accepted way to down-weight observations (strategy 2), and optimization of complex functions (strategy 3) may raise severe computational problems.

- Another fundamental shortcoming common to all the robust strategies described above is that their extension to complex problems such as those discussed in Chapter 2 (e.g. serial and spatial correlation, presence of multiple populations in the data) is difficult and requires ad hoc strategies for each specific problem. Therefore, no single available robust technique can deal simultaneously with all the complexity features introduced in the dissertation and provide the user with a unified view of the available data.
- Finally, the researcher loses the connection of the effect that each observation exerts on the estimates of the parameters in the proposed model. In the previous example of the estimation of μ , whatever approach is used the researcher loses the information that each observation, outlier or not, has on the final proposed estimate.

The Forward Search approach to data analysis and inference is different. Although, as in robust statistics, it concerns the effect that outliers may have on estimation and on other inferential problems, it consists in a tool that *preserves the interpretative and computational simplicity of the sample mean*, thus keeping its high efficiency when the basic uncontaminated model is true.

Moreover, the Forward Search tries to attack relevant inferential issues in a unified way by basing the inference on carefully chosen subsets of the data. The key difference with respect to the robust strategy 1 is that Forward Search does not choose just one sub-sample, but *fits a sequence of subsets and let the data decide which is best* for the model under study. This approach *preserves robustness* to departures from the underlying null model, because outliers and other observations not fitting this model are not present in the best sub-sample. It also *ensures high efficiency* because only those observations are discarded while all the ‘good’ ones are included.

Furthermore, Forward Search is very flexible and can be tuned to solve apparently different statistical issues through the *definition and monitoring of powerful problem-specific diagnostic quantities*. Such diagnostic investigation of the currently fitted model can be computationally demanding for real life datasets and, thus, was impossible to achieve up to few decades ago in the absence of sufficiently powerful computers. On the other hand, at that time also many traditional robust methods which minimize computationally intensive criteria different from the sum of squares were affected by severe practical limitations.

Both in the use of traditional robust and non-robust statistical methods, researchers end up with a picture of the data. In the estimation of the

mean, for example, non-robust methods produce even too simplistic pictures, by forcing all observations to be equally treated. Traditional robust methods allow differential treatment of the observations, but they use iterative processes whose output sometimes seems to come from a black box. On the contrary, the Forward Search approach involves watching a film of the data rather than a snapshot. The film is about the sequence of subsets built during the progression of the search and about the evolution of the fitted model and any other problem-specific diagnostic quantity whenever a new statistical unit is added to the subset. *By watching the film one can appraise the effect that any statistical unit of interest (e.g. an outlier or a leverage point) exerts on the fitted model once it is introduced into the subset.*

All the above motivate the claim in the introduction of the Chapter, that the Forward Search is not a simple new outlier detection algorithm but is rather a new philosophy of looking at the data.

3.3 Key phases in the Forward Search history

In presenting how the Forward Search has evolved in the years we think reasonable to roughly distinguish three phases in subsequent periods of time.

1992-1995: introduction of the new approach idea

To start with, it is certainly remarkable to note that the idea of forwarding detecting outliers in the data was introduced in engineering before than in statistics, at the beginning of the '80s, with a method known as *Random Sample Consensus* (RANSAC). The application context where RANSAC was born is image analysis, which is characterised by a high presence of errors in the data. We like to see this as a confirmation of the property of flexibility that we claimed about the forward approach in the previous Section. A reference article for RANSAC is [BF81].

In statistics, the basic ideas of the Forward Search originated with the work of Hadi and Simonoff [HS93], Hadi [Had92] and Atkinson [Atk94]. [HS93] and [Had92] proposed methods which face the problem of detecting multiple outliers, respectively in regression and multivariate contexts, by identifying in the data an outlier-free subset and by iteratively adding to this subset the observations which are less outlying according to some statistics. In [Atk94] the term 'forward search' is mentioned for the first

time with reference to both regression and multivariate problems. The author proposes to repeat a number of iterative procedures, similar to the ones described by [HS93] and [Had92], by starting from different random starting subset. Through specific plots, it is shown how the estimates obtained with these various procedures converge to a unique value.

1996-2000: introduction of diagnostic monitoring

The full power of the Forward Search approach became clear only towards the end of the 90's, when the idea of diagnostic monitoring was introduced in Cerioli and Riani [CR99], Riani and Atkinson [RA00] and in the book of Atkinson and Riani [AR00]. Diagnostic monitoring is a crucial ingredient of the Forward Search because it supplies the user with a wealth of clearly understandable graphical information about the structure of complex multivariate or multi-variable data.

In the context of spatial analysis, [CR99] showed how spatial data can be ordered through the Forward Search from those most in agreement with a specified autocorrelation model to those least in agreement with it. It is therefore possible to identify spatial outliers, that is extreme observations with respect to their neighbouring values, and of non-stationary pockets.

In 2000 [RA00] focused on transformation of the response in regression. The authors showed how the Forward Search provides an ordering of the data from those most in agreement with the proposed transformation to those furthest from it. The ordering was then used to understand the contribution of each observation to the estimated transformation.

2001-2009: wide application of diagnostic monitoring and inferential results

Many diagnostic procedures based on the Forward Search have been developed in different research fields, e.g. discriminant analysis and clustering methods, principal component analysis, multidimensional scaling, non-linear models, time series, spatial models, binary data, 3-D objects surveyed and so on. These contributions are described in the book by Atkinson, Riani, and Cerioli [ARC04] and in many scientific papers, among which we mention Riani and Atkinson [RA01], Cerioli and Riani [CR02], Riani [Ria04], Atkinson and Riani [AR07], Solaro and Pagani [SP07], Mavridis and Moustaki [MMss], Crosilla and Visentini [CV07] .

[RA01] is an extension of the multivariate transformations with the Forward Search to linear and quadratic discriminant analysis.

[CR02] extends the Forward Search to the analysis of spatial data through simultaneous autoregressive models and through the development of joint robust estimation of both trend and autocorrelation parameters in spatial linear models.

[Ria04] extends the Forward Search to the analysis of structural time series data. It shows how the Forward Search can detect the main underlying features of the series under study, e.g. masked multiple outliers, level shifts or transitory changes.

[AR07] shows how series of robust Mahalanobis distances based on increasing numbers of observations could be used to find the appropriate number of groups. The paper also shows simulated envelopes of the distribution of the test statistics in forward plots.

[SP07] extends the Forward Search to metric multidimensional scaling. In particular the authors use the Forward Search to monitor the influence of outliers on the extraction of dimensions and to compare multidimensional scaling solutions obtained with different multidimensional scaling techniques.

[MMss] extended the technique to latent factor analysis for binary data.

[CV07] used the Forward Search in the classification and segmentation of 3-D objects surveyed with high density laser scanning techniques.

In this period also many formal inferential results, that provide a sound statistical basis for the diagnostic tools of the Forward Search, have been developed. Important contributions to the statistical literature on this topic include Atkinson and Riani [AR02], Atkinson and Riani [AR06] and, more recently, Riani [Ria09], Proietti and Riani [PR09], Cerioli, Riani and Atkinson [CRA09], Riani, Atkinson, and Cerioli [RAC09].

[AR02] shows how it is possible to measure the importance of observations to the significance of the individual regressors with an added-variable test.

In [AR06] three straightforward methods using truncated samples are described for approximating the pointwise distribution of the test statistic in regression.

In [Ria09] a Forward Search score test is applied in time series in order to quantify the effect of each observation on the choice of the transformation.

In [PR09] the Forward Search is applied to the problem of seasonal adjustment of a non-linear transformation of the original time series.

[CRA09] shows in a multivariate context that the Forward Search provides good control over the size and high power measures of multiple outlier tests for various sample sizes and dimensions. Much worse results are obtained with minimum covariance determinant procedures.

With [RAC09], one of the most recent contribution, the Forward Search has gained full recognition to be considered as a truly new inferential paradigm, where the amount of trimming is not fixed in advance but is chosen from the data.

In the last two years the Forward Search was investigated in the context of anti-fraud control problems requiring accurate analysis of very large collections of complex datasets. The first steps in this direction were made by Perrotta and Torti [PT09].

3.4 Searching outliers backward and forward

Before introducing the matter with a mathematical formalism, it is useful to sketch the steps which distinguish the Forward Search from the traditional idea of monitoring the regression diagnostics backward.

Strategy 3.1 Backward Search of outliers. Iterate steps 2-3 until a certain number of potential outliers are excluded.

1. Start with the whole data set.
2. Compute a measure of outlyingness.
3. Remove the observation which is most extreme according to the measure used.

Strategy 3.2 Forward Search of outliers. Repeat steps 2-4 until all observations are included.

1. Start with an initial outlier free subset of the data;
2. Order all observations according to a measure of closeness to the subset of size m ;

3. Select the least aberrant $m + 1$ observation and refit the model;
4. Save the value of any problem-specific diagnostic quantity of interest.

Typically the measure of outlyingness used by Strategy 3.1 is the deletion residual regression diagnostic. An example of backward strategy will be given in Section 4.3. The next sections show how the problem-specific diagnostics monitored by the forward Strategy 3.2 are used for outlier detection and for other general inferential purposes.

3.5 The inferential approach to the forward detection of outliers

In the regression model

$$y = X\beta + \epsilon$$

y is the $n \times 1$ vector of responses, X is an $n \times p$ full-rank matrix of known constants, with i th row x_i^T , and β is a vector of p unknown parameters. The normal theory assumptions are that the errors ϵ_i are i.i.d. $N(0, \sigma^2)$. Note that this assumption is not critical, as the Forward Search output will show a possible lack of normality with specific patterns in the monitored trajectories that will allow to identify the non homogeneous observations (see for example Chapter 6).

The least squares estimator of β is $\hat{\beta}$. Then the vector of n least squares residuals is

$$e = y - \hat{y} = y - X\hat{\beta} = (I - H)y \quad (3.1)$$

where

$$H = X(X^T X)^{-1} X^T$$

is the ‘hat’ matrix, with diagonal elements h_i and off-diagonal elements h_{ij} . The residual mean square estimator of σ^2 is

$$s^2 = e^T e / (n - p) = \sum_{i=1}^n e_i^2 / (n - p).$$

The forward progression

The Forward Search in regression typically starts by fitting a small subset of m_0 of the n observations in the data. The subset is chosen robustly, using LMS or LTS estimators [Rou84]. For general m , with $m_0 \leq m \leq n$,

let $S_*^{(m)}$ be the optimum subset of size m , for which the matrix of regressors is $X(m^*)$. Least squares on this subset of observations yield parameter estimates $\hat{\beta}(m^*)$ and $s^2(m^*)$, the mean square estimate of σ^2 on $m - p$ degrees of freedom. Residuals can be calculated for all observations including those not in $S_*^{(m)}$. The n resulting least squares residuals are

$$e_i(m^*) = y_i - x_i^T \hat{\beta}(m^*). \quad (3.2)$$

The search moves forward with the augmented subset $S_*^{(m+1)}$ consisting of the observations with the $m + 1$ smallest absolute values of $e_i(m^*)$. Note that, therefore, the estimates of the parameters are based on only those observations giving the central m residuals.

In the forward process we obtain a series of parameter estimates for $p \leq m \leq n$, which progress from very robust (i.e. $\hat{\beta}_{LMS}$ or $\hat{\beta}_{LTS}$) at the beginning of the search to least squares at the end. The Forward Search estimator $\hat{\beta}_{FS}$ is defined as a collection of least squares estimators in each step of the Forward Search, that is,

$$\hat{\beta}_{FS} = (\hat{\beta}(p^*), \dots, \hat{\beta}(n)). \quad (3.3)$$

A remark should be made on the order of entry of the units, which is sometimes misinterpreted. In most moves from step m to $m + 1$ just one new unit joins the subset. However it may also happen that two or more units join $S_*^{(m)}$ as one or more leave. Such an event is quite unusual, only occurring when the search includes one unit that belongs to a cluster of outliers. When this happens, at the next step the remaining outliers in the cluster seem less outlying and so several may be included at once. In this case several other units then have to leave the subset.

Monitoring the search

During the search we can monitor any relevant statistic, typically, quantities indicative of model quality or inadequacy. Figure 3.1, for example, shows the evolution of some of such statistics monitored for the “perfect fit” dataset introduced in Figure 2.4 of Section 2.3.

To start with, consider the evolution of the estimate of the error variance $s_{S(m)}^2$ (top left panel). Because the search orders the observations by the magnitude of their residuals from the fitted subsets, the value of $s_{S(m)}^2$ increases during the search, although not necessarily monotonically. As a consequence, even in the absence of outliers and model inadequacies, the values of the t tests for the parameters in the model decrease during the search and are hard to interpret.

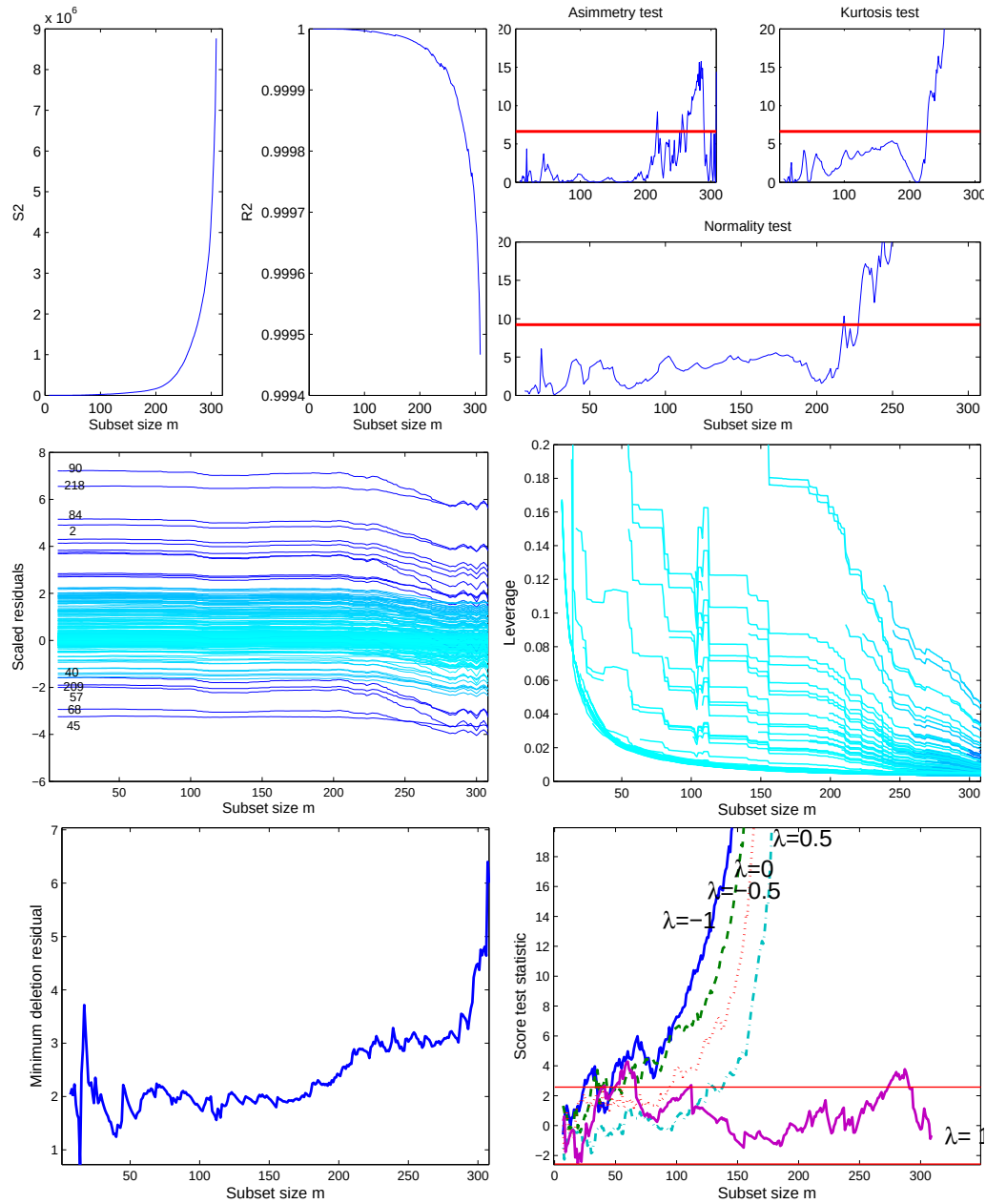


Figure 3.1: Various statistics monitored for the “perfect fit” dataset introduced in Figure 2.4 of Section 2.3. From the top left to bottom right: $s^2_{S(m)}$, $R^2_{S(m)}$, asymmetry, kurtosis and normality tests, scaled residuals, leverage values, minimum deletion residual and score tests for transformation (fan-plot using $\lambda = -1, -0.5, 0, 0.5, 1$). From the plots it is clear that the perfect fit dataset is not homogeneous.

In [AR02] the method of added variables is used to provide plots of t tests which are orthogonal to the search. In order to judge the importance of a particular variable (say w), the added variable plot shows the residuals of y against all predictors except w , versus the residuals of w on all other predictors.

In the context of response transformation, when the null hypothesis is on the Box-Cox transformation parameter [BC64], $\lambda = \lambda_0$, the added t tests based on constructed variables are known in the statistical literature as “score test for transformation”. A powerful robust tool to understand the percentage of observations which are in accordance with the different values of the transformation parameters is the forward plot of the score test statistic for transformation of the set of constructed variables for different values λ_0 , using a separate search for each λ_0 . These trajectories of the score tests can be combined in a single picture named the “fan plot” [AR00], displayed in the bottom right panel of Figure 3.1.

One of the most used plots monitors all residuals at each step of the search (middle left panel of Figure 3.1). Large values of the residuals among cases not in the subset indicate the presence of outliers, as do sudden changes in the value of the residual sum of squares. Because of the strong dependence of $s_{S(m)}^2$ on m , all residuals are standardized by the final root mean square estimate s^2 , so that their plot is not affected by a big change of scale from the start to the end of the search. One of the drawbacks of this plot is that when n is large there are too many lines and therefore this plot tends to become confused and misleading. This is why the residuals are represented with a colour with intensity proportional to the residual size, so that the smaller and therefore less interesting residuals are faint and almost put in the foreground.

A similar plot monitors a measure of the leverage of all observations at each step of the Forward Search (middle right panel of Figure 3.1). From the plot it is easy to see that observations with high leverage tend to enter in the last steps of the search. Also here the trajectories are coloured with an intensity proportional to the magnitude of the corresponding residuals.

Perhaps the most important plot monitors the minimum deletion residual among observations not in the subset at each step of the search (bottom left panel of Figure 3.1). Since the observations with high residuals enter last in the subset, this trajectory should have an ascending behaviour. In presence of outliers the minimum deletion residual curve in the last steps of the search will increase more drastically and assume high values.

Finally in the top right panels of Figure 3.1 we have reported the forward plots of some classical statistics: the R^2 , the asymmetry test, the Kurtosis test, and the Normality test. The first one indicates that, by including into

the subset the last observations, the linear correlation drastically decreases. The other three plots confirm that the last group of observations are not homogeneous.

So far the sequences of the quantities monitored are graphically represented as *forward plots*. Later, in Chapter 6, we will also present a set of interactive visualization tools which simplify their inspection and the extraction of relevant and often unexpected information from them. In other words, we monitor the search and use the forward plots for exploratory data analysis purposes.

However, there is another way to use the sequences of the quantities monitored. The rest of this Chapter will show how to use them for inferential purposes, i.e. to detect potential outliers. In this case the objective is to draw conclusions automatically, without the manual intervention of the analyst.

Testing for outliers

To test for outliers the deletion residuals are calculated for the $n - m$ observations not in $S_*^{(m)}$. Such residuals are

$$r_i^*(m^*) = \frac{y_i - x_i^T \hat{\beta}(m^*)}{\sqrt{s^2(m^*)\{1 + h_i(m^*)\}}} = \frac{e_i(m^*)}{\sqrt{s^2(m^*)\{1 + h_i(m^*)\}}}, \quad (3.4)$$

where the leverage of each observation

$$h_i(m^*) = x_i^T \{X(m^*)^T X(m^*)\}^{-1} x_i$$

depends on $S_*^{(m)}$. Let the observation nearest to those constituting $S_*^{(m)}$ be $i_{\min}$ where

$$i_{\min} = \arg \min |r_i^*(m^*)| \quad \text{for } i \notin S_*^{(m)},$$

the observation with the minimum absolute deletion residual among those not in $S_*^{(m)}$. If observation $i_{\min}$ is an outlier, so will be all other observations not in $S_*^{(m)}$.

To test whether observation $i_{\min}$ is an outlier we use the absolute value of the minimum deletion residual

$$r_{i_{\min}}^*(m^*) = \frac{e_{i_{\min}}(m^*)}{\sqrt{s^2(m^*)\{1 + h_{i_{\min}}(m^*)\}}} \quad (3.5)$$

which is very sensitive to the presence of outliers. In fact its denominator decreases when it is calculated with reference to an outlier and therefore the statistic (3.5) assumes high values. Note that the formula (3.5) is the

usual t test for the agreement of a new data point with a previous set of observations.

Unfortunately the statistical distribution of the minimum deletion residual is not known in the literature. However Atkinson and Riani [AR06] have proposed to simulate the envelopes for the minimum deletion residual. If the dataset to analyse has n observations and v explanatory variables, the envelopes are found by running the Forward Search on many samples of size n simulated from a standard v -variate normal distribution. Then the Forward Search is run on each simulated sample and the value of $r_{min}(m)$ is monitored. Then, for each value of m the empirical distribution of $r_{min}(m)$ under the hypothesis of normality is given. The envelopes are given by the appropriate quantiles of this distribution. For example, the 99% envelope is that value which is the 1% point of the empirical distribution. In a particular step of the search, with 10000 simulations, this is the 100th largest value, so that 99 of the simulated values are greater than it. These quantiles are computed for each value of $m \leq n$ that is of interest. Of course, the accuracy of the envelopes depends on the number of simulated samples on which we run the search.

As seen, the production of the simulation envelopes implies to run the Forward Search 10000 times or more. In principle this very demanding computational process could be done only once, for a reasonable range of sample sizes of practical interest. Nonetheless to rely on a huge dataset of quantile values is not very practical. For this reason, in [AR06] and [RA07] three approximation of the simulation envelopes have been introduced: one based on truncated samples, one derived using ordered observations and one based on ordering statistics. The pseudo code for the approximated envelopes based on truncated samples is given in the Appendix of the dissertation.

Using the envelopes

Once envelopes for a given monitored quantity are available, it is possible to carry out inference. Here we elaborate on how to detect outliers using the minimum deletion residuals envelopes. The envelopes are used with a two stage process. In the first stage a search is run on the data and the progression of the minimum deletion residual against the envelopes is monitored until a potential signal is obtained. This indicates that an observation, and therefore also the succeeding ones, may be outliers. For example, Figure 3.2 shows the forward plot of the minimum deletion residuals for the Fisheries trade dataset of Figure 2.5. A clear signal is found at step 266, where the trajectory of the minimum deletion residual start exceeding the upper

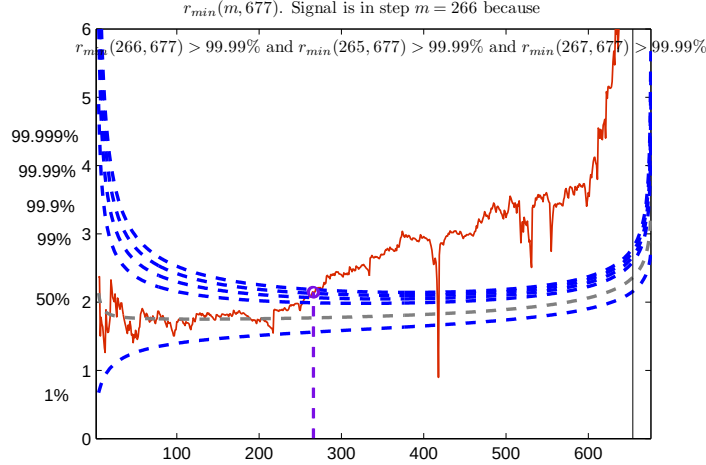


Figure 3.2: Fisheries data (677 observations, plot in Figure 2.5): Forward plot of minimum deletion residuals with 1, 50, 99, 99.9, 99.99 and 99.999% envelopes. There is a signal at step 266.

envelope.

However this is not sufficient to say that the group of 266 observations is homogeneous and all subsequent observations are outliers, because at step 266 the minimum deletion residual trajectory should be actually compared with envelopes based on a 266-sample size. Therefore, in the second stage the precise (outlier free) subset of homogeneous observations is identified by superimposing the new envelopes based on 266-sample size. The upper left-hand panel of Figure 3.3 shows the envelopes for $n = 267$, together with the deletion residuals up to this sample size. With these new, more curved, envelopes it is clear that the group is homogeneous up to this size. The same is true for the lower left-hand panel for $n = 284$. In the upper right-hand panel with $n = 304$, the observed values are approaching the 99% envelope. For $n = 321$ the 99.9% envelope is crossed, so there is evidence of non-homogeneity. When n is one less, namely 320, there is no exceedance and a group of 320 homogeneous observations is finally identified.

In synthesis, the standard strategy to detect outliers using the forward plots of the minimum deletion residuals and their envelopes is the following:

Strategy 3.3 Forward Search strategy to detect outliers at 1% confidence level.

1. Build a sequence of subsets of increasing size using the algorithm 3.2 (Section 3.4).

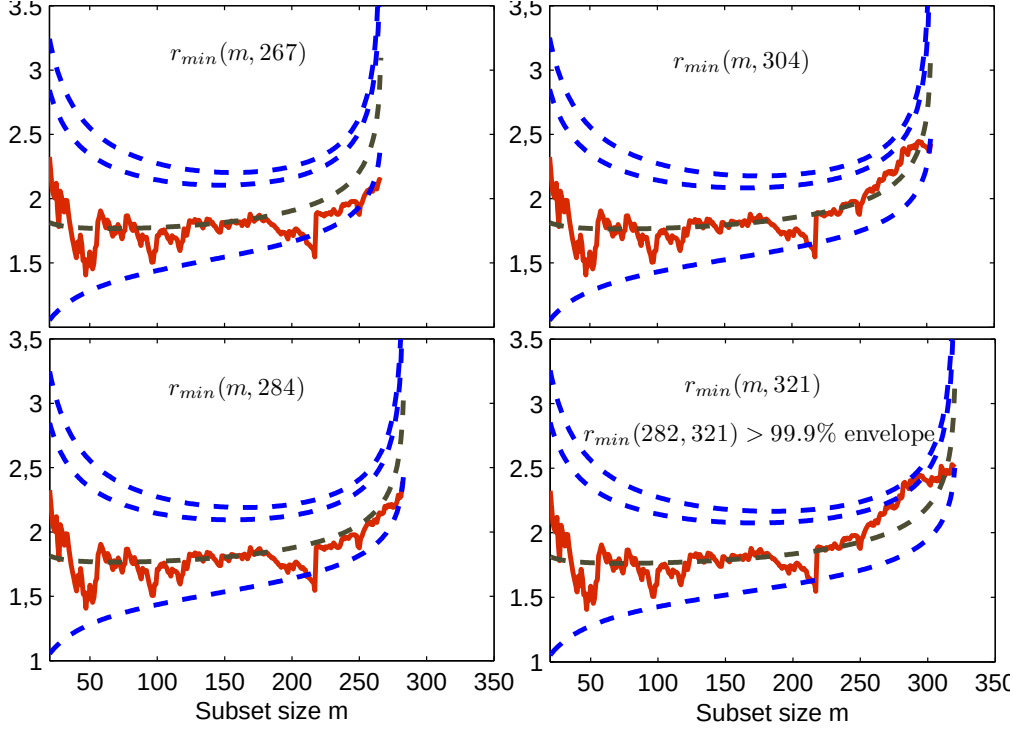


Figure 3.3: Fisheries data (Figure 2.5, 677 observations). The forward plot of minimum deletion residuals with 1, 50, 99 and 99.9% envelopes for various sample sizes. A group of 320 homogeneous observations is identified.

2. Monitor the minimum deletion residual (equation 3.5) at each step of the search.
3. Compare the minimum deletion residual sequence against the envelopes of its distribution. A step $m^\dagger$ where the minimum deletion residual exceeds the envelopes is a *potential* signal of non homogeneity.
4. Identify the precise subset of homogeneous (outlier free) observations by superimposing the envelopes from step $m > m^\dagger - 1$.
5. Observations outside the homogeneous subset are outliers.

The precise definition of a potential signal is given in the Section B of the Appendix, while Section C describes criteria to validate a potential signal into a true signal. Only if a signal is validated we start the (computational intensive) superimposition of the envelopes, which is detailed in Section D of the Appendix.

A signal is therefore indicating that the observation $m^\dagger$, and therefore also the succeeding observations, may be outliers. We want to stress here that more in general the signal may contain an indication of non homogeneous data (e.g. clusters) and violations of normality assumptions. This is the case of the just discussed Fisheries dataset, where the homogeneous subset identified is formed by less than half observations of the dataset and we are clearly in presence of multiple populations. We will use this fact in Chapter 5, where we generalise the use of the algorithm 3.3 to the estimation of linear mixtures.

3.6 On the nominal size of the Forward Search

The inferential approach based on envelopes is a totally automatic process, conceived to achieve a test size of 1%. The detailed description of this process, i.e. which patterns during the search are flagged as *potential* signals and how the potential signals are validated so that to guarantee the nominal 1% size, can be found in the Appendix of this dissertation, with a level of detail that for regression was never addressed so far in the literature. A confirmation of the excellent actual statistical performance of this procedure, in terms of empirical size and power, is discussed in Chapter 4.

Clearly, one may want to detect outliers with an arbitrary test size α , smaller or larger than 1%. This generalisation would require the distribution of the minimum deletion residual 3.5 and the related quantiles over the search. This involves the study of a stochastic process (an empirical process of residuals) which to our knowledge has not been studied yet.

We are only aware of a similar recent work in the multivariate context by Conti and Farcomeni [CF09], who proved that the sequence of the stochastic processes of the minimum Mahalanobis distance among observations not in the subset converges in law, as n tends to infinity, to a Gaussian process. Being a Gaussian process, the distribution of the minimum Mahalanobis distance throughout the search has therefore very well known properties and behaviours [VDV98]. The authors used techniques similar to [BDJ93] and [GEG05] and thanks to the parallel with such techniques they showed the asymptotic equivalence of the Minimum Covariance Determinant estimator and certain kinds of estimators obtained with the Forward Search.

Clearly a process of (minimum deletion) residuals may have totally different asymptotic properties and therefore similar results in the regression context may be hard to derive. Nonetheless, in absence of such desirable

asymptotic results, the limitation of the 1% test significance can be circumvented. We propose for this purpose the following strategy based on Forward Search robust and efficient model estimates. Such strategy is an original contribution of the thesis.

Strategy 3.4 Outlier detection strategy for arbitrary significance α (Forward Search re-weighted).

1. Use the standard Forward Search algorithm 3.3 to identify a homogeneous subset $S^{(m^*)}$ that excludes the most outlying observations at 1% confidence level.
2. Fit an ordinary least square regression line on $S^{(m^*)}$.
3. For each observation in the dataset compute the Studentized deletion residual, which is t -distributed with $n - p - 1$ degrees of freedom for observations inside $S^{(m^*)}$ and with $n - p$ degrees of freedom for observations outside $S^{(m^*)}$ [BKW80].
4. Test for outlyingness observations inside $S^{(m^*)}$ by comparing the Studentised deletion residuals with the critical value

$$t_{(1-\frac{\alpha}{2n}, n-p-1)}.$$

Test for outlyingness observations outside $S^{(m^*)}$ by comparing the Studentised deletion residuals with the critical value

$$t_{(1-\frac{\alpha}{2n}, n-p)}.$$

Now, it is important to remark that the Studentized deletion residual is computed differently for observations inside and outside $S^{(m^*)}$, the homogeneous subset of size m^* on which the fit is based. Let $X(m^*)$ be the matrix of regressors associated with $S^{(m^*)}$ and $\hat{\beta}(m^*)$ the parameter estimates of the fit on $S^{(m^*)}$. Let also $s^2(m^*)$ be the mean square estimate of σ^2 on $m^* - p$ degrees of freedom. Then:

- For observations in $S^{(m^*)}$ the Studentized deletion residual is:

$$r_i^* = \frac{y_i - x_i^T \hat{\beta}(m^*)}{\sqrt{s^2(m^*) \{1 - h_i\}}} \quad (3.6)$$

where the leverage of each observation

$$h_i = x_i^T \{X(m^*)^T X(m^*)\}^{-1} x_i$$

the mean square error estimated without the current observation

$$s^2(i) = \frac{(m^* - p)}{(m^* - p - 1)} s_{m^*}^2 - \frac{e_i^2}{1 - h_i}$$

and the mean square error

$$s_{m^*}^2 = e^T e / (m^* - p) = \sum_{i=1}^{m^*} e_i^2 / (m^* - p)$$

depend all on $S_*^{(m)}$. This means that the residuals e_i in the two previous formulas refer to observations inside the subset $S_*^{(m)}$ and therefore only such residuals contribute to the mean square error $s_{m^*}^2$ and $s^2(i)$ ¹.

- For observations outside $S^{(m^*)}$ the Studentized deletion residual becomes:

$$r_i^* = \frac{y_i - x_i^T \hat{\beta}(m^*)}{\sqrt{s^2 \{1 + d_i\}}} \quad (3.7)$$

where

$$d_i = x_i^T \{X^T X\}^{-1} x_i$$

and

$$s^2 = e^T e / (n - p) = \sum_{i=1}^n e_i^2 / (n - p).$$

now depend on the full dataset $S^{(n)}$. Note that d_i can take values greater than 1 and has not to be identified with the leverage h_i , although its definition and role are very similar to it.

We conclude with three remarks.

- The strategy repeatedly tests the null hypothesis of no outlyingness

$$H_0 : r_i^* \sim t \quad \text{or equivalently} \quad H_0 : e_i \sim N(0, \sigma^2) \quad (3.8)$$

at a Bonferroni significance level α/n , so that to get at a predefined overall empirical size α . In other words, the strategy has a multiplicity problem of size n .

¹Note that in the two last equations, the residual e_i should have been rewritten using the notation $e_i(m^*)$. Here we have avoided this notation because it could have generated confusion: $e_i(m^*)$ could have been interpreted as a residual computed without the m^* -th observation, as it is the case for $s^2(i)$

- When the desired overall test significance α is less than 1%, i.e. we want to be more conservative than the standard Forward Search in rejecting the null hypothesis, the step (4) de facto extends the homogeneous subset found by the Forward Search in step (1). In other words, the robust Forward Search estimate becomes even more efficient. With an abuse of terminology, inspired by a similar trick used for a variant of the LMS and LTS estimators (see Section 4.2), we will refer to this strategy as the “reweighted” version of the Forward Search. Of course, the properties of this new estimator have to be studied and will be subject of future work.
- On the contrary, the homogeneous subset found in step (1) will lose some of its observations if with step (4) we re-weight the standard Forward Search to overall test significance values α bigger than 1%. Also in this case the properties of the estimator have not yet been studied, but we argue they should bring to a sort of “early stopping” of the Forward Search.

3.7 On the simultaneity of the Forward Search tests

We have seen that the algorithm 3.4 involves a test multiplicity issue that advocates for Bonferroni corrections. We also have a problem of multiplicity with the standard Forward Search strategy 3.3, when at each step we test the minimum Studentised deletion residual (among observations not in the subset at that step) against its forward envelopes. In practice each test consists in comparing the minimum deletion residual with an appropriate quantile (e.g. the 99%) of the forward envelopes. However the multiplicity problem behind this process deserves more attention than one might intuitively imagine at first sight.

By construction, being the envelopes based on the idea of bootstrapping the Forward Search by simulation, the tests are point-wise and do not require Bonferroni corrections. In fact Riani et al. [AR06] have shown that the Bonferroni bounds are unrelated to the envelopes and, thus, to the (unknown) distribution of the minimum deletion residual (see Figure 3.4 for $n = 100$ and $p = 3$).

Nonetheless, each of these tests implicitly requires that the $n - m$ Studentised deletion residuals for the observations outside the subset are ordered, which in turn requires that also the m observations in the subset are ordered. Therefore at each step m we have a problem of multiplicity n (and

not m as one might be tempted to claim) because each sorting of the set of deletion residual values necessarily involves all the values.

In more formal terms, at step m the test statistic is the $(m+1)$ th ordered value of the Studentised deletion residuals. Its distribution involves the joint distribution of all the deletion residuals and the null hypothesis becomes more complicated than the (3.8):

$$H_0 : \{e_1 \sim N(0, \sigma^2)\} \cap \{e_2 \sim N(0, \sigma^2)\} \cap \dots \cap \{e_n \sim N(0, \sigma^2)\} \quad (3.9)$$

The size α of this simultaneous test, which is the same at each step of the search, tells us what is the probability that at least one of the individual hypotheses in the intersection hypothesis H_0 is rejected. The practice of superimposing the envelopes is to identify precisely which of the n individual hypotheses should be rejected and at which steps of the search.

In conclusion, the bands are *point-wise confidence intervals* for the minimum deletion residual and do not need to be corrected to account for the evident multiplicity problem. Nonetheless, they are used to make *simultaneous inference* about the significance of the minimum deletion residual values throughout the search. Independently from the step m , the simultaneity is of order n .

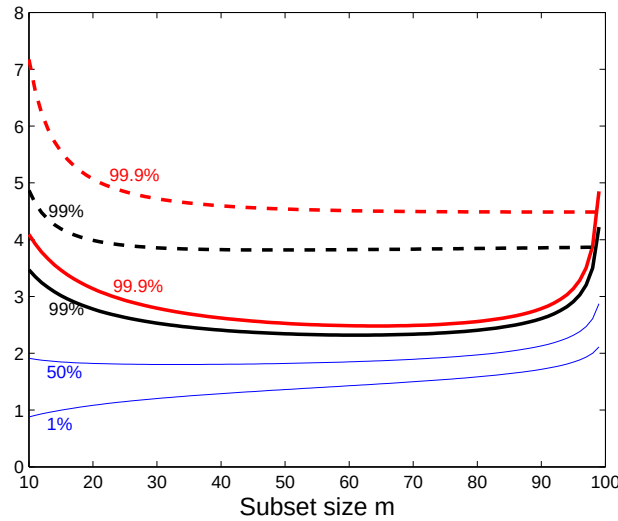


Figure 3.4: The Bonferroni bounds (dotted lines) are unrelated to the envelopes of the minimum deletion residual (continuous lines). They should coincide only at the end of the search, when we test the last observation left outside the subset. The small discrepancy here is due to the use of approximations of the correct simulated envelopes. This example is for $n = 100$ and $p = 3$.

Statistical Performances of the Forward Search in Regression

4.1 Introduction

In this Chapter I describe an extensive benchmark made for estimating the empirical size and power of the Forward Search in the regression context. The motivations for this benchmark are mainly two.

First, the actual size and power of the Forward Search in the multivariate context have been deeply studied by [RAC09], but no equivalent extensive evaluation exists for regression. In other words, the robust statistic literature has not assessed so far the Forward Search against the state of the art of robust regression methods.

Second, in the context of international trade analysis and anti-fraud in particular it is crucial to use statistical techniques that combine high power for outlier detection with reasonable good size. The reason is to find the best possible trade-off between two desiderata: (i) keep the number of false signals as small as possible, for not generating an untreatable number of cases to inspect by the investigators; (ii) detect more truly anomalous transactions as possible so that to minimise the risk of omitting potential frauds.

This strong requirement is not common to many other application contexts and, in fact, in general robust statistics gives much more emphasis on the power of a detection method than on its size. The need of balancing size and power of a statistical methods is not a recent subject and leads to interesting measures such the False Discovery Rate of [BH95].

The Chapter is structured as follows. A brief review of the most popular robust regression techniques in the literature is provided in Section 4.2. In

Section 4.3 I describe a traditional backward iterative algorithm which is concretely in use on anti-fraud problems at the Joint Research Centre of the European Commission, where I worked. Although it is well known that in general backward methods are affected by masking and swamping issues, in practice the method seemed to produce satisfactory results on the bivariate regression trade data analysed in the Joint Research Centre. This is why I decided to study more formally its statistical performances. Section 4.4 presents the benchmark experiment for estimating the empirical size and in sections 4.5 and 4.6 the corresponding results for the Forward Search, LMS, LTS and the backward method for large and small data samples. The corresponding benchmark experiments and results for the power are reported in Sections 4.7, 4.8 and 4.9. In Section 4.11 I reported the CPU time spent for executing the Forward Search and the other robust techniques assessed, so that to have a rough idea of their computational complexity and evaluate their applicability on real world problems. The Chapter concludes in Section 4.12 with some remarks and comments.

4.2 Reference robust regression techniques

The most popular outlier techniques in robust statistics are the Least Median of Squares (LMS) and Least Trimmed Squares (LTS) that were introduced respectively by Frank Hampel in 1974 [Ham74] and by Peter Rousseeuw in 1984 [Rou84].

LMS and LTS

The LMS and the LTS algorithms are articulated in the following steps:

1. Choose a number of random subsets of the data;
2. For each subset fit a least square regression line or hyperplane;
3. Compute the squared residuals $e_i^2, i = 1, \dots, n$ of all n observations in the dataset from the line or hyperplane fitted in step 2;
4. Keep the subset which minimises the following quantity:
 - a) **For LMS:** the median of the squared residuals e_i^2 :

$$\min_{\hat{\theta}_{LMS}} \text{med}_i e_i^2; \quad (4.1)$$

b) **For LTS:** the sum of the $(1 - \alpha) * n$ smallest squared residuals:

$$\min_{\hat{\theta}_{LTS}} \sum_{i=1}^k (e^2)_{i:n} \quad (4.2)$$

where $(e^2)_{1:n} \leq (e^2)_{2:n} \leq \dots \leq (e^2)_{n:n}$ are the ordered squared residuals;

5. Build the classical least square regression line or hyperplane on the subset selected in step 4.2 and compute the residuals of all the observations from this line or hyperplane;
6. Estimate the scale based on the squared residuals of the observations in the subset selected in step 4.2;
7. Compute the standardized residuals (see Equation 2.1) using the estimate of the scale obtained in step 6;
8. Declare as outliers the observations which have a standardized residual greater than a certain cut-off based on the t -distribution.

The LMS and LTS algorithm differ mainly in step where the first compares the median of the squared residuals, the second the sum of the smallest $(1 - \alpha) * n$ residuals.

The LTS estimator is considered by the same authors better than LMS. In fact its objective function (4.2) is smoother than LMS one (4.1) and therefore LTS is less sensitive to local effects than LMS. Also LTS statistical efficiency is better, because the LTS estimator has a higher convergence rate than LMS. Therefore a new LTS algorithm, the *fast LTS*, which is faster also than LMS, has been introduced.

Fast Least Trimmed Squares

In the following lines we summarise the fast LTS algorithm, a detailed description of which can be found in [RVD06].

1. Fit a least square regression line or hyperplane on a h -subset H_{old} of the data;
2. Compute the residuals $e_i, i = 1 \dots n$ of all n observations in the dataset from the line or hyperplane fitted in step 1;
3. Sort the absolute values of these n residuals;

4. Build a subset H_{new} with the observations associated with the h lowest residuals e_i ;
5. Fit a least square regression line or hyperplane on the subset H_{new} , obtained in step 4.

Repeat these steps until $H_{new} = H_{old}$. Then, in order to identify the outliers, apply to this dataset the last steps 5, 6, 7 and 8 of the standard LTS procedure.

The idea on which the fast LTS is based is that starting from any approximation to the LTS regression coefficients, i.e. the least squares estimator of step 1, it is possible to compute another approximation yielding an even lower objective function, i.e. the least squares estimator of step 5.

Rewighted versions of Least Median of Squares and Least Trimmed Squares

To increase the efficiency of the LTS and LMS, the reweighted versions of these estimators have been introduced by [RL87]. The reweighting process consists in adding to the LMS and LTS standard algorithms, just described, the following further steps:

1. Find an estimate of scale of the residuals excluding observations which were defined outliers with the standard LMS and LTS algorithms, described in the previous Subsections;
2. Compute the standardized residuals (see Equation 2.1) using the scale obtained in step 1;
3. Declare as outliers the observations which have a standardized residual greater than a certain cut-off based on the Normal distribution;

4.3 A backward iterative outlier method

This Section presents a traditional backward iterative method to detect outliers using regression diagnostics. The method follows the strategy adopted by most backward approaches, introduced in Section 3.5. The basic idea is that if an observation is substantially different from all other observations and its removal makes a large difference in the regression results (the observation is influential) we remove it and refit the regression model. The main problem of this approach is that it is not obvious to decide how many and

which observations should be removed and, in case of multiple deletion, to check all possible combinations is computationally infeasible.

We have tested different deletion schemes and adopted one in the spirit of [Mar85]: we identify the potential outliers as the observations for which the Studentised deletion residual exceeds a critical value, and remove the one which influences more the model estimate, i.e. the one with highest Cook distance. The regression model changes as the method proceeds and at some point converges to a stable estimate.

Note that the observations removed are only *potential* outliers, because some of them may be the result of regression diagnostics distorted at some backward iteration step by the presence of other outliers: the so called *swamping* problem. Therefore, at the end of the backward iteration each observation in the dataset is tested for agreement to the stable model estimate: like in the re-weighted version of the Forward Search, strategy 3.4, the observations finally declared as outliers will be those for which the statistics (3.6) and (3.7) are significant. Nonetheless a main problem remains. The stable model estimates may be biased by outliers positioned so that one or more are not identified as outliers: the well known *masking* problem.

I have extensively tested this backward strategy for the analysis of trade data at the Joint Research Centre, where it is in use and is producing rather satisfactorily results. This is why I decided to include this (non robust) method in this benchmark. However the assessment will be limited to data with a single predictor variable ($p = 2$), as it is in trade data, because it is well known that the masking problem becomes more and more severe as the dimensionality increases.

4.4 Benchmark setup for the statistical size

We now describe the configuration of the benchmark which assesses the statistical size of four linear regression methods: Forward search, Least Median of Squares, Least Trimmed Squares and a backward iterative method.

The benchmark took some months to complete on a computer platform consisting in two PCs with 32 bit CPU, Intel 2,3 GHz, 1,5 GBytes of RAM. The statistical programming environment used is Matlab, inclusive of the Statistical toolbox. LTS and LMS were run using the LIBRA toolbox for Matlab (Library for Robust Analysis) developed by Mia Hubert¹. For the Forward Search I used the FSDA toolbox for Matlab (Forward Search for

¹LIBRA is available at <http://wis.kuleuven.be/stat/robust/LIBRA.html>

Data Analysis)², of which I am one of the developers. The backward method based on regression diagnostics was implemented in MATLAB by myself, but the original version in SAS was conceived in the Joint Research Centre, also with my contribution.

The Forward Search was run at the standard nominal significance level of $\alpha = 0.01$. The subset from which the Forward Search starts its progression has been computed by running LTS on 1000 sub-samples of the dataset.

I used the re-weighted version of LTS and LMS (Section 4.2). To obtain an overall size of $\alpha = 0.01$, each LTS/LMS session was run at a Bonferroni-corrected size of $(1 - \frac{0.01}{n})$, where n is the sample size. Moreover, the initial robust LTS/LMS estimator was found by extracting 10000 sub-samples. This choice is motivated by the fact that the results obtained with 10000 sub-samples were considerably different from those obtained with only 1000 sub-samples, but very close to those obtained with 50000 sub-samples.

The backward method was also run to achieve an overall significance level $\alpha = 0.01$, with Bonferroni-corrected individual tests.

Independently from the method, each benchmark experiment was based on 10000 replications, i.e. on 10000 sets of data. To check if this number of replications was sufficient to achieve a reasonable accuracy on the size estimates, we have also performed few benchmark experiments with 50000 replications. The results were very close, as the table 4.1 shows for the Forward Search case, and therefore we have limited the benchmark experiment to 10000 sets of data.

Simulations were made for five sample size values $n = 1000, 500, 400, 200, 100$ and for four values for the number of explanatory variables 1, 2, 5, 10. Since in Section 3.5 we have defined the number of columns of X as p and since X has a number of columns equal to the number of explanatory variables plus 1 for the intercept, we can say that simulations have been made for $p = 2, 3, 6, 11$. Results were declared significant if *at least* one outlier was detected.

The null hypothesis to test is that data is normal, $H_0 : \{e_1 \sim N(0, \sigma^2)\} \cap \{e_2 \sim N(0, \sigma^2)\} \cap \dots \cap \{e_n \sim N(0, \sigma^2)\}$, therefore the response variable of the data has been generated from a normal distribution in each replication of the benchmark experiment.

The values of the explanatory variable were fixed for all the 10000 replications (sets of data). Such values were also generated from a normal distribution, so that to “mitigate” the effect of leverage points.

Other options to generate the explanatory variable values could be le-

² The FSDA toolbox can be downloaded from Marco Riani’s website <http://www.riani.it/MATLAB.htm>

gitimate. For example, one might generate them uniformly or, if the focus is on a specific application domain, one might estimate by sampling the distribution of the explanatory variable in a set of real data (e.g. trade data) and replicate by simulation that distribution. However this would bring to completely different benchmark experiments that we do not address in the thesis. We just report here that we checked the option of uniform explanatory values for few n and p combinations and such tests brought to results very similar to the normal case discussed in the Chapter.

Another even more specific benchmark experiment would result from not fixing the explanatory variable values in all replications, but generating them from some distribution (e.g. normal, uniform, etc.). This would bring us into the area of stochastic regression.

4.5 Results on the statistical size

We have summarized our findings on the test size in the tables of Figures 4.1 and 4.2 and in the plots in Figure 4.3.

The left table of Figure 4.1 shows the results on the Forward Search empirical α obtained with 50000 replications, i.e. sets of simulated data. The corresponding results on 10000 replications are on the right table. The difference between the comparable values in the two tables is negligible. On the contrary, a similar check on 1000 replications gave results considerably different and probably inaccurate. Therefore, we have run all other benchmark experiments using 10000 subsets of data.

In the Figure 4.2 we have reported the results for LTS (left table) and LMS (right table) α . Finally, in the four plots of Figure 4.3, we have reported for $p = 2$ (top left plot), $p = 3$ (top right plot), $p = 6$ (bottom left plot) and $p = 11$ (bottom right plot) the actual α of Forward Search (solid

n	$p = 3$	$p = 5$	$p = 11$	$p = 2$	$p = 3$	$p = 6$	$p = 11$
100	0.0105	0.0114	0.0262	0.0118	0.0099	0.0127	0.0283
200	0.0115	0.0106	0.0173	0.0171	0.0108	0.0112	0.0173
400	0.0122	0.0105	0.0154	0.0178	0.0121	0.0109	0.0151
500	0.0126	0.0112	0.0149	0.0175	0.0132	0.0095	0.0126
1000	0.0130	0.0115	0.0162	0.0167	0.0138	0.0117	0.0147

Figure 4.1: Empirical size of the nominal 1% outlier test for the Forward Search based on 50000 (left table) and on 10000 (right table) sets of data, for different sample sizes n and number of explanatory variables $p - 1$.

n	$p = 2$	$p = 3$	$p = 6$	$p = 11$	$p = 2$	$p = 3$	$p = 5$	$p = 11$
100	0.0142	0.0205	0.0377	0.0777	0.0063	0.0059	0.0055	0.0053
200	0.0142	0.0122	0.0152	0.0211	0.0054	0.0050	0.0072	0.0069
400	0.0105	0.0116	0.0133	0.013	0.0083	0.0082	0.0077	0.0080
500	0.012	0.0104	0.0124	0.0096	0.0082	0.0080	0.0086	0.0077
1000	0.0103	0.0079	0.0108	0.0117	0.0090	0.0081	0.0086	0.0097

Figure 4.2: Empirical size of the nominal 1% outlier test based on 10000 sets of data for the LTS (left table) and LMS (right table) for different sample sizes n and number of explanatory variables $p - 1$.

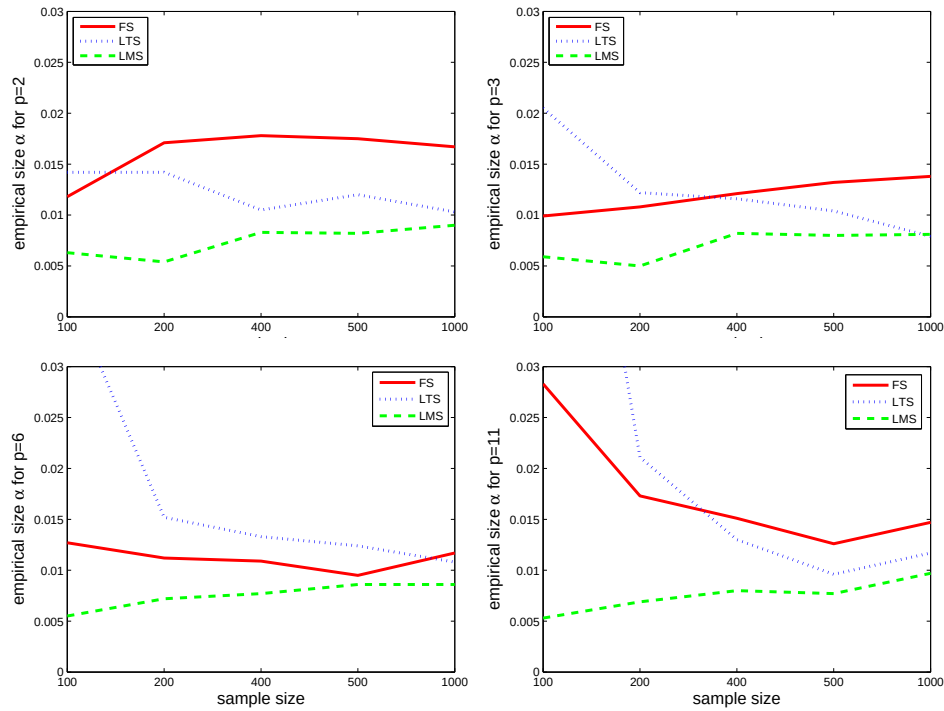


Figure 4.3: Empirical size of Forward Search (solid red line), LTS (dotted blue line) and LMS (dashed green line) for a nominal 1% outlier test for $p = 2$ (top left plot), $p = 3$ (top right plot), $p = 6$ (bottom left plot) and $p = 11$ (bottom right plot) and for sample sizes among 100 and 1000 (horizontal axes).

red line), LTS (dotted blue line) and LMS (dashed green line).

By looking at the four plots, it is clear that LMS has an empirical α which is constantly lower than Forward Search and LTS. In particular it is surprising that the LMS empirical α is even lower than the nominal size of 1%. Another aspect common to all p values is that the LTS α tends to decrease with n , and for small sample sizes it assumes rather high values. Finally, in the four plots the Forward Search empirical α is just above the nominal value: in particular for $p = 3$ and $p = 6$ the maximum value is 0.013, while for $p = 2$ and $p = 11$ the maximum values are 0.0178 and 0.0283 respectively.

Now, we remind that our aim is to analyse the empirical size of the three methods in relation to the corresponding power values. In order to compare the power values of different tests, it is necessary that these tests have approximately the same empirical size. Based on the plots in Figure 4.3 the combination of p and n for which the three size curves are closer is $(p = 6, n = 500)$. Here the LTS has the highest empirical α (0.0124), followed by the Forward Search (0.0095) and finally by the LMS (0.0086). Therefore the benchmark for estimating the empirical power (Section 4.8) will concentrate on the specific combination $(p = 6, n = 500)$.

4.6 Results on the statistical size for small samples

In the international trade scenario there are problems that require the analysis of few years of monthly aggregates of the trade declarations within two countries, the origin and the destination of a good. In this case the datasets to analyse are small, composed by less than 50 transactions each. We stressed already on the good properties of the theoretical envelopes that the Forward Search uses for the outlier tests, including the fact of accounting automatically for the multiplicity of the tests.

However, till now it has been never shown how good are in practice these approximated envelopes for small sample size values. Here we fill the gap by assessing the test size for datasets of sample sizes multiple of 5 in the range $15 \leq n < 100$. We limited this specific benchmark to one dependent variable ($p = 2$). There are two simple reasons for this choice. One is simply that in trade analysis typically the trade value is regressed only over the quantity. The second is that the deletion residual test is based on $(n - p - 1)$ degrees of freedom and it is not reasonable to reduce to too few degrees of freedom.

n	FS	LTS	LMS	BOR
15	0.0067	0.2389	0.0017	0.0214
20	0.0096	0.0297	0.0012	0.0177
25	0.0104	0.0244	0.0012	0.0199
30	0.0092	0.0262	0.0021	0.0212
35	0.0076	0.0183	0.0014	0.0205
40	0.0084	0.0195	0.0025	0.0204
45	0.0082	0.0191	0.0026	0.0207
50	0.0095	0.0180	0.0026	0.0197
55	0.0105	0.0180	0.0039	0.0220
60	0.0092	0.0169	0.0026	0.0241
65	0.0094	0.0162	0.0040	0.0214
70	0.0107	0.0173	0.0041	0.0193
75	0.0119	0.0151	0.0048	0.0205
80	0.0105	0.0162	0.0043	0.0236
85	0.0104	0.0136	0.0042	0.0195
90	0.0115	0.0130	0.0039	0.0200
95	0.0134	0.0166	0.0062	0.0217

Figure 4.4: Empirical size α of Forward Search, LTS, LMS and the backward method (BOR), for $p = 2$ and n among 15 and 95.

The Table 4.4 and the left plot of Figure 4.5 report the resulting α estimated for the Forward Search, LTS, LMS and the backward technique described in Section 4.3. In order to compare the performances for small sample sizes with those between 100 and 1000 discussed in the previous Section, the right panel of Figure 4.5 shows the top left plot of Figure 4.3, with the results of the backward method superimposed.

For small n (left plot) the Forward Search (solid red line) shows an α increasing from 0.6% to 1.13%, while, for larger n (right plot), it increases and stabilizes around 1.6%. The LTS (dotted blue line) for n small has a simulated α with high variability which assumes values greater than 3%, while, for larger n , it stabilizes around a value just above 1%. The LMS (dashed green line) always has very low α values, but for small n the size is even less than 0.5%. The backward method (black line with stars) is stable around 1% for all values of n .

By looking at the first plot in Figure 4.5 we can see that the simulated α of LTS is constantly higher than the one of the Forward Search, which is very close to that of the backward method and higher than the one of

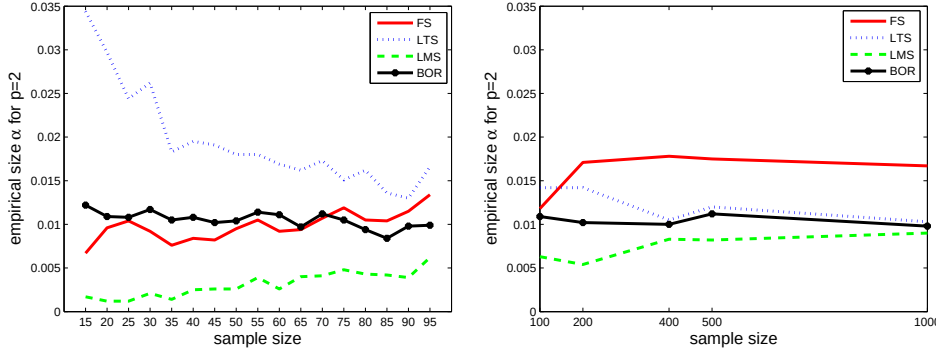


Figure 4.5: Empirical size α of Forward Search (solid red line), LTS (dotted blue line), LMS (dashed green line) and the backward method (BOR, black line with stars) when $p = 2$ and n among 15 and 95 (left plot) and among 100 and 1000 (right plot).

the LMS. There is no particular value of n for which the curves are much closer, thus we have decided to monitor the power for $n = 50$ (Section 4.9). Here the backward method has the highest α (0.0197), followed by the LTS (0.0180), followed by the Forward Search (0.0095) and finally by the LMS (0.0026).

4.7 Benchmark set-up for the statistical power

The benchmark on the statistical power inherits all the settings of the benchmark on the size. Of course for the evaluation of the power it is necessary to contaminate the response variable. Therefore, after having generated the response and the explanatory variables from a normal distribution we have contaminated 5% of the response values with a shift of increasing size, from 1 to 7.

Five different measures of power have been computed and analysed:

1. Average power: average over all the iterations of the number of true detected outliers with respect to the contaminated observations.
2. Simultaneous power: average over all the iterations of the number of detected (both true and false) outliers.
3. Family wise error rate: average number of iterations where at least one false outlier has been detected.

4. False discovery rate: average over all the iterations of the number of false detected outliers with respect to all the detected (both true and false) outliers.
5. Proportion of declared outliers in good data: ratio between the average number of false detected outliers with respect to the the number of non-contaminated observations.

As anticipated in Sections 4.5 and 4.6, we will present the results of our simulations on the power only for the combination $(n = 500, p = 6)$, where the simulated α of the methods are closer, and for the combination $(n = 50, p = 2)$.

4.8 Results on the statistical power

The tables associated with Figure 4.6 report the results obtained for the five measures of power introduced in the previous Section, for the combination $n = 500$ and $p = 6$. Figure 4.7 and Figure 4.8 are respectively the graphical representation of the measures of power and of their logit. The application of the logit function is to reduce the scale effects and make the power curves more comparable.

Considering average and simultaneous power, we can easily conclude that, for all shift levels, the Forward Search performs better than LTS: the Forward Search power measures are higher than the LTS ones even if LTS, as shown in Section 4.5, has a higher α . We cannot draw so neat conclusions on LMS, which was characterised by the smallest simulated α . However we think that the small difference in terms of size, 0.0009 and 0.0038 between LMS and respectively the Forward Search and LTS, is not sufficient to justify such a large distance among the two power curves. Besides, being widely accepted that LTS has larger power than LMS, we can conclude that the Forward Search (which, as seen before, has better power than LTS) is the most powerful among the three methods.

Let us focus now at the three last plots of Figure 4.7. The LTS and LMS curves are constantly very close to zero while the Forward Search curve appears far from the other two. Note however that in the plots of the false discovery rate and proportion of declared outliers in good data the y-axis scales are so small (from 0 to 10^{-4} and 10^{-3} respectively) that we can consider this difference not relevant for these two measures.

On the contrary, in the plot of the family wise error rate the range of the y-scale is not irrelevant and the bell shape of the Forward Search curve looks quite curious. The interpretation of this phenomenon is as follows.

<i>shift</i>	<i>FS</i>	<i>LTS</i>	<i>LMS</i>	<i>shift</i>	<i>FS</i>	<i>LTS</i>	<i>LMS</i>
1	0.0011	0.0003	0.0002	0	0.0095	0.0116	0.0082
2	0.0104	0.0037	0.0024	1	0.0142	0.0154	0.011
3	0.2111	0.0232	0.0116	2	0.1267	0.0904	0.0592
4	0.6261	0.1368	0.0250	3	0.8968	0.4109	0.2594
5	0.8771	0.5487	0.0546	4	1	0.9139	0.4893
6	0.9654	0.9187	0.1017	5	1	0.9999	0.789
7	0.9958	0.9949	0.2734	6	1	1	0.9332
				7	1	1	0.9884

<i>shift</i>	<i>FS</i>	<i>LTS</i>	<i>LMS</i>
0	0.0095	0.0116	0.0082
1	0.0075	0.0075	0.0053
2	0.0233	0.0028	0.001
3	0.1816	0.0007	0.0002
4	0.1958	0.0006	0
5	0.1049	0.0014	0
6	0.0432	0.0062	0
7	0.0276	0.0092	0

<i>shift</i>	<i>FS</i>	<i>LTS</i>	<i>LMS</i>	<i>shift</i>	<i>FS</i>	<i>LTS</i>	<i>LMS</i>
0	1.00E-04	1.00E-04	1E-04	0	8.9E-04	2.36E-05	1.6E-05
1	9.29E-05	4.84E-05	4.8E-05	1	0.0007	1.58E-05	1.1E-05
2	7.38E-05	2.93E-06	1.7E-06	2	0.0015	5.89E-06	2.1E-06
3	1.35E-05	1.20E-07	6.9E-08	3	0.0017	1.47E-06	4.2E-07
4	4.80E-06	1.75E-08	0	4	0.0017	1.26E-06	0
5	3.16E-06	1.02E-08	0	5	0.0015	2.95E-06	0
6	2.73E-06	2.74E-08	0	6	0.0014	1.33E-05	0
7	2.57E-06	3.74E-08	0	7	0.0014	1.96E-05	0

Figure 4.6: From the top: average power, simultaneous power, family wise error rate, false discovery rate, proportion of declared outliers in good data for Forward Search (FS), LTS and LMS, for sample size $n = 500$, number of explanatory variables $p - 1 = 5$ and levels of shift among 0 and 7.

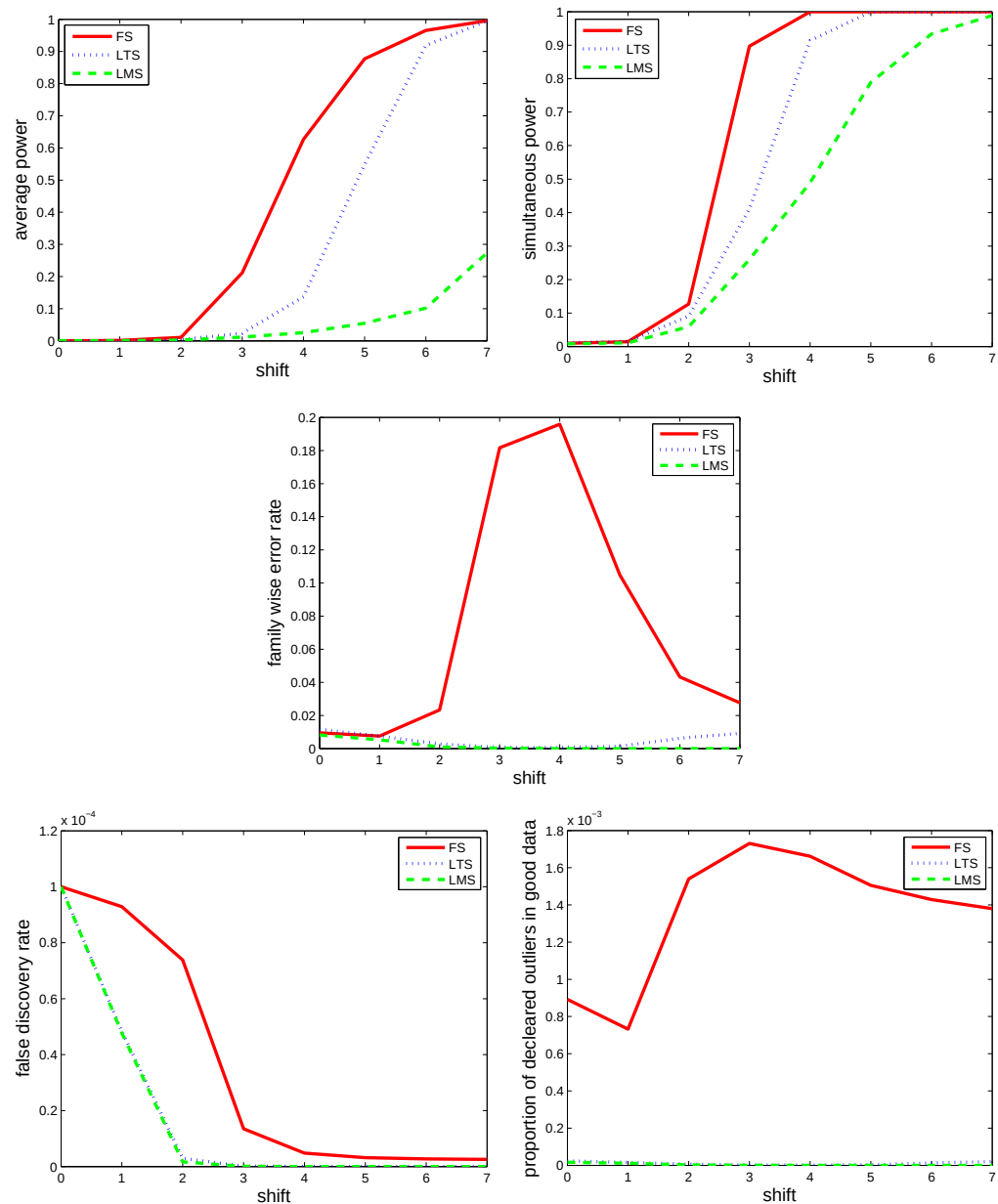


Figure 4.7: From the top: average power, simultaneous power, family wise error rate, false discovery rate, proportion of declared outliers in good data for Forward Search (FS, solid red line), LTS (dotted blue line), LMS (dashed green line) for sample size $n = 500$, number of explanatory variables $p - 1 = 5$ and levels of shift among 0 and 7 (horizontal axis).

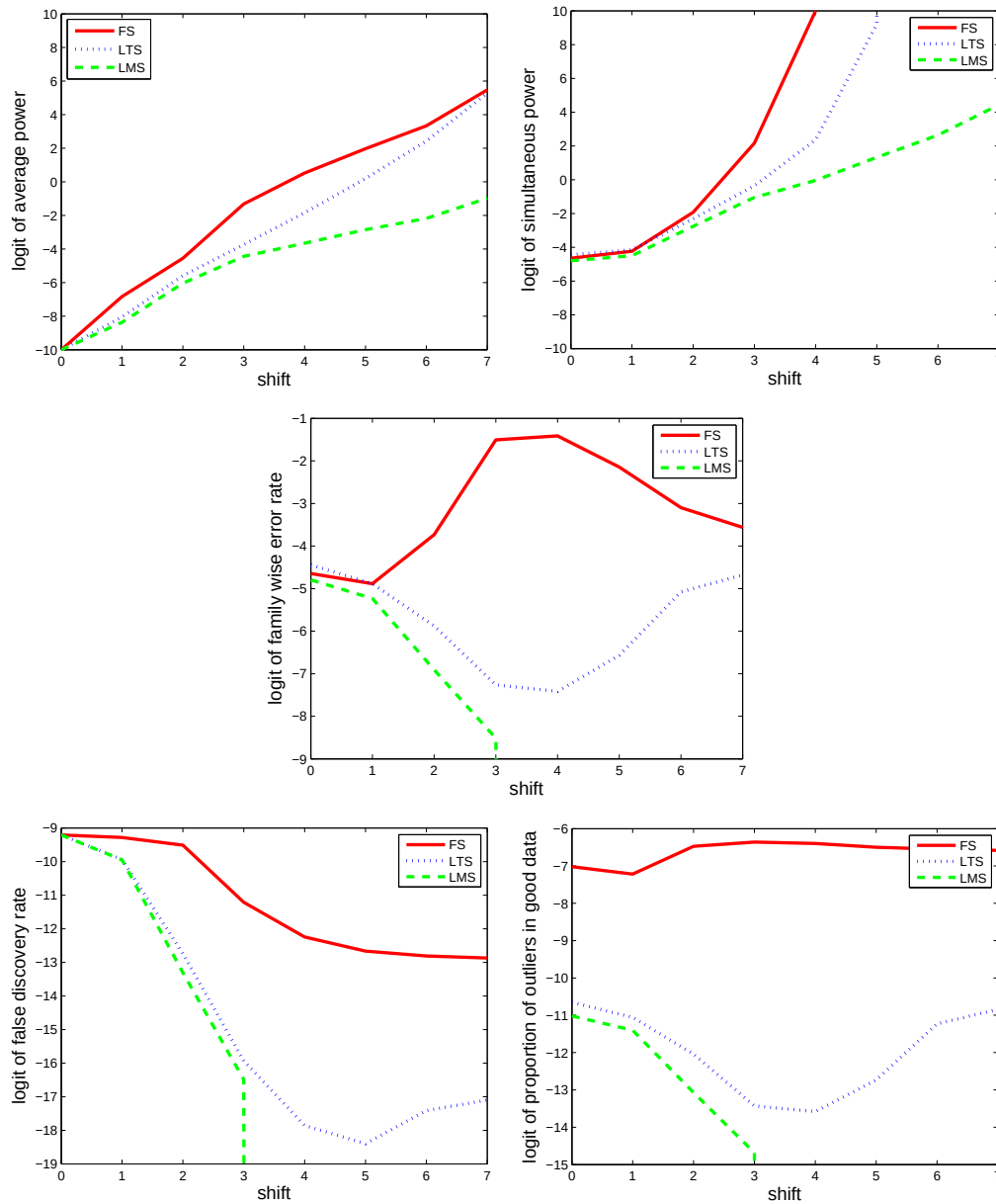


Figure 4.8: From the top: logit of average power, simultaneous power, family wise error rate, false discovery rate, proportion of declared outliers in good data for Forward Search (FS, solid red line), LTS (dotted blue line), LMS (dashed green line) for sample size $n = 500$, number of explanatory variables $p - 1 = 5$ and levels of shift among 0 and 7 (horizontal axis).

When a small shift of 1 or 2 is applied to values generated from a $N(0, 1)$, the contaminated observations are unlikely to be detected as outliers. In fact with contaminations of shift 1 or 2 the average power is very close to zero and the family wise error rate is limited. For larger shifts, 3 and 4, the family wise error rate increases: real outliers are not still well identifiable and can be confused with good observations. For the largest shifts 5, 6 and 7, the real outliers are well identifiable and so the family wise error rate decreases.

4.9 Results on the statistical power for small samples

This Section discusses simulations and results for the power of the methods when the sample size and dimensionality are small. The choice of $n = 50$ and $p = 2$ was justified in Section 4.9.

Like in the previous Section, we report in Table 4.9 the values of the five measures of power of interest, and in the Figures 4.10 and 4.11 we give the graphical representations of such results and their logit transformation, which provides a more interpretable scale.

The average and simultaneous power curves for LTS are just a bit higher than the Forward Search curve, while the LMS curve is considerably lower than the previous two. Note also that, if LTS has the highest simulated power, it also has the highest α (Figure 4.5). Therefore, in this case we cannot establish a clear performance ranking among the methods, as we could do in the previous Section for the more standard combination of $n = 500$ and $p = 6$ and for the various levels of shift.

For the false discovery rate and proportion of declared outliers in good data the y-axis scales are so small (from 0 to 10^{-4} and 10^{-3} respectively in Figures 4.10 and 4.11) that such cases do not deserve further attention.

Finally, concerning the family wise error rate, we remark that in this case we do not find again the bell shape that the Forward Search curve shown for the combination $n = 500$ and $p = 6$. For the moment we do not have a clear interpretation of the phenomenon.

<i>shift</i>	<i>FS</i>	<i>LTS</i>	<i>LMS</i>
1	0.0057	0.0026	0.0007
2	0.0152	0.0244	0.0045
3	0.0577	0.1122	0.0241
4	0.2479	0.3309	0.0412
5	0.6254	0.6714	0.138
6	0.9149	0.9071	0.3512
7	0.9880	0.982	0.6103

<i>shift</i>	<i>FS</i>	<i>LTS</i>	<i>LMS</i>
0	0.0095	0.018	0.0026
1	0.0106	0.0148	0.0015
2	0.0078	0.0084	0.0005
3	0.0094	0.0064	0
4	0.0133	0.0074	0
5	0.016	0.0095	0.0002
6	0.0162	0.0111	0.0005
7	0.0163	0.012	0.0009

<i>shift</i>	<i>FS</i>	<i>LTS</i>	<i>LMS</i>
0	1E-04	1E-04	1E-04
1	9.3E-05	6.7E-05	4.0E-05
2	8.1E-05	1.1E-05	3.6E-06
3	5.5E-05	2E-06	0
4	1.9E-05	7.6E-07	0
5	1E-05	5.0E-07	4.8E-08
6	7.7E-06	4.3E-07	4.7E-08
7	6.5E-06	4.1E-07	4.9E-08

<i>shift</i>	<i>FS</i>	<i>LTS</i>	<i>LMS</i>
0	0.0035	3.9E-04	5.2E-05
1	0.0049	0.0003	3.2E-05
2	0.0042	0.0002	1.1E-05
3	0.0045	0.0001	0
4	0.0036	0.0002	0
5	0.0046	0.0002	4.3E-06
6	0.0049	0.0002	1.1E-05
7	0.0044	0.0002	1.9E-05

Figure 4.9: From the top: average power, simultaneous power, family wise error rate, false discovery rate, proportion of declared outliers in good data for Forward Search (FS), LTS and LMS, for sample size $n = 50$, number of explanatory variables $p - 1 = 1$ and levels of shift among 0 and 7.

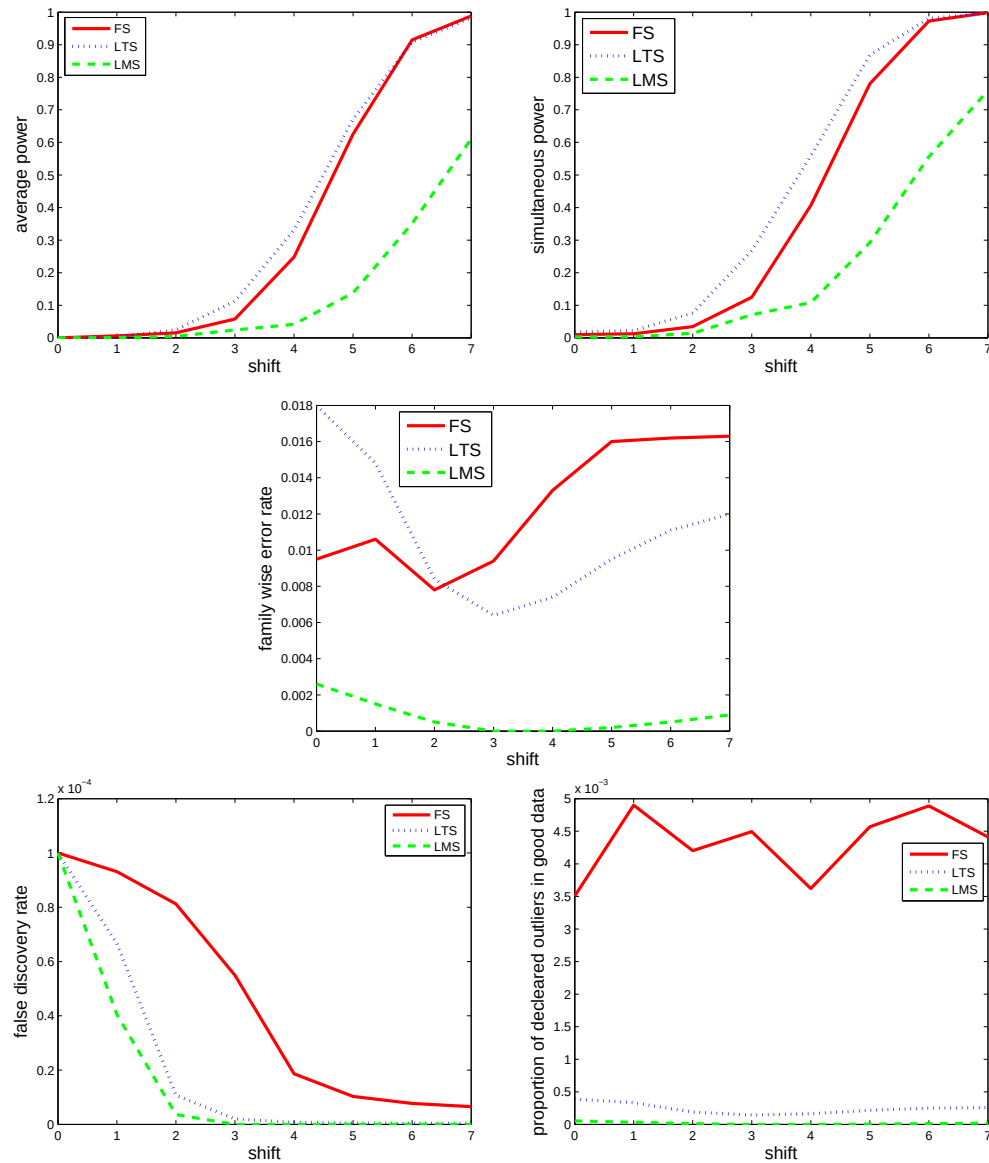


Figure 4.10: From the top: average power, simultaneous power, family wise error rate, false discovery rate, proportion of declared outliers in good data for Forward Search (FS, solid red line), LTS (dotted blue line), LMS (dashed green line) for sample size $n = 50$, number of explanatory variables $p - 1 = 1$ and levels of shift among 0 and 7 (horizontal axis).

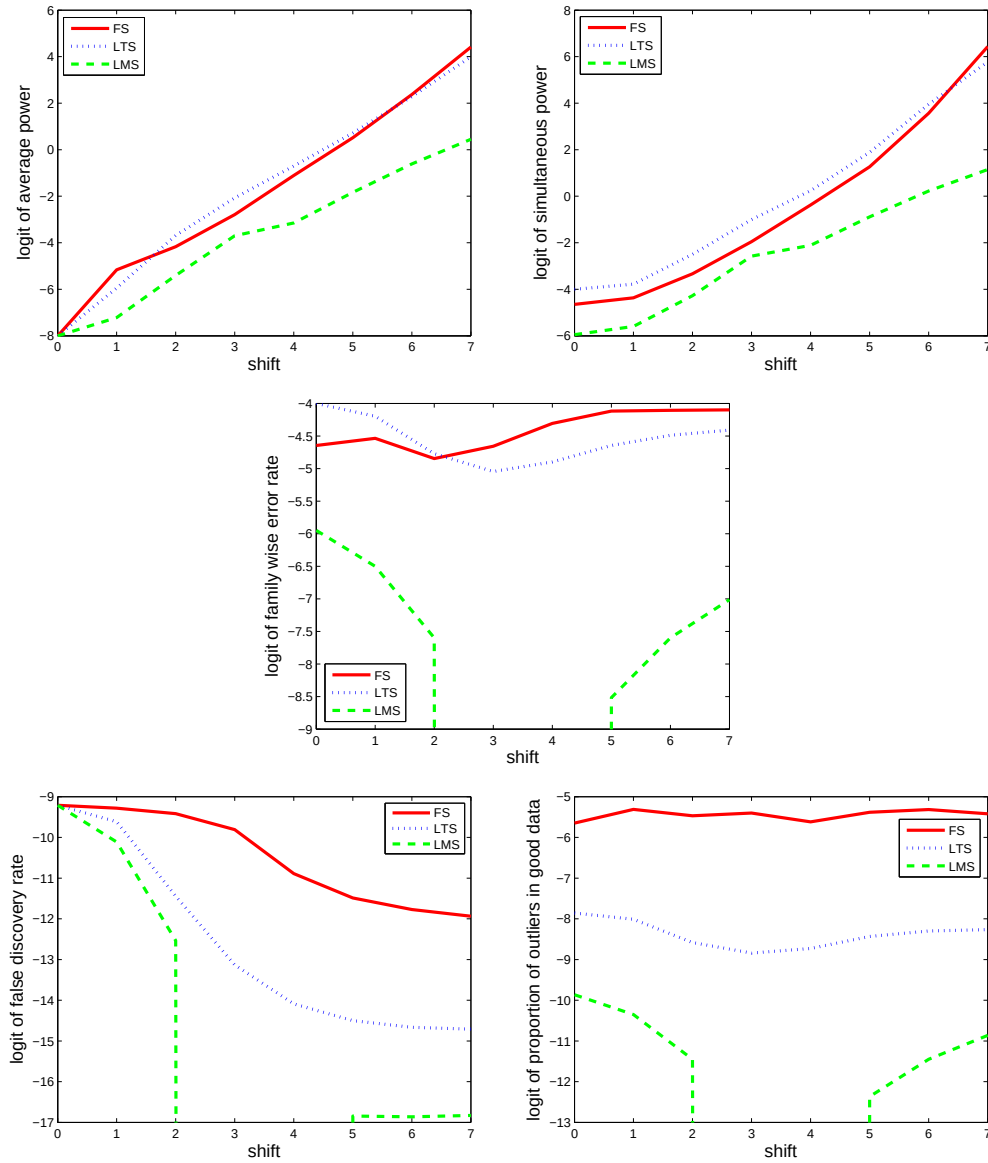


Figure 4.11: From the top: logit of average power, simultaneous power, family wise error rate, false discovery rate, proportion of declared outliers in good data for Forward Search (FS, solid red line), LTS (dotted blue line), LMS (dashed green line) for sample size $n = 50$, number of explanatory variables $p - 1 = 1$ and levels of shift among 0 and 7 (horizontal axis).

4.10 The statistical power of the backward method

It is well known that the backward iterative approaches suffer of low power as the dimensionality increases. Nonetheless we dedicate a separate section to the average power of the backward method for the case $p = 2$, which is of interest in the analysis of trade data.

We start with the average power for the combination $n = 50$ and $p = 2$, that we already discussed for the other methods with the top left plot of Figure 4.10. We superimpose to this plot the average power found for the backward method for the same (n, p) combination. The overall picture, in the left plot of Figure 4.12, shows a rather acceptable power performance of the backward method in relation with the Forward Search, that has comparable size according to Figure 4.5. Difficult to draw conclusions in relation to LMS/LTS that, compared to the backward method, have quite bigger/smaller size values.

Let us consider now the combination $(n = 1000, p = 2)$, in the right plot of Figure 4.12. For this combination the results for shift values equal to 1 and 2 are not available at the time of this writing. However, for the most relevant shift values, the backward method has smaller power than the Forward Search, although it should be considered that the Forward Search for the combination $(n = 1000, p = 2)$ has a slightly bigger size (see Figure 4.5).

We omit to report results on the other power measures for the backward method.

4.11 Considerations on computational efficiency

In the benchmark experiments we have also monitored the CPU time needed to execute the four methods, for a range of sample sizes n and number of explanatory variables $p - 1$. Here we report the time estimates monitored during the size benchmark. One might expect for the Forward Search some degradation in the time performances in presence of contamination, because of the process of superimposing the envelopes. In practice the Forward Search time values monitored during the power benchmarks did not show a significant increase, and we omit them.

The estimate of the computational time of a given technique, for given n and p , is the mean of the CPU time values monitored for each of the

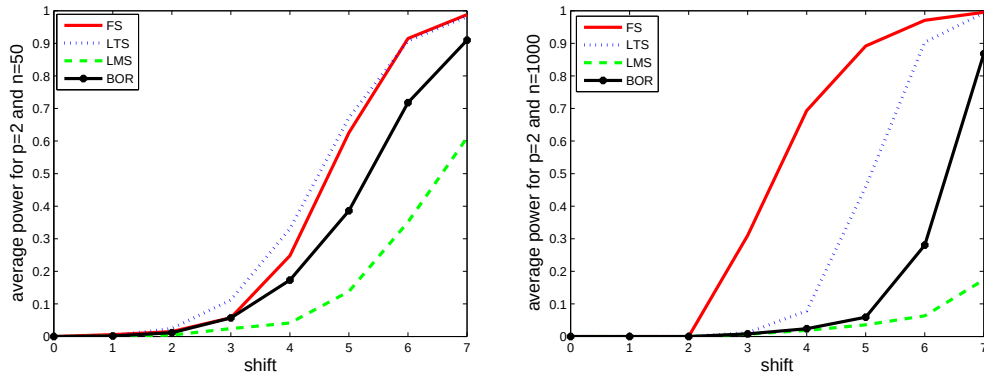


Figure 4.12: The average power of the backward method (BOR, Backward Outlier Regression, in the legend) is compared with that for Forward Search (FS, solid red line), LTS (dotted blue line) and LMS (dashed green line), for one explanatory variable ($p = 2$) and levels of shift among 0 and 7 (horizontal axis). On the left the results are for the sample size $n = 50$, on the right for $n = 1000$.

10000 replications. The results are summarised in Figure 4.13.

The backward approach (black line with stars in the first plot of Figure 4.13) is based on a computationally simple and efficient algorithm, which is therefore very fast compared to the other methods. Among the robust approaches, computationally the Forward Search is the best performing technique: the average time curve lies below LTS and LMS for all sample size values n and independently from the number of explanatory variables $p - 1$. Although the Forward Search finds an initial robust estimator using the LTS algorithm, the LTS is applied only once to find a small and sub-optimal subset of $p + 1$ observations (at most 1000 random subsets are considered). This explains why the call to LTS does not increase significantly the Forward Search computational time.

The LTS time estimates are those for the standard re-weighted LTS. In Section 4.2 we have discussed a fast version of the LTS, introduced precisely to overcome the not satisfactory computational performances of the standard method. Therefore we also have made an assessment of the LTS through the fast LTS algorithm. However the time estimates of the fast algorithm were still not comparable with those of the Forward Search, and therefore we do not report them here.

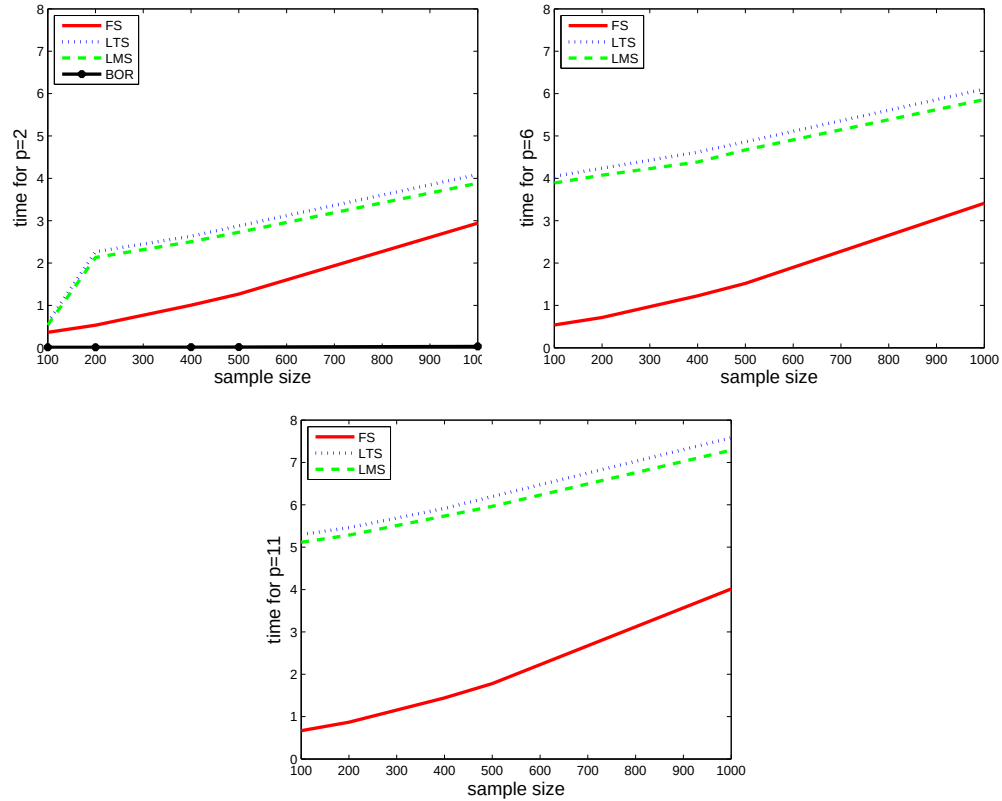


Figure 4.13: Estimated elapsed time for Forward Search (FS, solid red line), LTS (dotted blue line), LMS (dashed green line) when the sample size n is among 100 and 1000 and $p = 2$ (top left plot), $p = 6$ (top right plot) and $p = 11$ (bottom left plot). For $p = 2$ the results of the backward method (BOR) are also represented with a dotted blue line with stars.

4.12 Conclusions

The results of this Chapter show first of all that the Forward Search tests have good statistical size for a wide range of sample sizes ($15 \leq n \leq 1000$) and dimensions ($2 \leq p \leq 11$). The good performance for small sample sizes confirms the excellent properties of the theoretical envelopes introduced by [Ria09].

The results also show a very low size of LMS for the full range of dimensions n , and a good size for LTS but for $n > 100$.

The methods can be conveniently compared in terms of power for the specific combination $n = 500, v = 5$, in correspondence of which the sizes of the three methods are the closest. In this case, it is clear that the performance of the Forward Search is superior, followed by LTS and then

by LMS.

The interpretation of the corresponding results for smaller n are not so neat. In fact for n between 15 and 100 the LMS has a statistical size unduly small, followed by the Forward Search and by LTS. However for $n = 50$ and $p = 2$ LTS has the highest average power, just followed by the Forward Search and LMS.

Again for small sample size and one explanatory variable, we note a rather acceptable power performance of the backward method in relation with the Forward Search, that has comparable size. However the backward power deteriorates for bigger sample sizes and is not even considered here for higher dimensions, where the masking effect is much stronger.

An empirical analysis of the computational complexity of the four methods has also shown that the Forward Search in practice is the most convenient from a computational point of view (excluding the backward method, that is not comparable to the other methods in terms of statistical performances). This is because LMS/LTS have a severe combinatorial problem to face with, while the Forward Search can use such methods for the choice of the initial subset using a very reduced number of random subsets.

A last comment on the fact that this benchmark still deserves more attention and work. Thanks to the detailed analysis of this Chapter, we are now in the position to monitor the frequency of the types of Forward Search signals described in the Appendix of the dissertation, identify those which determine more false alarms, and therefore fine-tune the algorithm to improve its statistical performances.

Extending the Forward Search to Linear Mixtures: a Proposal

5.1 Introduction

The detection of outliers in trade data has to face several levels of complexity that I started to discuss in Section 2.3. In this Chapter I address one that occurs quite often in trade data and that is not possible to address with traditional robust methods: the aggregation of points around an unknown number of regression lines, i.e. the presence of data generated by a mixture of linear components.

In addressing the problem, I propose a new methodology that extends the Forward Search to the estimation of linear regression mixtures, for the general case where the number of components and the membership of the observations to one *or none* of the components are both unknown. The methodology is conceived in two stages. In the first stage the observations are divided into tentative groups and a preliminary fit is done on each of the tentative groups (Section 5.2), in the second stage the clustering and possible outliers (i.e. unassigned observations) are confirmed or reassigned (Section 5.3).

The corresponding problem of clustering multivariate data with the Forward Search was thoroughly addressed in the book of Atkinson *et al.* [ARC04], but in regression the issue was not yet studied. The methodology proposed here is therefore an original contribution of the thesis.

More in general the identification and estimation of linear regression mixtures in data is a problem very well studied in the statistical literature. In the Chapter, I will refer for example to the works of DeSarbo [DC88], Neykov *et al.* [NFDN07] and Hennig [Hen03], which I will assess against

the proposed Forward Search solution. However, typically in trade data the mixture estimation problem shows peculiar aspects that are very difficult to address because of issues that I briefly recall here from Section 2.2.

First of all, in trade data the regression components are not well separated. Secondly, there are observations that do not follow any of the regression lines in the mixture. These outlying observations may be isolated, caused for example by recording errors during the data collection process, or clustered, when they represent a systematic behaviour. These groups of outliers contain useful information for fraud detection purposes and therefore their identification is of primary importance. Also inliers that do not follow any specific regression structure and are intermediate among different regression lines are of interest, especially if they are economically relevant, that is those with large values of either variable. Of course, in the absence of fraud, large values of both variables occur together. In synthesis, the statistical problem addressed in the Chapter is:

1. To provide an estimate of the unknown number of regression lines.
2. To estimate the parameters of these lines and identify the observations clustering around each of them.
3. To identify clusters of anomalous transactions, if any, and inliers that do not belong to any regression line.

All the three steps must be performed robustly to ensure that multiple outliers do not mask each other. The estimated regression lines of a mixture are usually called mixture components. However, since here such lines are also used to partition the data into groups of observations, I may refer to one of such groups using the term ‘mixture component’ as a synonym of ‘cluster’.

In application areas where the data are highly complex, like in trade data, the information to be extracted is strongly driven by the commitment of the final user, i.e. the subject matter analyst, who grounds his judgement on background knowledge or any other information available. Sometimes this information is coded into concomitant variables in the data that are not used in the regression analysis, e.g. in trade data the code of the product, or its origin and destination countries. In other cases this information even comes from sources external to the dataset under analysis, e.g. in the trade context information on the Community measures relating to imports and exports, e.g. import duties applied to specific goods, are found in a publicly accessible database called TARIC.

Therefore, by assuming that in complex real world applications it is often reasonable to rely on the subjective judgement and final aim of the analyst

or subject matter expert, I accept the possibility of different conclusions for the same clustering.

This means that the Chapter will only marginally touch upon formal or objective criteria to validate the proposed methodology (Section 5.9). The only sensible requirement that will be followed is that *a valid cluster should be formed by homogeneous observations, separated from the rest of the observations in the dataset*, where the concepts of “homogeneity” and “separation” will be substantiated case by case for each of the methods discussed.

5.2 Tentative allocation of the groups

Basic idea: repeated group identification and trimming

We suppose that the observations come from g regression models with different and unknown parameter values. Our aim is to allocate each unit to its true model and to estimate the corresponding parameters. Also the number g of component models is not known in advance. Our proposal is to iteratively use the Forward Search for fitting the g components of the regression mixture. More precisely our forward algorithm is summarized as follows.

Strategy 5.1 Tentative Forward Search allocation of the groups to the mixture components.

1. Let $n^*(j)$ be the size of the sample to be analysed at iteration j . At the first iteration $n^*(1) = n$.
2. The Forward Search for regression is applied to these $n^*(j)$ observations. The search is initialized robustly through the least median of squares criterion and progresses using the squared regression residuals $e_i^2(m), i = 1 \dots, n^*(j)$.
3. At each step m of the Forward Search, we test the null hypothesis that there are no outliers in the $n^*(j)$ observations. The test is performed using the minimum deletion residual among observations not in the subset (equation 3.5) and the technique developed by Riani *et al.* [RAC07] to keep simultaneity into account.
4. If the sequence of tests performed in step 3 does not lead to the identification of any outlier, the sample of $n^*(j)$ observations is declared

to be homogeneous and the algorithm stops by fitting a regression model to this sample. Otherwise go to step 5.

5. Let m^* be the step of the Forward Search in which the null hypothesis of no outliers is rejected by the sequence of tests of step 3. Then the observations in $S(m^*)$ identify one mixture component, i.e. one cluster of m^* observations following the regression model found in step 3. Fit a new regression model to this cluster.
6. Remove the cluster identified in step 5. Return to step 1 with a reduced sample size, by setting $n^*(j+1) = n^*(j) - m^*$.

The strong points of this approach can be summarised as follows.

- The algorithm leads to the identification of g regression models, one for each iteration.
- The tests performed in step 3 ensure that each component of the mixture is fitted to a homogeneous subset.
- The tests are robust and are not influenced by outliers or by observations falling between the groups. Indeed such observations, which in trade data may be relevant for fraud detection, are clearly revealed by the forward diagnostic plots during the search.
- The method does not force all observations to be firmly clustered into one of the g components.
- Borderline units are recognized as intermediate between clusters and can thus be inspected separately.

We now demonstrate these properties, and discuss possible drawbacks of the algorithm, on the Fisheries dataset already addressed in Sections 2.4 and 3.5. Later, in Section 5.5, we will compare this Forward Search solution with other mixture estimation methods on the more complex trade dataset (imports of technological products) introduced in Section 2.2.

The four groups that the algorithm finds in the Fishery dataset and the mixture components fit on the groups are represented in Figure 5.1. Figure 5.2 shows the plots of each identified group. From these plots it is clear that the algorithm identifies groups of observations with very small variability, which therefore are extremely aligned to the corresponding components. Each of these groups is homogeneous in the sense of the Forward Search,

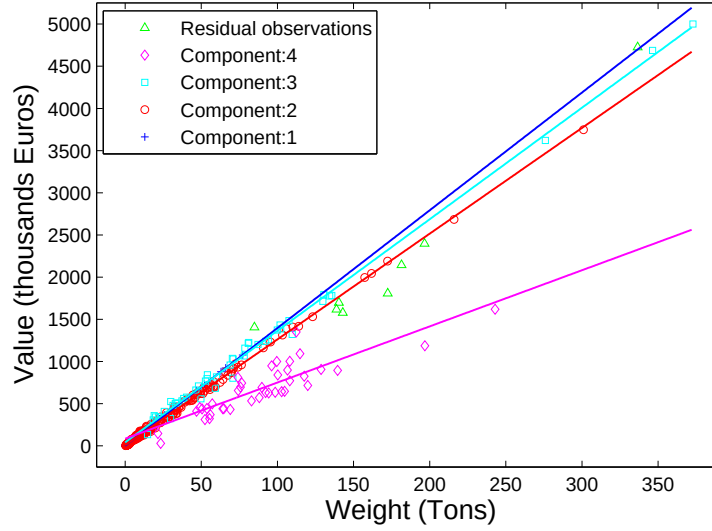


Figure 5.1: Fishery data (677 observations): the four tentative groups with the associated linear fits and the unassigned observations (green triangles) detected by the algorithm 5.1. The components are formed by 310, 274, 46, 39 observations (from top left to bottom right). The 8 unassigned observations may be outliers or may have to be reassigned.

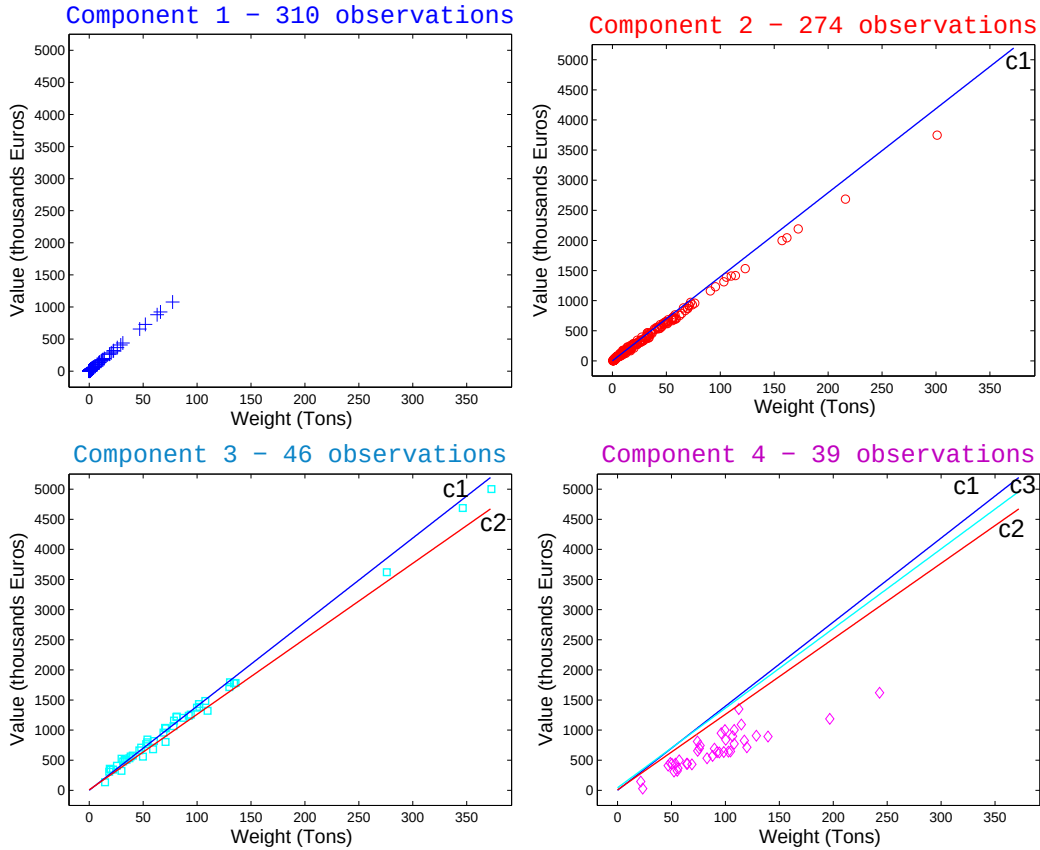


Figure 5.2: Fishery data: the four tentative groups are highlighted in separate plots. Note that for the components identified from the second step, we have superimposed the line(s) fit on the component(s) detected in the previous steps.

i.e. the residuals of the associated observations from the corresponding mixture component are normal.

In general it is not possible or there is no unique way to formally determine whether this is the “correct” clustering or, in other words, if the observations in each group come from the “right” mixture model. Probably, in a real world application context like the one under discussion, it is much more reasonable to ask oneself if the clustering is “good”, relative to a subjective judgement that can only rely on the final aim of the clustering exercise. If the goal of the researcher is to detect homogeneous groups, even if they have very small variance, these mixture components are certainly a good result. On the contrary, if we leave the judgement to investigators who check the dataset for macroscopic anomalies in view of detecting frauds, it is likely that they would judge significant the difference between component 4 and the others but they would prefer to see groups 1, 2 and 3 merged together into a unique larger groups with a single regression component fit on them.

In other terms, to obtain “better” groups it may be convenient to allow for higher within-group variance. In order to achieve this task, in the next two Sections we propose to relax our algorithm 5.1 in two natural ways.

Before addressing the relaxed algorithms, a remark: we implemented and tested a strategy similar to 5.1 in the multivariate context, where the Mahalanobis distance is monitored. In this case the strategy produces bad results because of the presence of high dense and overlapping areas ([PRT09]).

Adaptation through re-weighting the Forward Search

Each homogeneous group is identified by the Forward Search with a specific test size of 1%, as stressed in Section 3.6. We then proposed Strategy 3.4 to extend the use of the Forward Search to any pre-specified confidence level. It is therefore natural to modify Strategy 5.1 along the lines of Strategy 3.4 to α values smaller than 0.01, so that to relax the homogeneity requirement and hopefully merge components 1, 2 and 3. This idea can be formalised in the following algorithm:

Strategy 5.2 Tentative allocation of mixture components by re-weighted Forward Search.

1. Let $n^*(j)$ be the size of the sample to be analysed at iteration j . At the first iteration $n^*(1) = n$.
2. Choose a significance level α .

3. Apply Strategy 3.4 to these $n^*(j)$ observations and obtain the subset of $m^*(j) < n^*(j)$ observations that are homogeneous at the desired level α .
4. The mixture component c_j is the fit on the homogeneous group of $m^*(j)$ observations.
5. Trim the homogeneous group j identified in step 3 and return to step 1 with a reduced sample size, by setting $n^*(j+1) = n^*(j) - m^*(j)$.

The application of this algorithm to the Fishery dataset with different levels of α , 0.0001%, 1% and 20%, gives the results in Figure 5.3. Visually, the result for the small significance value $\alpha = 0.0001\%$ is acceptable. Coherently, for a given step, bigger α values produce smaller groups. However, the results are not of easy interpretation. A common aspect is that

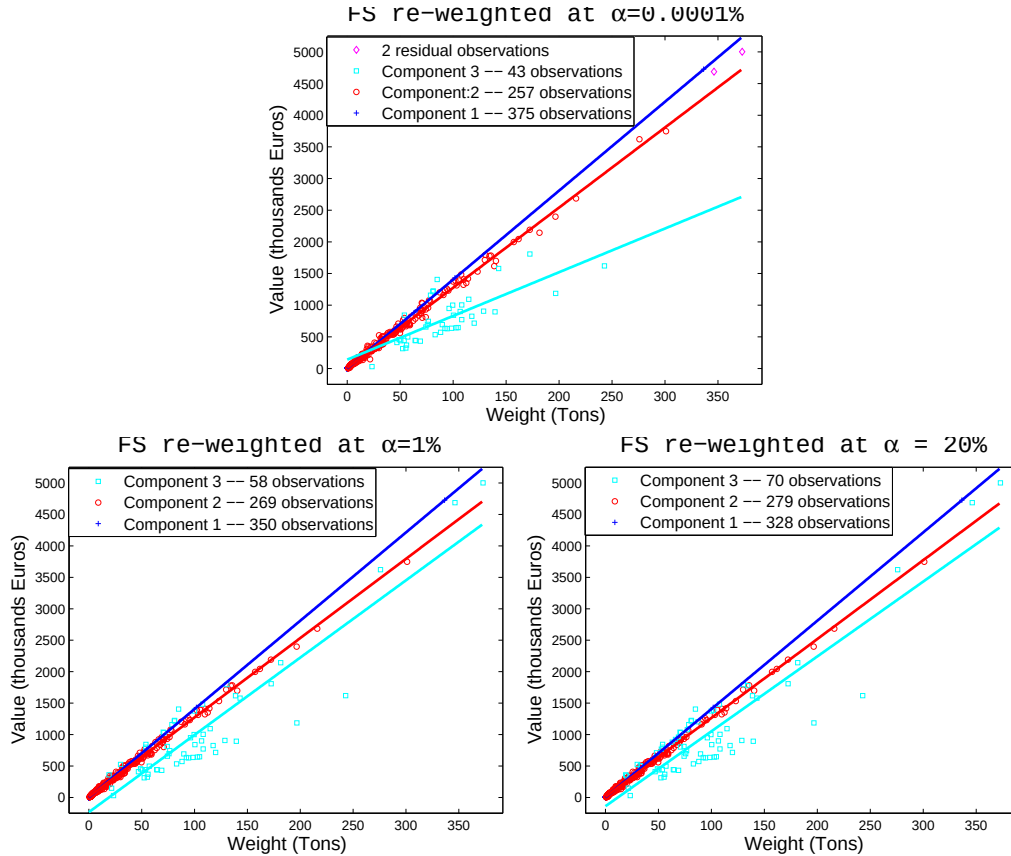


Figure 5.3: Fishery data: tentative linear components estimated by the Forward Search re-weighted algorithm 5.2 for different levels of α : 0.0001, 1% and 20%.

the re-weighted algorithm detects one component less than the standard algorithm for any α value. Moreover for α equal to 1 and 20% (two bottom panels of Figure 5.3) the estimated mixtures are very similar. This counter-intuitive behaviour could be explained by the presence of homogeneous subgroups characterised by the perfect-fit problem where, as said in Section 2.3, slightly deviating observations may have extremely high Studentised deletion residuals. Therefore reducing the size can only partially help to find components with higher within-group variability. Note also that the Forward Search re-weighted at 1%, i.e. the standard significance level of the Forward Search, produces results which are not equivalent with those of the standard Forward Search (Figure 5.1). This is also a counter-intuitive result that will be considered when we will study the formal properties of the re-weighted Forward Search strategy.

Adaptation through R^2 control

A second natural way to relax the basic algorithm 5.1 is through the control of the value of the linear correlation coefficient

$$R^2(x, y) = \frac{\hat{y}'\hat{y}}{y'y}$$

where $\hat{y}'\hat{y}$ is the regression deviation and $y'y$ is the total deviation.

The idea is to run the Forward Search only from steps where the R^2 computed for the subset is smaller than a convenient threshold which is indication of strong alignment, say $R_{th}^2 = 0.95$. In other words we leave the subset to grow until R^2 keeps high values, independently from the values of the minimum Studentized deletion residuals. We first detail the methodology for a generic sample on size n and dimension p (Strategy 5.3) and then we use it for the purpose of mixture estimation (Strategy 5.4).

Strategy 5.3 Forward Search relaxed through R^2 monitoring.

1. Choose a threshold R_{th}^2 for R^2 (for example $R_{th}^2 = 0.95$).
2. Run the Forward Search from step $3(3p + 1)$ to monitor the R^2 and minimum Studentized deletion residual.
3. Let $i_{>Bth}$ be the first step where the minimum Studentized deletion residual exceeds the value of the 99.99% envelope at the last step, the so called Bonferroni threshold (see Figure 3.4).
4. Let $i_{<R^2}$ be the first step where $R^2 \leq R_{th}^2$.

5. **if** $i_{<R^2} = \emptyset$ **or** $i_{>Bth} > n - 3$
 - 5a **then** the outliers are the units that are not included in the subset $i_{>Bth} - 1$, i.e. the step prior to the one in which there was the first exceedance of Bonferroni threshold.
 - 5b **else** the outliers are those found by a run of the Forward Search from step $i_{>Bth} - 2$, using the standard strategy 3.3.

Note that this algorithm typically requires two additional monitorings, the first associated with the R^2 and deletion statistics (Step 2) and the second associated with the determination of the subset of homogeneous units by superimposing the envelopes (Step 5b). Also note that the strategy starts monitoring the R^2 and the minimum deletion residual from step $3(3p + 1)$ rather than from the standard $(3p + 1)$ step (see Section 7). This is because at the very beginning of the search there may be minimum deletion residual values well above the Bonferroni threshold and, because of the strategy step 3, this would result in declaring as outliers all observations subsequent to the spurious peak.

Now, we have the tool to relax the tentative allocation of mixture components as follows:

Strategy 5.4 Tentative allocation of mixture components through R^2 monitoring.

1. Let $n^*(j)$ be the size of the sample to be analysed at iteration j . At the first iteration $n^*(1) = n$.
2. Choose a threshold for R^2 , typically $R_{th}^2 = 0.95$.
3. Apply Strategy 5.3 to these $n^*(j)$ observations and obtain the subset of $m^*(j) < n^*(j)$ observations that are homogeneous unless $R^2 > R_{th}^2$.
4. The mixture component c_j is the fit on the homogeneous group of $m^*(j)$ observations.
5. Trim the homogeneous group j identified in step 3 and return to step 1 with a reduced sample size, by setting $n^*(j + 1) = n^*(j) - m^*(j)$.

With the threshold $R_{th}^2 = 0.95$ the algorithm 5.4 on the Fishery data gives the results in Figure 5.4. The method has now detected only two components. Note that one corresponds to the lowest pink component of Figure 5.1 and the other includes the higher blue, cyan and red components of the same Figure 5.1. From a practical point of view the algorithm has produced

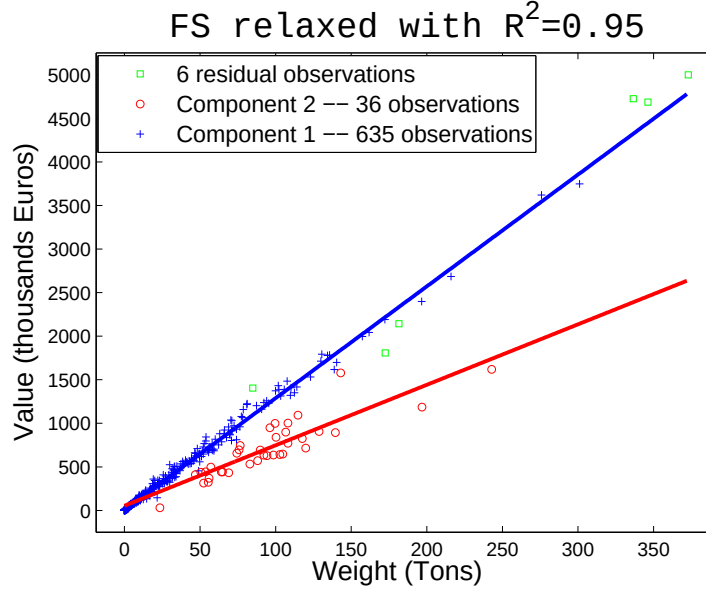


Figure 5.4: Fishery data (677 observations): two linear components detected by algorithm 5.4 relaxed by controlling the R^2 . They are composed respectively by 635 and 36 observations. 6 observations remain unassigned.

very sensible results. In fact the second (red) component is formed almost exclusively by import declarations of a particular Member State which, as seen in Section 2.4, under-declared systematically the price of the fishery product.

5.3 Confirmation of groups and outliers

The groups identified by the algorithm 5.1 and its modified versions 5.2 and 5.4 are only tentative, and so are the corresponding least square regression fits. It is desirable to have a procedure that validates the tentative groups by eventually reassigning observations to groups. In fact it can happen that an observation tentatively assigned to a component at some iteration step, becomes closer to the fit of another component detected at later steps. We propose to address this problem with the following algorithm that reassigns the observations to the tentative components from which they are closer in term of raw residual, and then refits the regression lines. Formally:

Strategy 5.5 Confirmation of the mixture components.

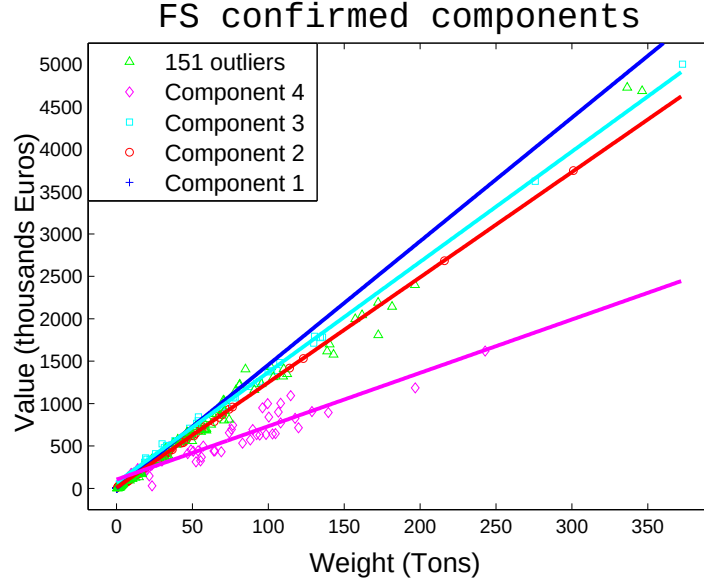


Figure 5.5: Fishery data: the 677 observations have been reallocated from the four tentative components of Figure 5.1 to these confirmed components. They are composed respectively by 291, 139, 49, 39 observations and there are 159 outliers.

1. Let $c_1, \dots, c_J$ be the tentative components identified with one of the three strategies of Section 5.2 in a dataset S_n of n observations.
2. For each observation i in the dataset compute the raw residuals $e_{i,j}$ from each regression line c_j , $j = 1, \dots, J$.
3. Partition S_n in J groups $g_1, \dots, g_J$ such that observation $i \in g_k$ if $e_{i,k} = \min_{j=1, \dots, J} (e_{i,j})$.
4. Apply the Forward Search to each partition part g_j . Let $g'_1, \dots, g'_J$ be the J groups of homogeneous observations identified.
5. The regression lines $c'_1, \dots, c'_J$ fitted on the homogeneous groups $g'_1, \dots, g'_J$ are the estimated mixture components. The outliers are the observations in $S_n \setminus \bigcup_{j=1, \dots, J} g'_j$.

Of course, the term “partition” is used here as in set theory, i.e. as a division of the set of observations S_n into non-overlapping and non-empty parts that cover all S_n . By applying this confirmation procedure 5.5 to the components obtained for the fishery dataset by the basic tentative allocation algorithm

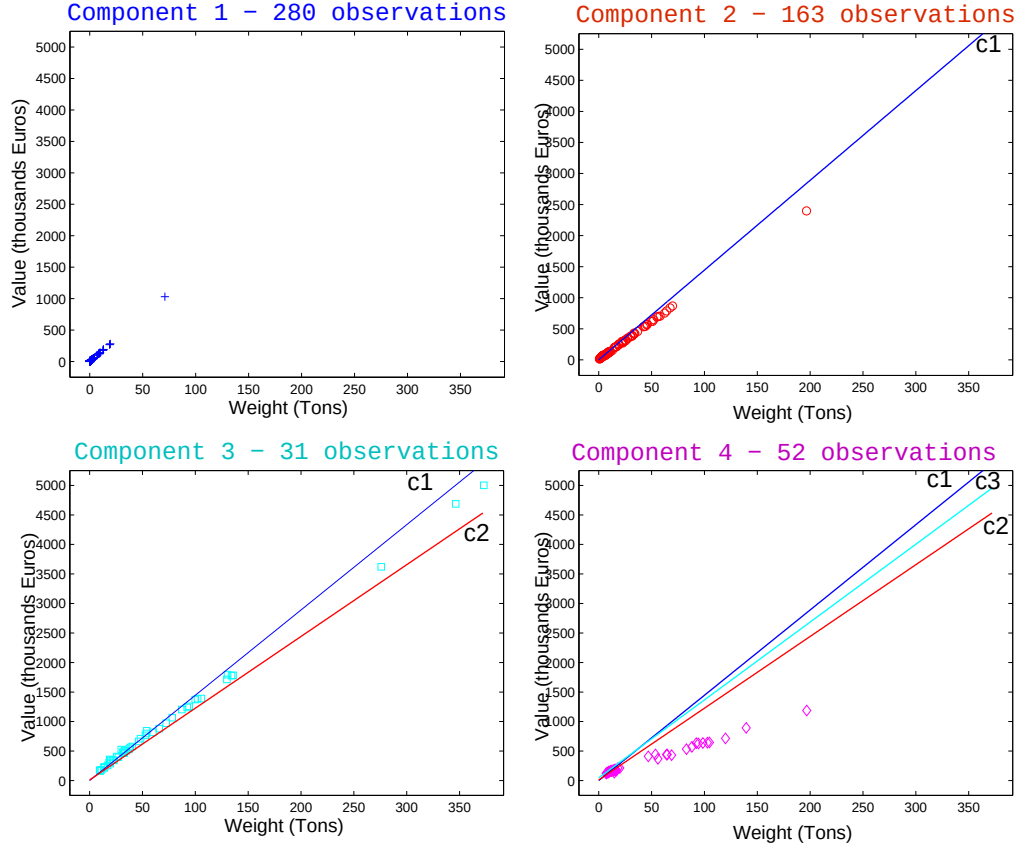


Figure 5.6: Fishery data: the four final components reallocated by the confirmation algorithm 5.5. Note that for the components identified from the second step, we have superimposed the line(s) fit on the component(s) detected in the previous steps.

5.1, we obtain the results in Figures 5.5 and 5.6. The final estimated mixture looks reasonable. Note the considerable number of observations declared as outliers. Most of them seem concentrated in proximity of the dense overlapping area. This can be explained by the fact that in that area there is a very low within-group variability and the chance for an observation to have significant deletion residual is higher.

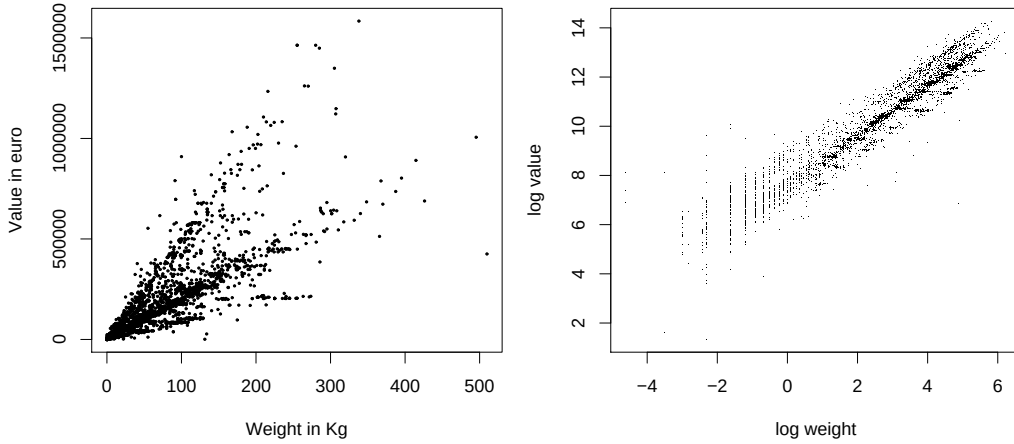


Figure 5.7: On the left the trade dataset of Section 2.2 and on the right its logarithmic transformation. The 4719 import declarations follow different linear populations that are clearly overlapping. In the logarithmic scale the clusters are more separated than in the original scale.

5.4 Experimenting different mixture solutions on a complex dataset

This Section introduces an experimental comparison of the new Forward Search mixture solution with other three mixture estimation methods. We have applied the methods on the trade dataset which opened the dissertation in Section 2.2, which is bigger and more complex than the Fishery dataset used so far to illustrate the concepts of the Forward Search approach.

We propose again the dataset in the left scatterplot of Figure 5.7 and recall the features that are relevant for the discussion. The 4719 observations (import declarations of a technological product) appear roughly distributed along three lines departing from the origin of the coordinate axes, where they are extremely concentrated and overlapping. One of the three groups extends mostly over the upper part of the plot (upper group), a second group over the central part (central group) and a third over the lower part (lower group).

More than 50% of the observations (precisely 2669) are below 30 Kg weight and cover only 10% of the total trade volume considered in this dataset.

The presence of the dense and overlapping area complicates the statisti-

cal analysis. In fact, different methods can produce very different mixture estimates and clusters depending on whether the data are separated into well distinct groups or not. For this reason we will assess the methods also on the data transformed in the logarithmic scale for both variables, which is represented in the right panel of Figure 5.7. In the logarithmic scale the original upper, central and lower groups become distributed over three parallel and rather separated groups of observations. Besides, the dense area now expands over a larger area along the central group.

The dataset contains information that we have not used to estimate the mixtures and to identify the clusters, e.g. the country of the exporter/importer and a masked identifier of the importer. We will use this additional information to validate the quality of our statistical analyses, by subjectively looking over meaningful structures (patterns) in the generated clusters.

The literature on mixture estimation is wide, even if we remain in the domain of linear regression mixtures. We have decided to experiment three methods that are comparable with our Forward Search solution in the way the concepts of cluster homogeneity and separability are conceived. The methods are the Clusterwise Linear Regression of DeSarbo and Cron [DC88], the Trimmed Likelihood estimator of Neykov *et al.* [NFDN07] and the Fixed Point Clustering of Hennig [Hen03].

The three methods are based on maximum likelihood estimates of the parameters of the regression lines in the mixture. Following the notation of [NFDN07], we assume that our sample $S(n)$ of n observations (y_i, x_i) is independent and identically distributed, and that y_i is generated from a mixture of g distributions ψ_j in proportions π_j given by

$$\varphi(y_i|x_i, \Psi) = \sum_{j=1}^g \pi_j \psi_j(y_i|x_i, \theta_j) \quad (5.1)$$

In our context the ψ_j are linear distributions, but more in general other reference distributions can be adopted. $\Psi = (\pi_1, \dots, \pi_{g-1}, \theta_1, \dots, \theta_g)$ is the unknown parameter vector which maximizes the log likelihood:

$$\log L(\Psi) = \sum_{i=1}^n \log \left\{ \sum_{j=1}^g \pi_j \psi_j(y_i|x_i, \theta_j) \right\} = \sum_{i=1}^n \sum_{j=1}^g \log \pi_j + \log \psi_j(y_i|x_i, \theta_j) \quad (5.2)$$

If we want to associate each observation to some mixture component, we need an extra variable z_{ij} which should be 1 if y_i belongs to the component j

and 0 otherwise. Then the objective is to maximize the new log-likelihood:

$$\log L(\Psi) = \sum_{i=1}^n \sum_{j=1}^g z_{ij} \{\log \pi_j + \log \psi_j(y_i|x_i, \theta_j)\} \quad (5.3)$$

This maximum likelihood estimate problem is complex, but can be decomposed in simpler estimation problems using the expectation-maximization (EM) algorithm. Here we skip the details of this general EM approach: they can be found, e.g., in [MP00] or clearly sketched in [NFDN07]. In the following, we will summarise the distinguishing features of the three methods and then we will present the linear components detected by each method on the import technological data discussed in Section 2.2.

Clusterwise Linear Regression maximises (5.3) for a given number of components g . The method fits the whole dataset to the distributions ψ_j of the mixture 5.1. Therefore, all observations are assigned to some component.

The idea of *Trimmed Likelihood estimator* is to identify the subset of k observations with the largest likelihood. The trimming is implicit in the fact that the fit is done on the subset of (homogeneous) observations that are more likely to occur if the assumed model is true, while the other (unlikely) observations are excluded. In this sense, Trimmed Likelihood estimator considers the observations excluded from a component as outliers for that component. The problem is combinatorial: k is of course unknown and all possible $\binom{n}{k}$ data combinations should be fitted. Thus, approximate solutions are found using an adaptation of the FAST-TLE algorithm of [NM02]. The algorithm alternates a number of times a trial and a refinement steps based on random sampling of the data, which can be implemented in a reasonably easy manner using the R package **FlexMix** of [Lei04].

Both Clusterwise Linear Regression and Trimmed Likelihood Estimator assume that the data are formed by clusters of the same parametric form, namely linear regression distributions, and therefore they fit the whole dataset to the distributions Ψ_j of the mixture 5.1.

The key idea of *Fixed Point Clustering* is to relax this assumption by searching for a single subset of homogeneous observations at a time, under the assumption that the rest of the data consists of outliers separated from the current subset. A Fixed Point Clustering cluster will be of the form

$$(1 - \epsilon)P_0 + \epsilon P^*, \quad 0 < \epsilon \leq 1, \quad (5.4)$$

where P_0 is the reference distribution for a Fixed Point Clustering cluster and P^* arbitrarily defines the rest of the data. In our context P^* can be

the mixture of distributions given in 5.1. ϵ is an important constant which sets the level of separation of a Fixed Point Clustering cluster from the rest. A Fixed Point Clustering cluster is therefore more determined by the local properties of the observations in the cluster, and not by a global mixture model such as 5.1. In this sense, the concept of cluster in Fixed Point Clustering and in the Forward Search are similar. Another communality between Fixed Point Clustering and the Forward Search is that both can be used as mixture estimation methods but, as Hennig, the inventor of the Fixed Point Clustering, claims for his approach, they are very suitable to be used as exploratory analysis tools. The details of Fixed Point Clustering can be found in [Hen98] and [Hen03].

In the next Sections we discuss the mixture solutions obtained with the four methods on the dataset and its logarithmic transformation.

5.5 Testing the new Forward Search solution

We start with two remarks that are common to the experiments of this Section.

In the Figures that illustrate the estimated mixtures, the number that identifies each component also indicates the detection order.

As already stressed, the forward progression starts with an initial LTS or LMS regression estimate, which is based on a number of random subsets of the data (Sections 3.5 and 4.2). Normally the Forward Search is very robust to the choice of the initial subset and therefore the random choice is limited to 1000 sub-samples. However, for the dataset in question, the relatively big size and presence of highly overlapping areas are such that different starting points chosen on the basis of 1000 random subsets often give rise to different results. Therefore, in order to obtain more stable components, we have chosen the starting LTS/LMS subset among 10000 random sub-samples.

Tentative components in original-scale data

We start with the application of the basic mixture component detection algorithm (5.1) (see Section 5.2) on the original data. This produces the nine mixture components shown in Figure 5.8.

- Components 1, 2, 3, 5 and 7 (yellow circles) partially overlap and capture well most of the central group of observations.

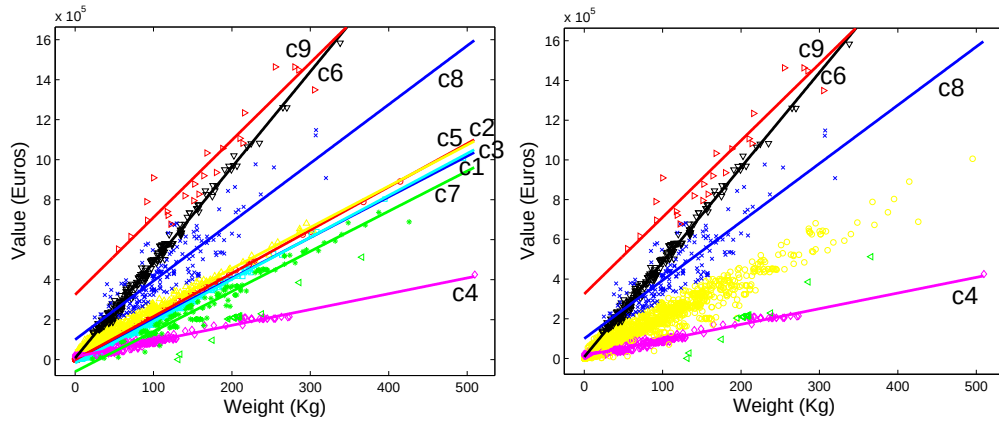


Figure 5.8: Application of the Forward Search mixture strategy 5.1 (standard tentative components). On the left the 9 tentative components detected in the original scale data. On the right the components 1, 2, 3, 5 and 7 have been merged into a unique component (yellow $\circ$). Composition:

Comp.	1	2	3	4	5	6	7	8	9
Obs.	1175	1691	432	353	453	243	152	175	25
Color	blue	red	cyan	pink	yellow	black	green	nblue	red
Symbol	+	$\circ$	$\square$	$\diamond$	$\triangle$	∇	$\diamond$	$\times$	$\triangleright$

- Component 8 (navy blue $\times$) extends over the central-upper group, but it is much more dispersed.
- Components 6 and 9 (black ∇ and red $\triangleright$) cover most of the upper group.
- Finally, components 4 (pink $\diamond$) covers rather well the observations of the lower group.
- The residual observations (green $\triangleleft$) that are not assigned by the Forward Search to any cluster, are dispersed mostly in the lower part of the plot. We will see how the confirmation procedure will reassign them to the existing tentative clusters.

Table 5.1 gives a few basic statistics for the components. Since more than half of the dataset is made up by declarations below 30 Kg weight, the average price (value/weight) is computed excluding such small observations. This is to concentrate on the trade of bigger operational interest but also to avoid the effect of rounding in the small declarations of few Kilograms, that under certain conditions are contemplated by the Customs data collection procedures. Note that when we regress values over weights, rounding the

comp	average price	#obs W > 30	#obs	R ²	slope	intercept
1	2038	68	1175	0.999	2035.3	80.2
2	2138	447	1691	0.996	2161.9	-684.2
3	1856	314	432	0.997	2095.7	-17129.3
4	934	288	353	0.953	788.7	14065.9
5	2728	326	453	0.972	2059.8	43618.7
6	4858	235	243	0.991	4775.6	7524.4
7	1570	150	152	0.947	2005.4	-59830.5
8	3977	174	175	0.757	2939.6	100058
9	6260	25	25	0.900	3865.6	324721.8

Table 5.1: Technological product dataset: statistics on the tentative components detected with the Forward Search strategy 5.1. The statistics are computed excluding declarations of weight smaller than 30 Kg.

weight of small declarations is in general not a problem, as this affects observations of low leverage, but this may produce very inaccurate prices.

The average price estimates confirm that, for operational purposes, components 1, 2, 3, 5 and 7 can be merged into a single group. As expected, the first three components are highly homogeneous and capture observations densely concentrated. The corresponding estimate of the squared correlation coefficient R^2 is very close to 1. The progression of this estimate during the searches is shown in Figure 5.9.

We may get insights from the mixture components by exploring the trade dynamic for the product in question, e.g. by focusing on the country of origin of the goods. The “cheapest” goods, those in component 4, originate mainly from a single third country (Malaysia, MY). For anti-fraud purposes one may also remark that they are imported mostly by a single company. The most expensive goods (components 6 and 9) have origin mainly in Korea (KR). The biggest exporting country is China (CN), which mainly occurs in the central components (1, 2, 3, 4 and 6). Sometimes there are unexpected patterns in the components. For example, by looking at the date of the transactions in component 4, we note that gradually the price of imports from MY raised from a lower to a higher value. Such patterns are difficult to detect using alternative techniques.

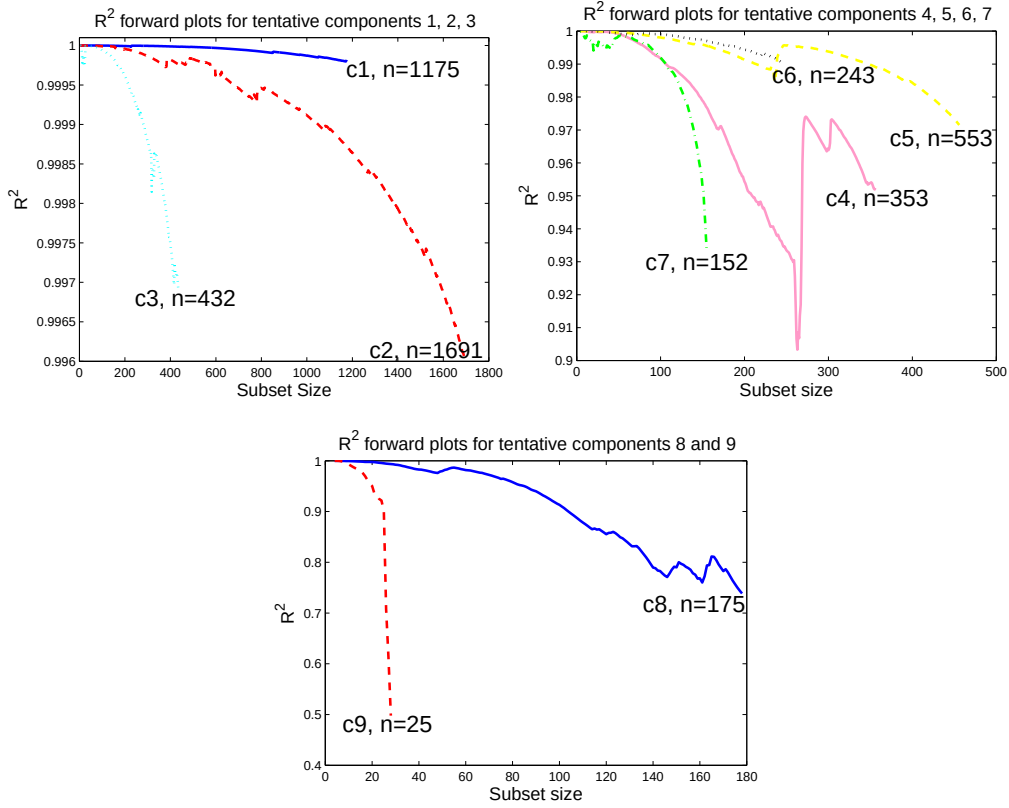


Figure 5.9: Forward plots of the estimates of R^2 for components 1, 2, and 3 (top-left plot), for components 4, 5, 6 and 7 (top right plot) and for components 8 and 9 (bottom plot).

Tentative components in the logarithmic-scale data

Figure 5.10 shows the results of applying the procedure to the data after log transformation of values and weights, which reproduces a scenario with greater separation between the groups of observations. In this case the components found by the Forward Search are only six. To simplify the interpretation of the results, we also plot in Figure 5.10 the representation of the six clusters in the original scale. We can see that:

- The central components 1 (blue +) and 4 (pink $\diamond$) correspond quite well to the group formed by merging the components 1, 2, 3, 5 and 7 previously found in the original scale.
- The observations that in the original scale were assigned to the almost overlapping components 6 and 9, in the log scale are split in two well

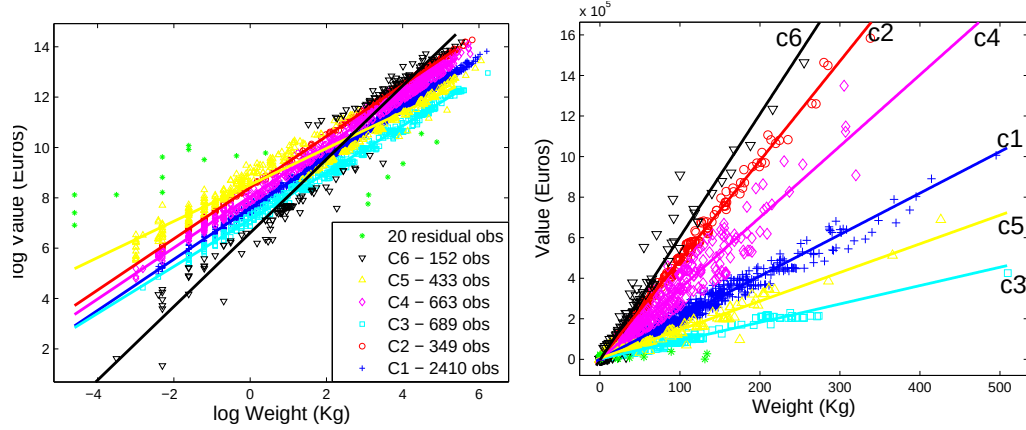


Figure 5.10: Application of the Forward Search mixture strategy 5.1 (standard tentative components). Components are detected in the log-transformed data (plot on the left) and the results are also represented in the original scale (plot on the right).

distinct components, 2 and 6.

- The 20 residual observations in the log-scale appear to be clear outliers. However their operational interest is low, being the quantities and values involved small.

Final components in the log-transformed data

In this Section we show what we achieve with our confirmation strategy 5.5. We apply the confirmatory analysis only to the tentative components detected in the logarithmic scale (Figure 5.10), as they are better defined than those identified in the original scale (Figure 5.8). Of course the same analysis can be done for the other case.

In Figure 5.11 the six final components identified in the log transformed data (left plot) are also represented in the original scale (right plot). By inspecting the plots and the composition of the six components shown in the plot legend, we can conclude that the confirmatory procedure 5.5 has redistributed the observations into components that are less dispersed than those detected by the tentative procedure 5.1.

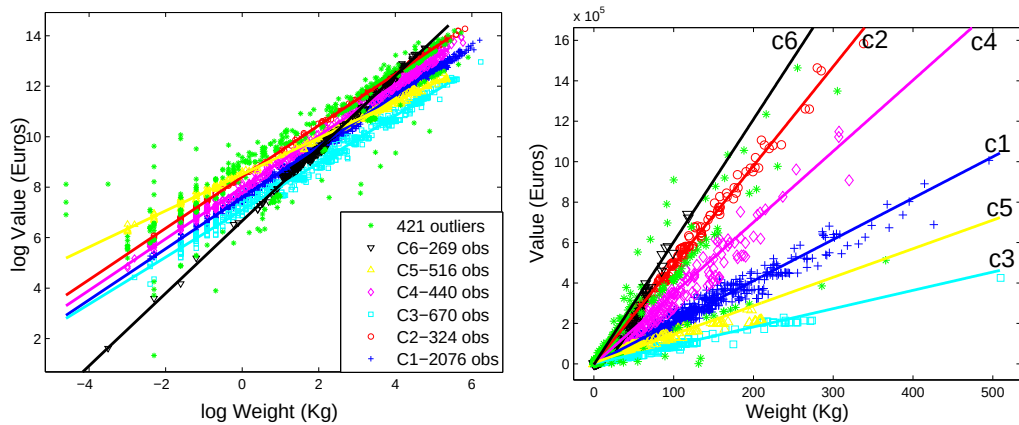


Figure 5.11: Application of the confirmatory procedure 5.5 (final components). The components are detected in the log-transformed data (plot on the left) and the results are also represented in the original scale (plot on the right).

Tentative and final components with the R^2 -relaxed strategy

Figure 5.12 shows the results obtained with the R^2 -relaxed algorithm 5.3 on the dataset in the original scale.

The tentative components here look much more homogeneous than those detected by the standard algorithm 5.1 (Figure 5.8). In this sense, these results strengthen our intuition that the R^2 -relaxed algorithm 5.1 can lead to better results when data are highly linearly correlated, in particular if populations overlap. Nevertheless, the results of Figure 5.12 are not completely satisfactory. In fact, it is not clear if components 4 (pink), 5 (yellow) and 6 (black) are necessary or are just spurious. The impression is that still too many components have been identified.

Components driven by the expert opinion

All the algorithms discussed up to now are fully automatic, i.e. the subjective judgement of the analyst or the subject matter expert is not taken into account. However, we might want to consider the expert opinion to improve the estimation of the final components. This can be done before applying the confirmatory procedure, by accepting or rejecting the appropriateness of one or more tentative components. In practice, it is sufficient to intervene at step 1 of the confirmatory procedure 5.5 by removing the undesired

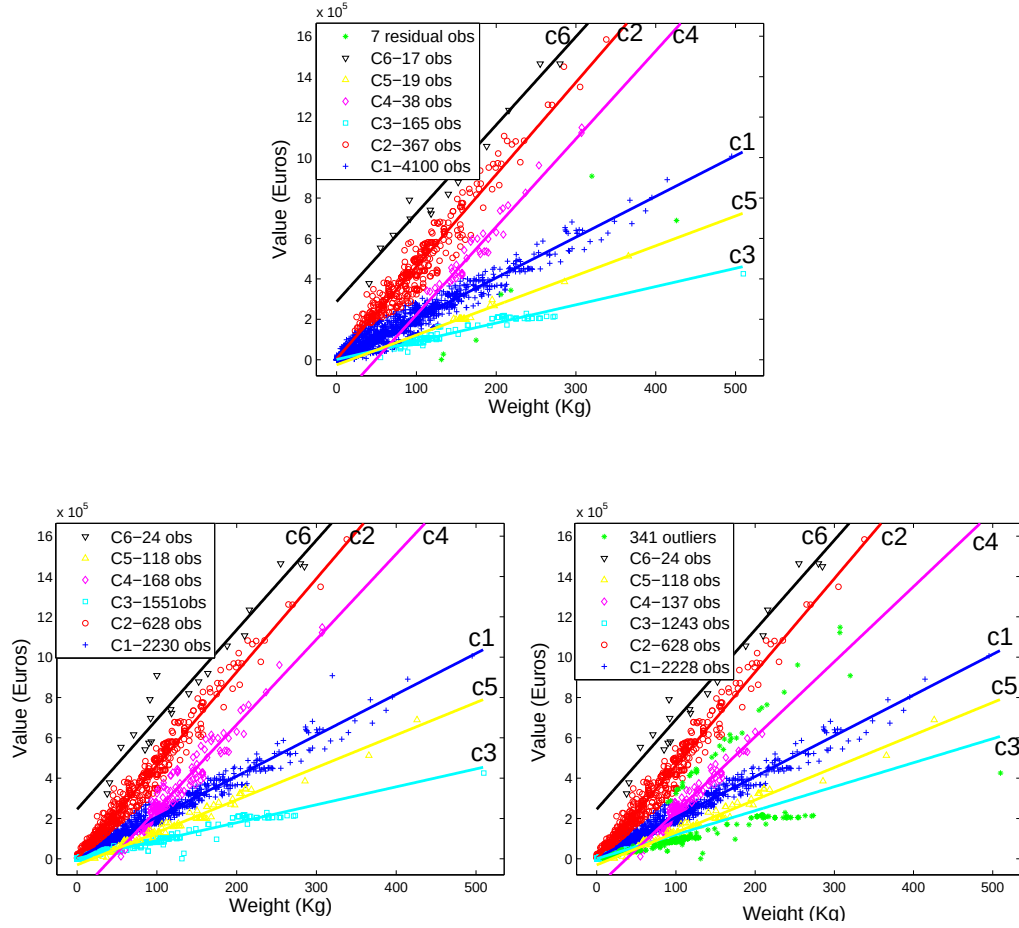


Figure 5.12: Application of the R^2 -relaxed algorithm 5.3. Tentative components (top), partition (bottom left) and final components (bottom right) detected on the technological trade data in the original scale.

components, say $\tilde{c}_i$ for some indexes $i \in \{1, \dots, J\}$, and proceed with partitioning and redistributing the observations to the remaining components, which will become final.

As an example, if we believe that the tentative components c3, c4 and c5 in Figure 5.12 are unreliable and apply the partitioning and confirmation steps only based on c1, c2 and c3, we would obtain the final components that are plot in Figure 5.13. The distribution of the observations among the final components and the outliers appear now very reasonable.

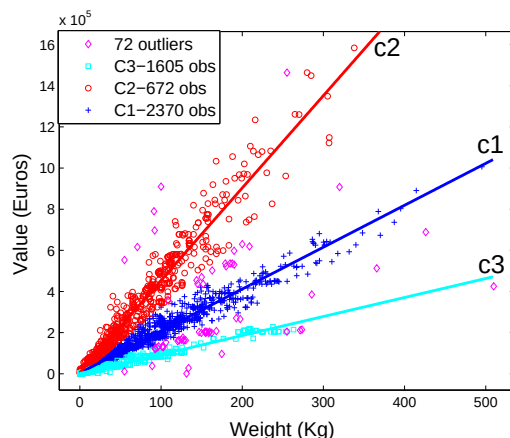


Figure 5.13: Components driven by the expert opinion. Three of the six tentative components detected by the R^2 -relaxed algorithm 5.3 (Figure 5.12) are judged inappropriate and removed before the confirmation step.

5.6 Testing the Clusterwise Linear Regression solution

To run Clusterwise linear regression on our trade data we used the R function `regmix` which is part of the `fpc` library written by Hennig. The function requires indication of the maximum number of expected components and the stopping criterion is decided on the basis of the Bayesian Information Criterion (BIC). We tried first by largely overestimating the number of expected components: 30.

With this choice, for the log transformed data BIC suggested 10 mixture components, while for the original data the BIC value was monotonically decreasing and produced 30 components. Therefore, we decided to force the method at fitting a maximum of 10 components. Figure 5.14 reports the results, which are not easy to interpret and compare them with those produced by the Forward Search.

Other runs to fit a smaller number of components to the data would be necessary. However this try and error approach in general is not reasonable, especially if the datasets to analyse are many.

A second drawback of the method is that, as already remarked, by construction there are no outliers: all observations are assigned to some component.

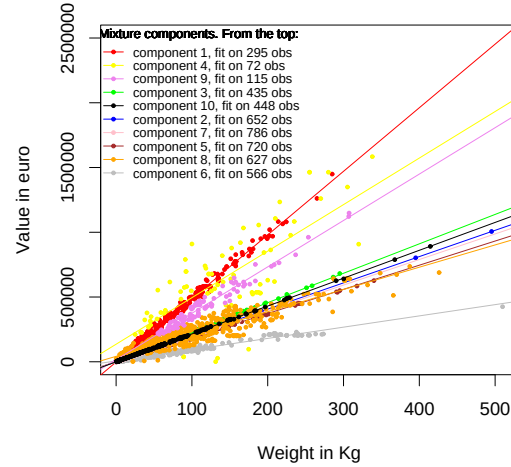


Figure 5.14: Clusterwise Linear Regression applied to the original data.

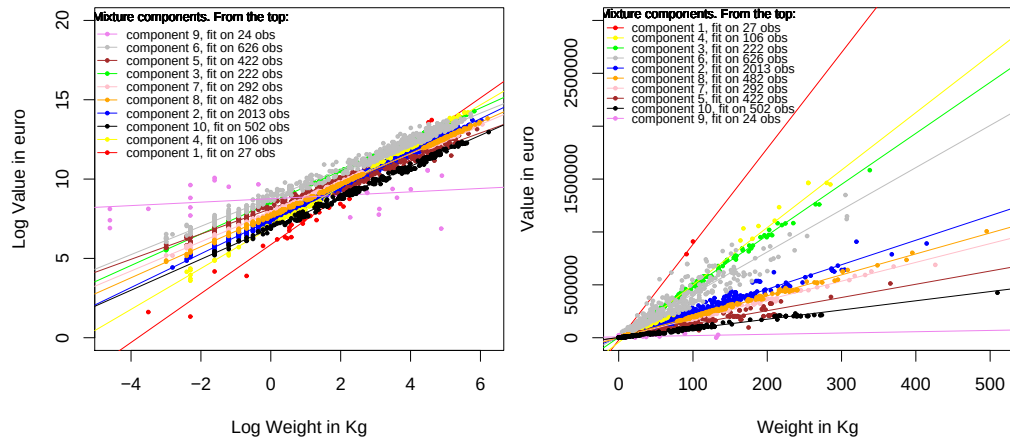


Figure 5.15: Clusterwise Linear Regression applied to the log-transformed data (left) and results represented in the original scale (right).

5.7 Testing the Trimmed Likelihood Estimator solution

We have run Trimmed Likelihood Estimator using R routines for `FlexMix`, kindly provided to us by Neykov, Neytchev and Filzmoser. We fixed the most relevant parameters of the method as follows: trimming percentage, $k_p = 0.2$; random sample size in the trial step, $k_s = 0.1$; number of FAST-TLE repetitions, $nit = 200$. We run the Trimmed Likelihood Estimator also with other parameters but the results were worse or at most similar.

Trimmed Likelihood Estimator is also used in combination with the BIC to determine the number of mixture components. Trimmed Likelihood Estimator gave three well separated and sensible components in the log-transformed data, similar to some of those given by the Forward Search. A large number of outliers (943) positioned between the components were left out, some of which appear to be distributed quite linearly. The method does not have a natural confirmation procedure to check this impression.

In the data in the original scale, the three components found only capture the central highly dense area.

5.8 Testing the Fixed Point Clustering solution

Fixed Point Clustering was run using the R function `fixreg` that is part of the `fpc` library written by Hennig. The indication of the number of expected clusters is not required. In addition, the library implements a criterion to estimate the similarity among Fixed Point clusters and to group similar clusters together in *stable groups*. Then a *representative Fixed Point cluster* is selected for each stable group.

From a theoretical point of view, stable groups are Fixed Point Clusters themselves and should be visually connected. However, surprisingly, in this case they are not, as we can see in the right plot of Figure 5.18. The reason is that stable groups are based on a number of points in the intersection of Fixed Point clusters. However in our case the intersection is totally dominated by the clump in the lower left area of the plot. Therefore, Fixed Point clusters that may look quite different on the plot are actually merged together in the same stable group which, therefore, looks disconnected¹.

¹ I thank Christian Hennig for the comments on this peculiar dataset and results.

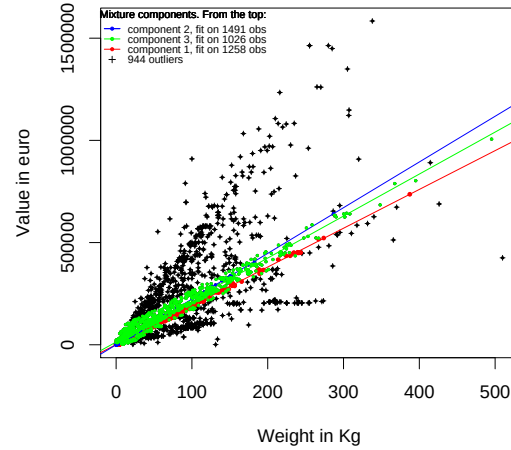


Figure 5.16: Trimmed Likelihood Estimator applied to the original data.

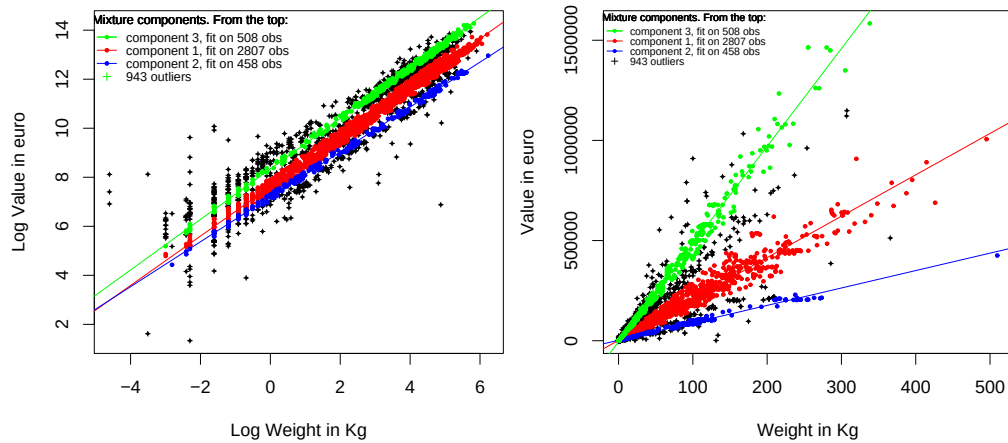


Figure 5.17: Trimmed Likelihood estimator applied to the log-transformed data (left) and results represented in the original scale (right).

On the contrary the Fixed Point clusters that are selected by the algorithm as representative from each stable group are well separated, as shown in the left plot of Figure 5.18. However, such representative Fixed Point clusters result very “thin”, with very small within-group variability. They are not well delimited and a large number (366) of unassigned observations are left out.

In Figure 5.19 we also plot the results obtained on the log-transformed-data (left) together with their representation in the original scale (right). The results are not reasonable: only a big representative component was found and an even bigger number of observations (1970) remain unassigned.

5.9 Towards a rigorous assessment of mixture estimation methods

The Forward Search and the other mixture estimation methods discussed in the Chapter produce different data clusterings. First of all, it should be said that our empirical comparison of such clusterings relies on the fact that each method produces a different number of components. The comparison of the four methods should be also repeated by assuming that the number of components is known, e.g. by choosing the number of components suggested by the Forward Search. This is in our research agenda.

This Section is dedicated on how to assess “objectively” different clusterings, which is a problem that we have not considered so far. We will just introduce this difficult issue and propose directions for future work.

Perhaps the most important notion for cluster validation is *stability*. The idea behind stability is that good clusterings should not vary much from one sample to another if the samples differ for a reasonable number of observations. For example, Hennig [Hen07] proposed to measure cluster stability for a given dataset and set of clustering algorithms by:

1. Generating new datasets from the original one, e.g. by bootstrapping.
2. Apply the clustering method to the new datasets.
3. Compute an appropriate statistic to measure the similarity of the new clusterings compared to the one on the original dataset. High similarity would be indication of stable clustering.

The statistic proposed by Hennig is the Jaccard coefficient, which is a standard similarity measure between sets.

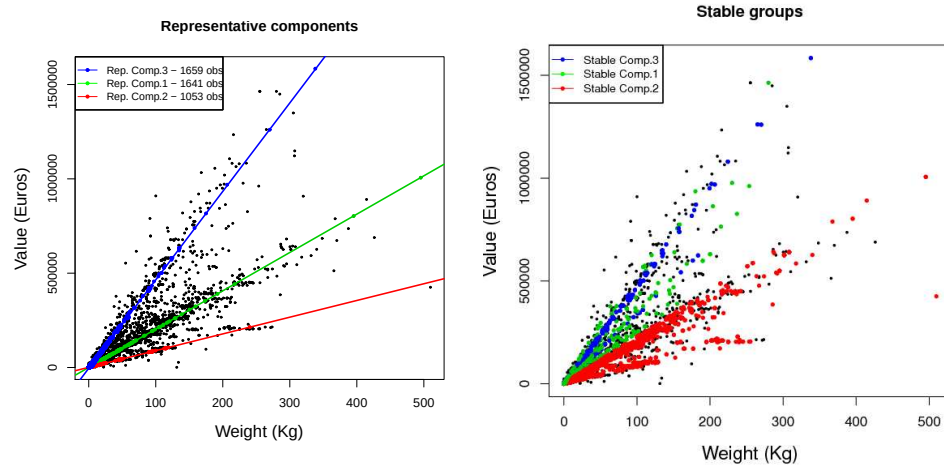


Figure 5.18: Fixed Point Clustering applied in the original scale. One stable group, on the right, looks disconnected. The representative components of each stable group, on the left, are composed respectively by 1641, 1053 and 1659 observations. 366 observations remain unassigned.

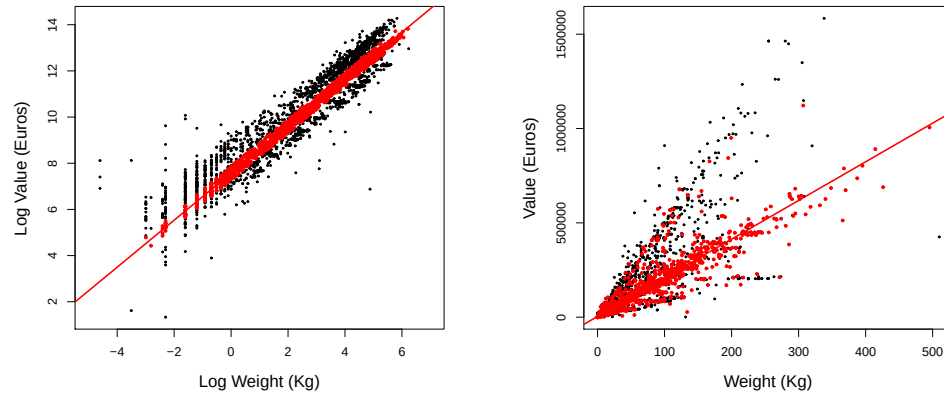


Figure 5.19: Fixed Point Clustering applied in the log-transformed data (left) and represented in the original scale (right). It detects only one big cluster of 2749 observations. 1970 observations remain unassigned.

More abstract definitions of cluster stability have been proposed by Ben-David and Ackerman in [ABD08] and [BDVLP06]. A remarkable result from the application of such formal definitions is that, as the authors claim, stability seems not well-suited as a tool to determine the number of clusters, because it is determined by symmetries in the data that may be unrelated to the clustering model.

Again Hennig [Hen05], but also Ng and Huang [NH02], have proposed a visual cluster validation approach, more subjective, that may be effectively used in our application context.

It is in our future plans to study the properties of our mixtures estimation and clustering proposal with a benchmark experiment on synthetic and real data. The generation of synthetic data from bi-variate distributions, so that to reproduce the most common patterns found in trade data or more general patterns of interest, can be done for example along the lines of Banfield and Raftery [BR93]. Once synthetic data are available, different measures of cluster validity, including those mentioned above, can be measured for the algorithms to assess an objective and systematic comparison among them.

5.10 Discussion and back to the perfect fit problem

In this Chapter I have proposed several new ideas and methods.

- The main method is to identify and estimate the components of a linear regression mixture using the standard Forward Search algorithm. The idea consists in finding a group of homogeneous observations with the Forward Search, trim the group, and repeat the procedure on the remaining data (Strategy 5.1).
- Then I have proposed two variants which adopt a relaxed definition of homogeneous group. The first variant is based on the re-weighted Forward Search (Strategy 5.2) that consists in using two test size levels, one to identify the robust regression line (the standard 1% of the Forward Search envelopes) and another to identify the outliers. The second variant consists in introducing a control on the R^2 during the progression of the Forward Search (Strategy 5.3).
- Furthermore, I have proposed a procedure to confirm, or validate, the tentative components detected with one of the previous algorithms.

The basic idea of the confirmatory step is to reallocate the observations of the dataset to the component to which they are closer.

- Finally, I have proposed a natural way of using the subjective opinion of the analyst about the appropriateness of the tentative components (5.5).

The rationale for proposing all these variants is that the concept of clustering, at the present stage of the research, is not completely formalised. No objective and universal (i.e. applicable to all clustering approaches) validation method exists and the quality of a clustering is at the end driven by the opinion and goals of the analyst. In our specific case, if the researcher wants to know whether the data is structured in multiple normal populations, then he will use the standard algorithm 5.1. If he is driven by operational goals, such as the identification of fraudulent transactions in data, he should look for very clear patterns and anomalies and use one of the relaxed versions 5.2 or 5.4 to identify rougher components and stronger outliers.

We can demonstrate the concept using the “perfect fit” dataset of Section 2.3, which at first sight is far from the mixture problems discussed in this chapter. This will show the big potential of the algorithms that I have proposed, for addressing problems of apparently different nature.

We defined as “perfect fit” a phenomenon occurring when the random fluctuations around a line are so small that minor deviations lead to large deletion residuals and so to declare as anomalous transactions which in absolute terms lie very close to the regression line. I checked the perfect fit dataset for possible sub-population first using the Forward Search standard linear mixture algorithm 5.1. It contains three almost perfectly parallel strips of observations, as shown in Figure 5.20. The zoom on the right highlights the three evident parallel linear components (a highest one represented by red circles, a central one represented by blue crosses and a lowest one represented by cyan squares). Even if the three components are very close one to the others, the Forward Search identifies them with high precision.

If we now analyse the same dataset with the R^2 -relaxed algorithm 5.4, with threshold $R_{th}^2 = 0.95$, we obtain one component and only three outliers, represented in Figure 5.21. In this case the algorithm is forced to disregard the real fine structure of the data because of the high linear correlation. However this result would be much more suitable for anti-fraud purposes, as the three tight groups of Figure 5.20 are of no operational use and to highlight them would even produce disaffection for the method in the inspectors.

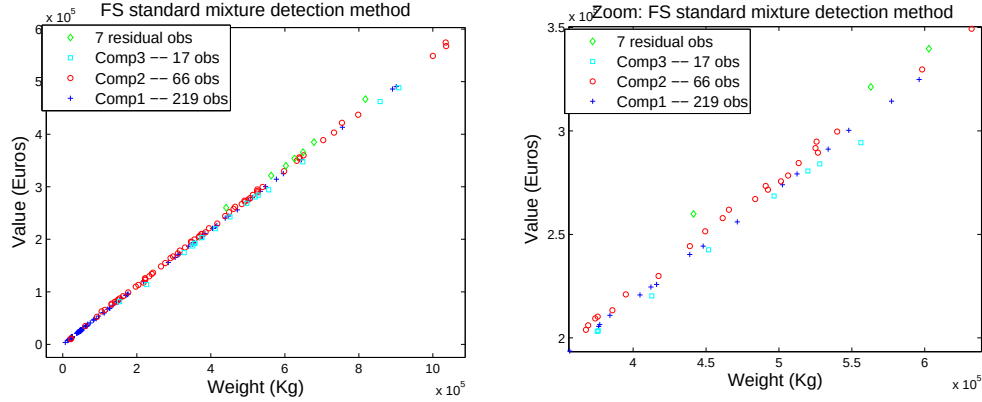


Figure 5.20: Perfect fit dataset. Tentative components detected by the Forward Search standard linear mixture algorithm 5.1. The right plot is a zoom of the plot on the left.

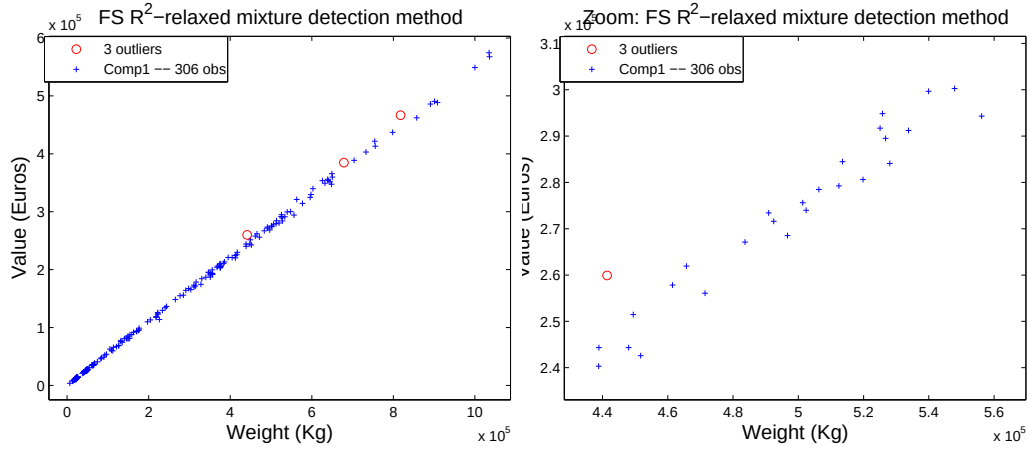


Figure 5.21: Perfect fit dataset. Tentative components detected by the R^2 -relaxed mixture algorithm 5.3, when $R_{th}^2 = 0.95$. The right plot is a zoom of the plot on the left.

The use of the second relaxed approach, based on reweighed Forward Search 5.2 with $\alpha = 0.0001\%$, brings to two components represented in Figure 5.22. As α is lower than the standard 1% the first component includes a large number of observations (255) and the second most of the remaining. The result is therefore somehow in between the previous two.

We conclude by coming back to the outlier detection context. Of course the application of the standard Forward Search outlier detection algorithm 3.3 brings to the first homogeneous group detected by the mixture algorithm 5.1 (the component in red in the left plot of Figure 5.20), and to 90 outliers.

It is interesting to compare this with the results of Least Trimmed Squares, which are similar as we can see in Figure 5.23.

An interesting paradox happens if we apply the non robust backward method 3.1 to the dataset. As shown in Figure 5.24 the method detects only a small number of outliers. This result, which would be much appreciated in the anti-fraud context, is the effect of a clear masking case: the groups of observations distributed along different almost parallel components mask each other and therefore no outliers are in practice detected. A problem that we try to avoid with sophisticated robust techniques becomes in this case an advantage.

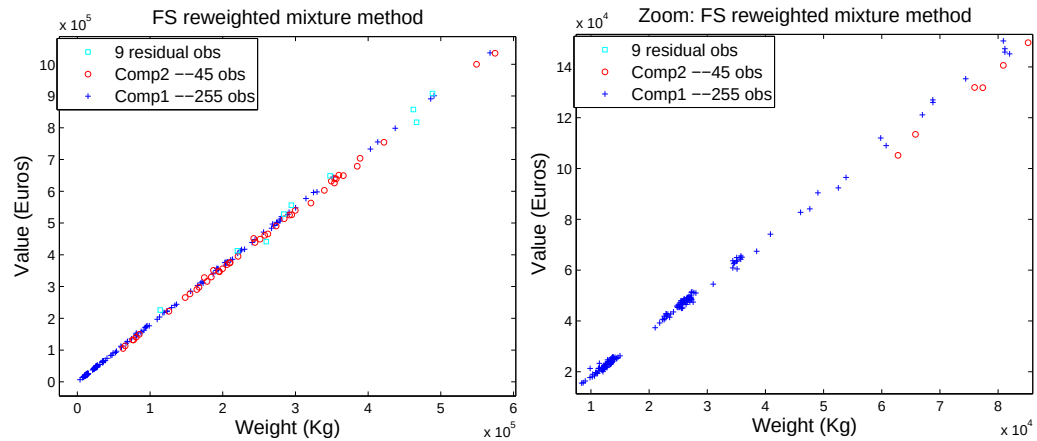


Figure 5.22: Perfect fit dataset. Tentative components detected by the reweighted Forward Search algorithm 5.2. The right plot is a zoom of the plot on the left.

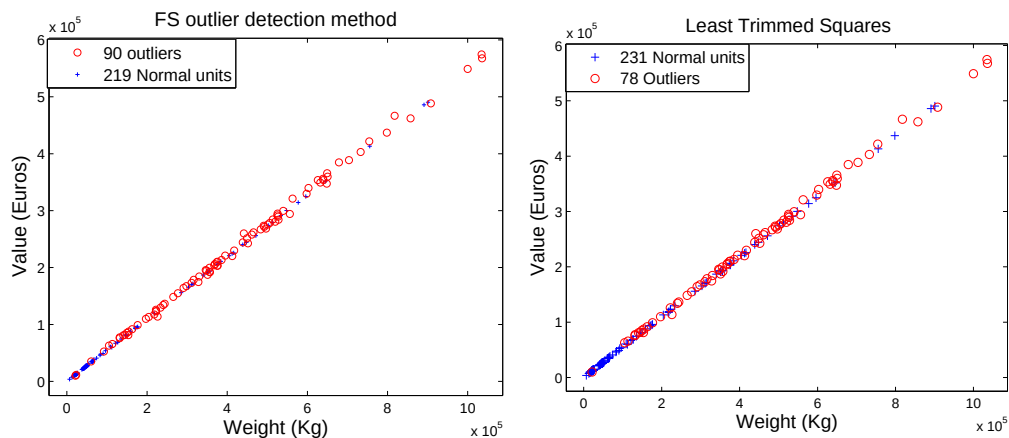


Figure 5.23: Perfect fit dataset. Outliers detected by the Forward Search (left plot) and by Least Trimmed Squares (right plot).

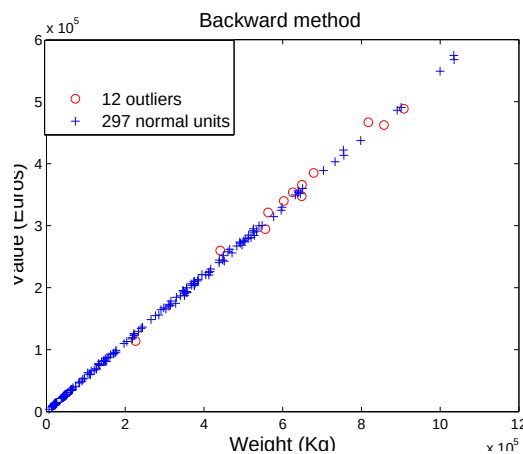


Figure 5.24: Perfect fit dataset. Outliers detected by the Backward algorithm 3.1.

Exploring and interacting with forward plots

6.1 Introduction

Throughout the dissertation I have presented the Forward Search mainly as an inferential tool. However, many researchers find that the major appeal of the method is still in its powerful exploratory data analysis capacity.

The big exploratory data analysis potential comes from the information provided by the wide variety of forward statistics that one can monitor during the search. In fact, often it happens that a lot of interesting features emerge unexpectedly during the progression of the Forward Search only when a specific combination of forward plots is inspected at the same time. Thus, *the analyst should be able to interact with the plots and redefine or refine the links among them*. In absence of dynamic linking and interaction tools, the analyst risks to miss relevant hidden information.

From these considerations, I have enthusiastically worked in a demanding software development project for integrating the Forward Search algorithms with a graphical library with flexible interaction tools, applicable to a variety of standard and forward plots. The result is the Forward Search MATLAB toolbox (<http://www.riani.it/MATLAB.htm>) that I used for the experiments described in the dissertation and that I quickly present in Section 7.

This work has already been brought at the attention of the statistical community, in particular with a recently published journal article [PRT09]. Of course for the details I suggest consulting the article. However, considering the importance of this subject and the considerable efforts that I have dedicated to it in the last year, I want to make a synthesis of the key

features in this Chapter. In doing so, I will mainly use the fishery data presented in Figure 2.5 of Section 2.4.

6.2 From static to dynamic linked robust forward plots

During the Forward Search process it is possible to monitor any relevant statistic. Section 3.5 has described some of them and the “perfect fit” dataset has been used to exemplify some forward plots (Figure 3.1).

This Section illustrates what we mean by “linking plots dynamically” and shows the advantages of such innovation. We will do it with a concrete example, the fishery dataset of Figure 2.5, which has a complex structure with more populations and such populations have an interesting interpretation for the end-user. We will show how the end-user with our tools can easily combine subject matter knowledge with information retrieved from the statistical plots and draw conclusions of relevant operational value.

We start in the context of data response transformation. Here, the null hypothesis is on the Box-Cox transformation parameter [BC64], $\lambda = \lambda_0$, and the added t tests based on constructed variables are known in the statistical literature as “score test for transformation” ([AR02]). The tool to understand the percentage of observations which are in accordance with the different values of the transformation parameters, is the forward plot of the score test statistic for transformation of the set of constructed variables for different values λ_0 , using a separate Forward Search for each λ_0 . These trajectories of the score tests can be combined in a single picture named the “fan plot” [AR00], that we have mentioned in Section 3.5.

Now, we observe that, being the relationship between trade value and quantity supposedly linear, trade data should not require transformation. The need of transformation may be due to the presence of multiple populations. For example, in the fishery dataset there are “fair” (market price) trade declarations together with price under-declarations (frauds).

The left panel of Figure 6.1 shows that the inclusion of the last observations, which enter at some point in the subset, causes strong rejection of the hypothesis of no transformation. After a simple brushing action on the trajectory of the fan plot for $\lambda = 1$ (i.e. no transformation required), it is immediate to see that the brushed units form a well defined cluster below the main cloud in the scatter plot (right panel of Figure 6.1).

On the other hand, the inclusion of the last observations causes acceptance of the hypothesis of square root transformation (left panel of Fig-

ure 6.2). However by looking at the scatterplot of the transformed data (right panel of Figure 6.2), it is evident that there is no linear relationship among the transformed response and the independent variable. Therefore the square root transformation is not appropriate.

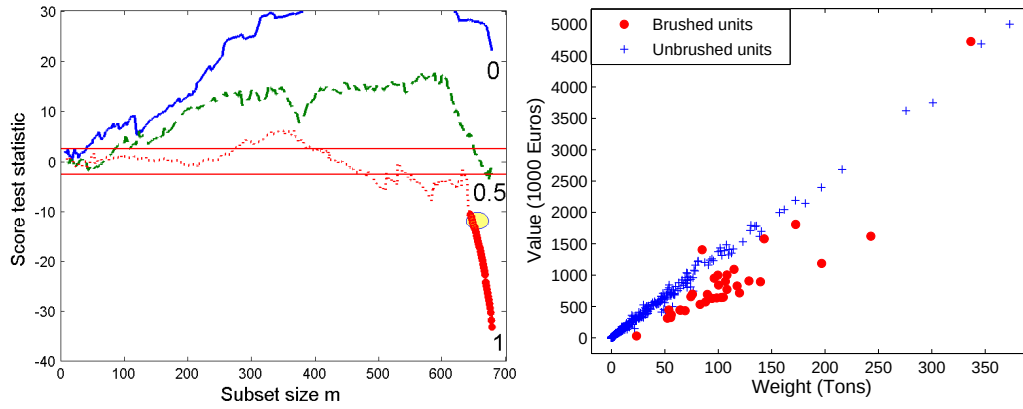


Figure 6.1: Fishery Trade Data. Fan plot for $\lambda = (0, 0.5, 1)$ (left panel) and the scatter of values (in thousands of euros) against quantities (in tons). The brushed units in the fan plot are automatically highlighted in the scatter plot. The brushed units form a separate cluster from the bulk of the data.

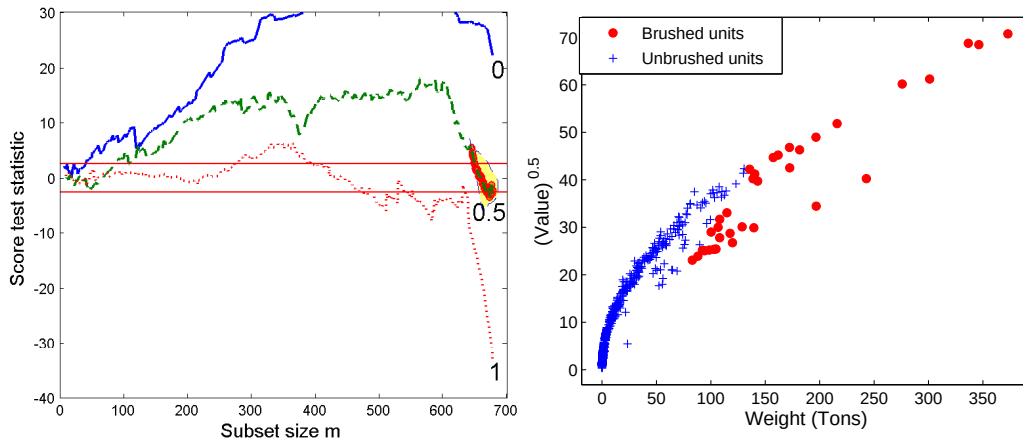


Figure 6.2: Fishery Trade Data. Fan plot for $\lambda = (0, 0.5, 1)$ (left panel) and the scatter of square root values (in thousands of euros) against quantities (in tons). The brushed units on the $\lambda = 0.5$ trajectory in the fan plot are automatically highlighted in the scatter plot.

Using repeated application of the Forward Search, [PT09] could identify the group of the “fair” import flows (which in the scatterplot of Figure 6.1 are represented with symbol ‘+’) to the other Member States, estimate their price (13 €/Kg) and finally divide the group of abnormal flows in other two linear clusters (which in Figure 6.3 are represented respectively with red circles and cyan squares) of significantly different estimated price, 9.17 vs 6.55 €/Kg. These two groups correspond to imports into a specific EU Member State (say MS7). In this case the prices declared look suspiciously low if compared to those of the other Member States.

However, there is a relevant fact that becomes clear only by inspecting in the input table (see bottom panel of Figure 6.3) the transactions corresponding to the red circles and cyan squares, by comparing their prices and the time periods when the transactions were recorded. In fact, the two abnormal clusters include flows that took place respectively in the first 14 consecutive months of the period analysed (red circles) and in the following consecutive 21 months (cyan squares). A natural conclusion of clear operational value is that in the period analysed the traders in MS7 gradually lowered their import price declarations, up to half of the reported price by the other Member States.

In [PT09] such unexpected pattern was found almost incidentally, by off-line inspection of the data table. Clearly this way of proceeding is not efficient: it requires big efforts and the chance of finding unexpected patterns of this type is limited. For example, without interactive tools it would be difficult to realize that near the origin of the axes there is a large number of linear groups of data, visible only by zooming that region.

Our solution for the end-user consists in linking the familiar tools such as the scatterplot of the import values and quantities and the table of the original data, where all relevant trade variables can be found. The brushing and linking tools can start from the forward plot of scaled residuals, minimum deletion residual, fan plot or scatter plot matrix. For example, if in Figure 6.3 we brush the units corresponding to the cyan squares symbols, we can see that they have the largest negative residuals throughout the search (left panel of Figure 6.4) and enter the search in the final steps (right panel of Figure 6.4). Inside this group, the observations with the highest absolute scaled residuals value throughout the search correspond to flows 200 and 212. These two observations enter the subset only in the last two steps of the procedure. Similarly, the brushing of the red circles in Figure 6.3 reveals that this group is characterized by trajectories with negative residuals intermediate to those of the other two groups and is formed by units which enter the search just before the group of the cyan squares.

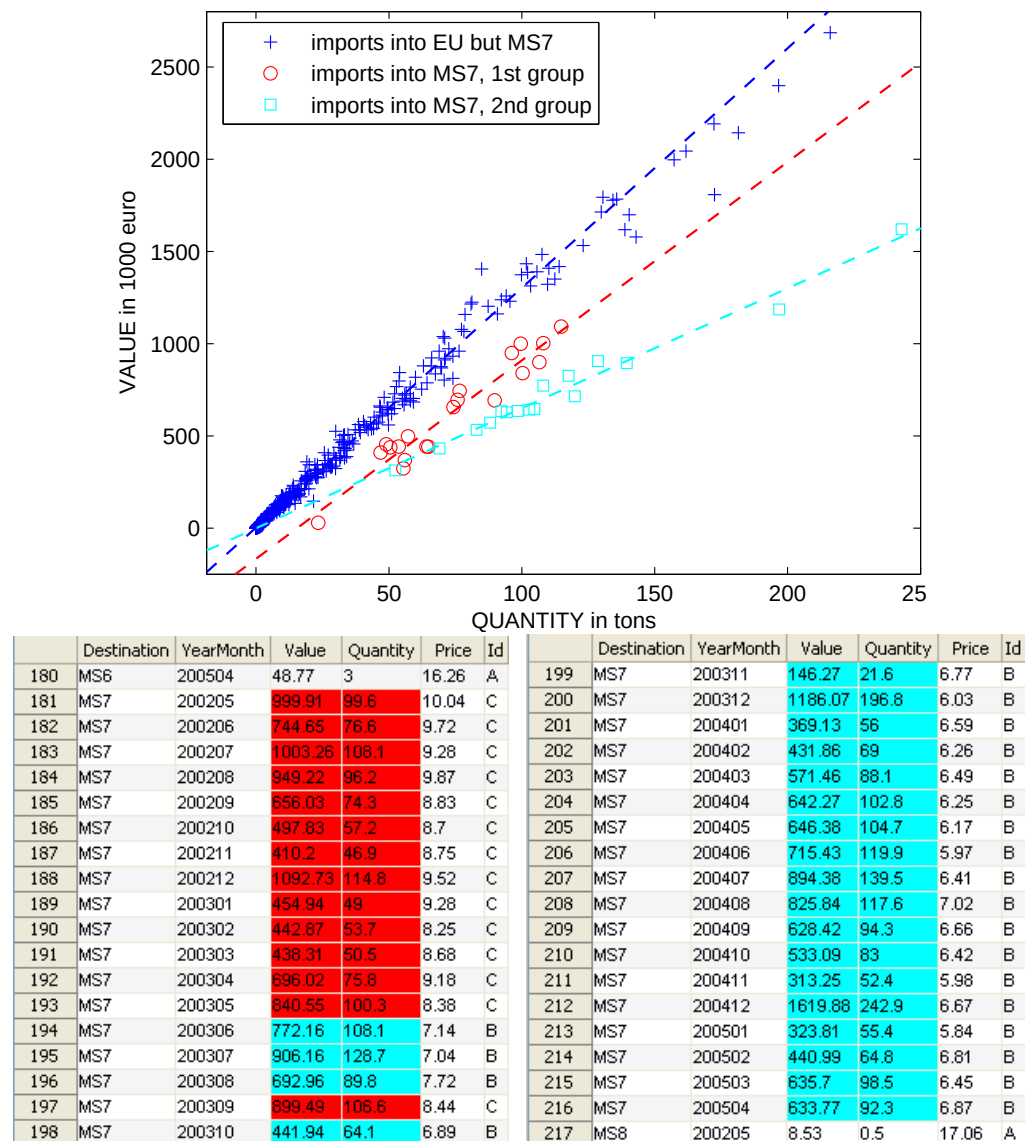


Figure 6.3: Fishery Trade Data. The scatterplot of values and quantities (top panel) and the data table with other relevant variables (bottom panels). Selected units in the scatterplot are automatically highlighted in the table and vice-versa. The ‘Id’ variable in the table indicates how the three groups of flows were identified by the Forward Search: the blue + in the scatterplot has id ‘A’ in the table, the cyan squares correspond to a ‘B’ and the red circles to a ‘C’.

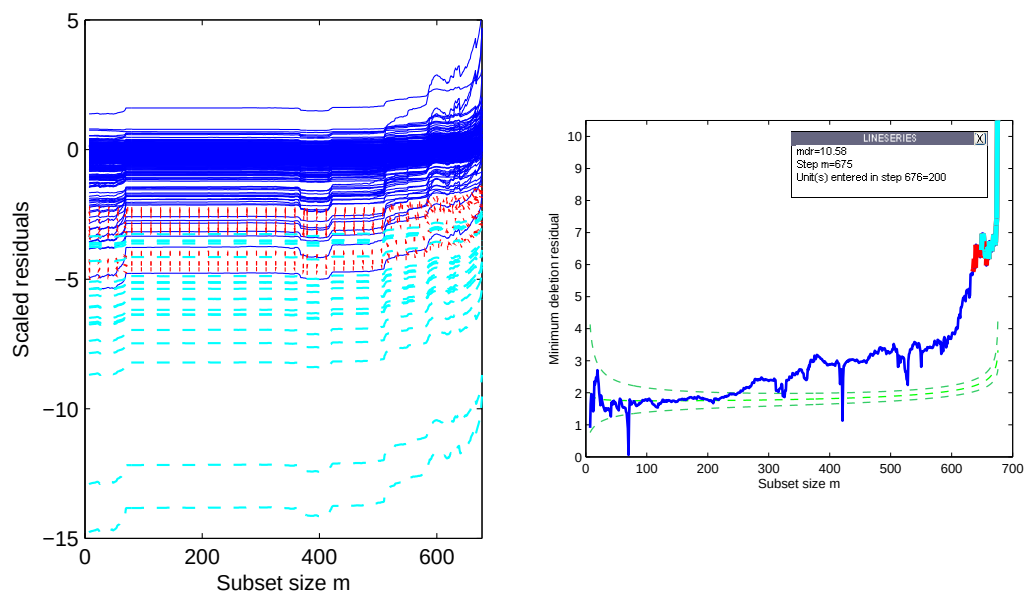


Figure 6.4: Fishery Trade Data. The scaled residuals trajectories (left) and the minimum deletion residual curve at each step of the FS (right). The units highlighted in the two plots refer to those selected in the scatterplot. The link is automatic also in this case.

Conclusions and Future Perspectives

In this thesis I have approached the Forward Search (FS) in the regression context under different perspectives, motivated and inspired by issues encountered in a concrete application context: the analysis of international trade data.

In the *work of systematisation* of the literature, almost necessary for a doctoral dissertation, I went beyond a standard survey of the state of the art of the subject. In trying to find reliable outlier signals among a very large number of trade flow combinations, it became clear how important is the issue of simultaneous testing for controlling the size of an inferential method and how difficult is to find an acceptable trade off between size and power. I found that these aspects, although concern numerous applications, have not been sufficiently addressed in the literature of robust statistic. Therefore, one of the contributions in this thesis (Chapter 3 and its integration in the Appendix) has been to investigate how the Forward Search achieves its nominal size, how it faces with the multiple testing issue and how it can be used to make inference with an arbitrary confidence level.

Addressing the statistical performance of the Forward Search tests naturally leads to the *validation perspective*. In Chapter 4 I discussed the results of a rigorous and extensive assessment of the actual size and power of the Forward Search and other classical robust methods in regression. The original contribution of this part of my work consists in having empirically proved, for the first time in regression, that the Forward Search can achieve at the same time high power and small size. I used a carefully defined benchmark platform to estimate empirically the actual computational complexity of the method. This has also contributed to estimate the current

limitations of the method in data mining applications.

A first *methodological contribution* of the thesis relies in the idea of relaxing the Forward Search in order to identify outliers with arbitrary significance levels, other than the standard 1% that is inherent to the method. This idea has the effect of re-weighting the robustness-efficiency trade off of the method, similarly to the re-weighted versions of LMS and LTS. I also proposed a different form of relaxing the Forward Search by monitoring and taking into account the R^2 value of the dataset analysed. This was introduced in the context of what I called “perfect fit” problem, which to my knowledge is not addressed in the literature.

Another methodological contribution of the thesis comes from having addressed the Forward Search from the point of view of clustering. In chapter 5 I proposed an extension of the Forward Search that can be used to identify and validate homogeneous sub-populations in the data that, in regression, manifest as mixtures of linear components. The new method is also capable of detecting outliers of the mixture, if present. This extension of the Forward Search algorithm has shown its potential in the analysis of international trade data, affected by different kinds of patterns at the same time.

Given that the inferential algorithm at the basis of the outlier detection, which I detailed in the Appendix, was ported to regression strictly following step by step the multivariate version of [RAC09], I found remarkable that such good results in terms of size and power are obtained in regression without any adjustment of the algorithm. Note that this is likely at the basis of the good behaviour of my extension of the Forward Search to mixture estimation, which is in fact based on the basic inferential outlier detection algorithm.

Lastly, I have addressed the Forward Search from the *exploratory data analysis perspective*. It is striking that, although the potential of the Forward Search in exploratory data analysis was clear since the introduction of the method, in practice its use in this context has been obstructed by the factual difficulties in inspecting the many plots produced by the method. This is why I contributed with enthusiasm to a project aimed at enabling the Forward Search with new dynamic and interactive graphical tool for extracting information from the forward plots. I described in Chapter 6 the proposed new functionalities and I demonstrated their effectiveness in extracting information and communicating statistical results to end-users.

In terms of future research, a number of new directions are now open. I mention those that I find of more interest.

For example, to address the Forward Search as a stochastic process (an empirical process of residuals for regression) is a challenging problem that

I have contributed to define in relation to the simultaneity and multiplicity issues inherent in the forward approach, and that others are now already studying.

I have indicated other benchmark experiments in the regression context, based on generating the explanatory variable values according to non normal distributions (uniform or, having still in mind trade data, by sampling the distribution of the explanatory variable in a set of real data so that to better account for leverage problems) or in the area of stochastic regression, where the explanatory variable values are not fixed in all replications, but generated each time from some distribution (again, several options are possible: normal, uniform, etc.).

The extension of the forward approach to the estimation of regression mixtures started from observing that the corresponding multivariate problem, based on monitoring the Mahalanobis distance, fails in presence of groups overlapping in proximity of dense areas. This has opened an interesting problem in the existing multivariate Forward Search methodology.

Besides, my strategy to validate/confirm the mixture components tentatively estimated by repeated application of a Forward Search and trimming steps, is only a first promising proposal. For example, I am studying an alternative confirmation phase that replaces the approach of partitioning the dataset and confirming each partition's part with the Forward Search, with the new idea of starting a search from a tentative component and let it progress so that *any* unit of the dataset, once entered in the subset, cannot leave it. I expect from this approach a more precise identification of the homogeneous subgroups in the dataset.

Finally, in the area of interactive visualisation there is lot to do, in terms of optimization of the graphical routines which are computationally very demanding and may require substantial revision of the graphical algorithms, but also in terms of introducing and implementing new concepts of statistical visualization.

Appendix to Chapter 3

Algorithmic Details of the Forward Search Inference Approach

A Introduction and formalism

This Appendix gives the algorithmic details of the Forward search in the regression context. The deep level of detail is perhaps superfluous if one aims at understanding the basic concepts and the mathematics of the Forward Search, but it is indispensable if the objective is to implement the method in a statistical programming environment.

There is no contribution in the literature that documents the Forward Search algorithm in such a detailed level. This was made possible here by a continuous feedback with the authors of the method.

We recall some notations and concepts treated in the dissertation. Assume to have one univariate response Y and $p - 1$ explanatory variables $X_1, \dots, X_{p-1}$ satisfying the usual regression model $E(y_i) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{p-1} x_{ip}$ with assumptions of independent, additive errors of constant variance. We run the Forward Search on a set of n observations on Y and on the explanatory variables X and we monitor the minimum deletion residual among the observations not in the subset.

The envelopes are used in the Forward Search for detecting outliers, non homogeneous data (e.g. clusters), violations of normality assumptions, etc.. This is done with a two stage process.

1. In the first stage (Sections B and C) we run a search on the data and we monitor the progression of the appropriate statistic against the envelopes bounds until we obtain a potential signal indicating that an observation $m^\dagger$, and therefore also the succeeding observations, may be outliers.
2. Then, in the second stage we identify the precise (outlier free) subset

of homogeneous observations by superimposing the envelopes with the procedure described in Section D.

If matrix X is in a general position (that is all subsets of size p are full rank) the search starts with p , that is the minimum number of observation needed to estimate the model parameters. The monitoring of the minimum deletion residual may start from $p + 2$, because its computation is based on estimate of S^2 at step $p + 1$. However, we are not interested in monitoring the very first steps because of the high fluctuations of such values. We thus monitor the minimum deletion residual from step $3p + 1$.

Let us divide the progression of the search in three parts:

- The **initial** part which goes from $initialsubsetsize = p + 2$ to $init = 3p$.
- The **central** part formed by the central steps from $init + 1$ to $istep - 1$, where $istep = n - \left\lceil 13\sqrt{\frac{n}{200}} \right\rceil$. We assume that $istep > init + 4$ for ensuring that the central part includes at least three units (see below the sections on signal detection).
- The **final** part formed by the last steps from $istep$ to $n - 1$ (i.e. about 6% of the points fall here).

We denote with $r_{\min}(a, b)$ the value of the minimum deletion residual (among the observations not in the subset) at step a , evaluated using an envelope based on b observations, where a goes from $3p + 1$ to $b - 1$.

B Stage 1a: signal detection

The signal detection is based on monitoring consecutive triplets or single extreme values.

In the central part of the search we have either:

1. **An extreme triplet** i.e. there are 3 consecutive values above the 99.99% envelope. These values are: $r_{\min}(m^\dagger - 1, n)$, $r_{\min}(m^\dagger, n)$ and $r_{\min}(m^\dagger + 1, n)$, where $m^\dagger = init + 2, init + 3, \dots, istep - 2$.

Therefore, the first and last triplets which are considered are

$$\{r_{\min}(init + 1, n), r_{\min}(init + 2, n), r_{\min}(init + 3, n)\}$$

and

$$\{r_{\min}(istep - 3, n), r_{\min}(istep - 2, n), r_{\min}(istep - 1, n)\}.$$

- 2. An extreme single value** i.e. a value above the 99.999% envelope.
We check this condition for $m^\dagger = \text{init} + 2, \text{init} + 3, \dots, \text{istep} - 1$.

Thus, first extreme value which is considered is $r_{\min}(\text{init} + 2, n)$ and the last is $r_{\min}(\text{istep} - 1, n)$.

Remark. If the sample is homogeneous, the probability of having a value above the 99.99% envelope at any step is much larger than the nominal 0.01%, because the envelopes are simultaneous but pointwise. On the other hand, the probability of observing a stretch of consecutive values above the 99.99% threshold is much smaller. This explains the choice of the 3 consecutive values.

In the final part of the search there is:

- 3. An extreme couple adjacent to an exceedance**, i.e. there are 2 consecutive values above the 99.9% envelope and 1 (which precedes or follows the two consecutive) above the 99% envelope. Again, these three values are referred to steps $m^\dagger - 1, m^\dagger, m^\dagger + 1$.

The first and the last triplets which are considered are

$$\{r_{\min}(\text{istep} - 2, n), r_{\min}(\text{istep} - 1, n), r_{\min}(\text{istep}, n)\}$$

and

$$\{r_{\min}(n - 4, n), r_{\min}(n - 3, n), r_{\min}(n - 2, n)\}.$$

This means that $m^\dagger \in [\text{istep} - 1, \dots, n - 3]$.

Remark. This relaxes the previous condition 2 because, when the sample is homogeneous, the probability of observing a stretch of consecutive observations above an extreme threshold is smaller in the final part of the search.

- 4. A potential couple of outliers**, i.e. the value at penultimate step, $m^\dagger = n - 2$ of $r_{\min}(n - 2, n)$, is above the 99.9% threshold.

Remark 1. The probability to have such a signal is 0.1%.

Remark 2. Note that in item 3, $m^\dagger$ was considered up to step $n - 3$, because an extreme couple adjacent to an exceedance at step $n - 2$ implies that $r_{\min}(n - 2, n)$ is certainly above the 99.9% threshold.

- 5. A single outlier**, i.e. the value at last step, $m^\dagger = n - 1$ of $r_{\min}(n - 1, n)$, is above the 99% threshold. In this case, given that there is only one unit left outside the subset, such unit is detected as isolated outlier and the procedure terminates.

Remark. If we consider 10000 replicates from a normal distribution, we expect to find this signal in about 100 searches.

We want type-I errors to occur mainly in the last steps of the Forward Search: a signal before the last steps should be clear indication of structure in the data. With the above strategy, the size of the test is slightly larger than the nominal 1% which comes from B.

C Stage 1b: signal validation

The potential signal, detected at step $m^\dagger$, becomes a true signal when one of the following conditions is true:

- A The signal takes place in the final part of the search.
 - B There are three consecutive values, $r_{\min}(m^\dagger - 1, n)$, $r_{\min}(m^\dagger, n)$ and $r_{\min}(m^\dagger + 1, n)$, above the 99.999% threshold (uncontrovertible signal).
 - C There is a certain number of values h , not necessarily consecutive, in the central and final part of the search above the 99.999% threshold (uncontrovertible signal). So far it has been put $h = 10$. In the future it may be reasonable to define h according to the sample size.
- Remark.* When we have two groups which overlap considerably, or a heavy tailed distribution, the value $r_{\min}$ may not manifest visible peaks during the search and lies steadily just above the 99% threshold. In presence of lot of small exceedances, it is therefore reasonable to consider the tentative signal, which takes place at $m^\dagger$, as real.
- D The signal is above the last value of the 1% envelope based on $m^\dagger + 1$ observations. Little is lost if, instead of considering this threshold, we rather use as an approximation the final value (step $n - 1$) of the 1% envelope based on n observations. In general, the approximated threshold is slightly larger than the correct one based on $m^\dagger + 1$ observations, because the minimum outside the subset in the final step, that is the distribution of the n -th ordered statistic, increases as n increases. The quality of the approximation improves as $m^\dagger$ gets close to $n - 1$. For example, when $n = 100$ and $m^\dagger = 83$ the difference is $3.835 - 3.808 = 0.0265$, while when $n = 1000$ and $m^\dagger = 600$ the difference is $4.367 - 4.238 = 0.1290$.

Remark 1. In an homogeneous sample this can hardly occur.

- E There is a peak followed by a dip, i.e. there are at least two values, not necessarily consecutive, below the 1% envelope based on n observations within the steps going from $m^\dagger + 1$ to $m^\dagger + \min(30, n - m^\dagger)$.

Remark. If at a certain point of the search the subset only contains units from one group, the search will tend to include first the units coming from that group. Therefore, $r_{\min}$ will start to lie outside the envelope based on n observations and it will continue to increase up to when all units of the group are included. Finally this curve will reach a peak and then suddenly dip below the lowest envelope, as relatively remote observations from other groups enter the subset used in fitting and change the slope of the regression hyperplanes. According to our experience the entry of observations from other groups forces the value to fall below the 1% threshold for a certain number of steps just after the peak: that is the reason why we verify this condition in a reasonably short interval (i.e. not greater than 30 steps) and we require at least two values below the threshold. In presence of this pattern, the signal must be considered true.

D Stage 2: superimposition of the envelopes

Once a true signal takes place at $m = m^\dagger$, in the second stage of the inferential procedure we start superimposing 99% envelopes for $n = m^\dagger - 1, m^\dagger, m^\dagger + 1, \dots$ until one of the two following stopping rules is fulfilled, say at step $m^\dagger + k$:

- S1 The final, or the penultimate, or the antepenultimate values, that is $d_{\min}(m^\dagger + k, m^\dagger + k + 1)$, or $d_{\min}(m^\dagger + k - 1, m^\dagger + k + 1)$, or $d_{\min}(m^\dagger + k - 2, m^\dagger + k + 1)$, are above the 99% threshold.
- S2 We have a value above the 99.9% threshold for any $m \geq m^\dagger$, say m^* , that is $d_{\min}(m^*, m^\dagger + k + 1)$ exceeds the 99.9% envelope.

Note that it is preferable to start superimposing at $m^\dagger - 1$ rather than at $m^\dagger$, because the signal at $m^\dagger$ may be based on the triplet starting from $m^\dagger - 1$ and therefore the first candidate unit to be declared as outlier is the one joining in the search at step $m^\dagger - 1$.

Finally, the number of homogeneous units is determined as follows:

- H1 When the stopping rule takes place at step $m^\dagger + k$ (i.e. the envelope is based on $m^\dagger + k + 1$ observations), and $m^* = m^\dagger + k$, we declare $m^\dagger + k$ homogeneous observations.
- H2 On the other hand if $m^* < m^\dagger + k$:
- a We declare $m^\dagger + k$ homogeneous observations if from step $m^* + 1$ to step $m^\dagger + k$ all values are above the 1% threshold based on $m^\dagger + k + 1$ units.
 - b If from step $m^* + 1$ to step $m^\dagger + k$, there is at least a value, say at step $m^{1\%}$, below the 1% threshold based on $m^\dagger + k + 1$ units, this means that the subset of $m^\dagger + k$ observations is not homogeneous: there are likely to be at least two overlapping groups. Tentatively, the homogeneous set is made up of the units in correspondence of the step where the maximum value takes place over the interval which goes from $m^\dagger$ to $m^{1\%} - 1$. Typically, this maximum takes place at step m^* .

E Bonferroni signal

Up to now we have described the typical patterns giving rise to a signal. However, there is an extreme situation that may happen at any point in the search. More precisely, it may occur that $r_{\min}(m, n)$ for any m is above the 99% envelope based on all the observations. This is the so called Bonferroni threshold.

If such an extreme event occurs at step $m^\dagger$ and there is no extreme triplet associated with $m^\dagger$ (cases 1 or 3, i.e. extreme triplet or extreme couple adjacent to an exceedance), then $n - m^\dagger + 1$ observations are declared outliers. We call this situation **isolated Bonferroni signal**.

In case a Bonferroni signal and an extreme triplet both occur at step $m^\dagger$, it is necessary to proceed in the usual way, by superimposing the envelope at step $m^\dagger - 1$ to check whether the number of outliers is $n - m^\dagger + 2$ or $n - m^\dagger + 1$.

It is clear that, in presence of an isolated Bonferroni signal, there is no need of superimposing the envelopes because, if the value of $r_{\min}(m, n)$ is greater than the 99% envelopes based on n observations, it is clearly also greater than the final step of the 99% envelope based on $m^\dagger$ observations. In other words, the stopping rule takes place at $m^\dagger$.

Appendix to Chapter 6

Forward Search Data Analysis Matlab toolbox

A Introduction

In this Appendix I will describe the main characteristics of the Matlab Forward Search Data Analysis (FSDA) Toolbox software, that I have contributed to develop and that I have used for the experiments described in the thesis. The FSDA toolbox extends the statistics toolbox of Matlab to *robust analysis of complex data sets* and to *data visualization functions*. From the perspective of the statistical analysis, the toolbox is composed by the regression routines and the multivariate routines. The last ones are under construction, while the first ones are rather stable. Therefore, in the following of this Annex, I will refer only to regression routines. The dataset that I will use in all the examples is the fishery dataset introduced in Figure 2.5 of Section 2.4. The FSDA toolbox, that can be freely downloaded in <http://www.riani.it/MATLAB.htm>, is equipped with a help program to which I refer for details, examples and demos of the toolbox. Figure A shows the introductory page of the FSDA help program with, on the left, the structure of the whole help.

B Robust analysis of complex data sets

The toolbox extension to the robust analysis of complex data sets consists in a series of Matlab functions that implement (1) the Forward Search and other robust regression techniques, (2) a transformation of the response variable and (3) a robust forward variable selection through the added t tests.

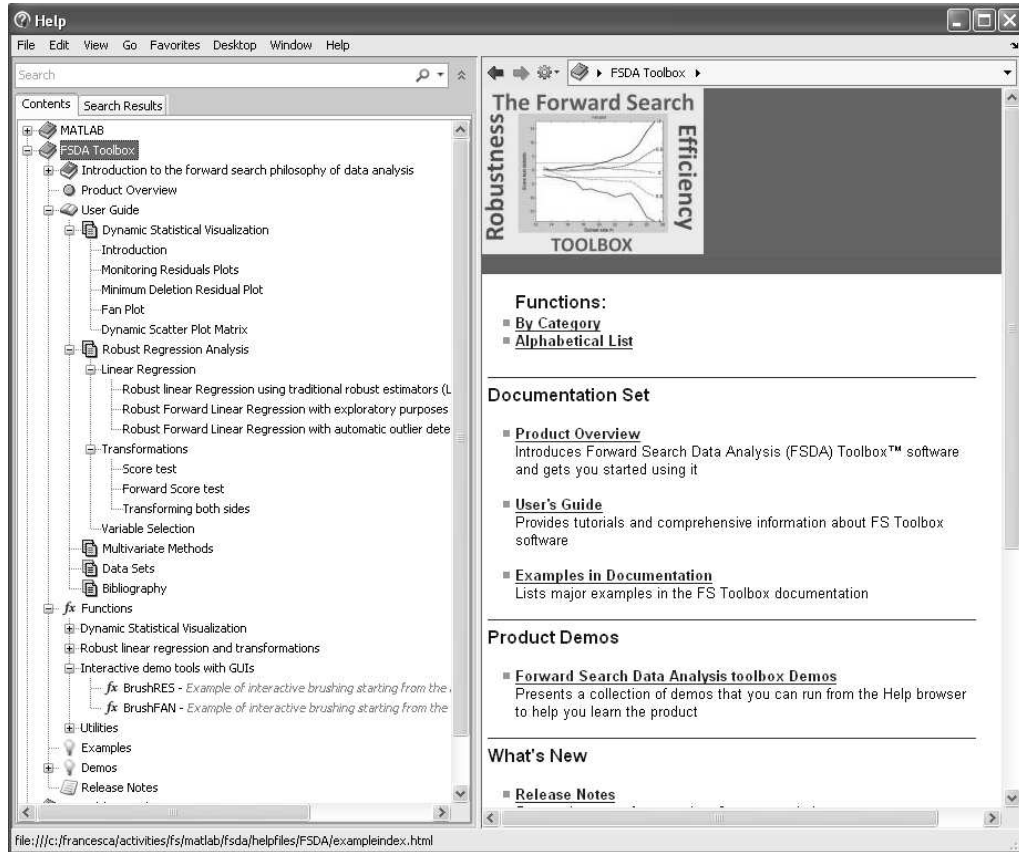


Figure A: FSDA Help. Structure (on the left) and introductory page (on the right).

B.1 Robust regression techniques

The robust regression techniques that we have implemented in the FSDA toolbox are on one hand the two standard robust techniques that we have presented in Section 4.2, i.e. the Least Trimmed Squares and the Least Median of Squares, on the other hand the Forward Search. The description of the main functions that implement the three techniques follow:

- *LXS* function implements Least Trimmed Squares and Least Median of Squares estimators. Using option “rew” it is also possible to consider their raw version or their reweighted version (example in Figure B);
- *FSR* function implements the Forward Search automatic outlier detection procedure (example in Figure C);

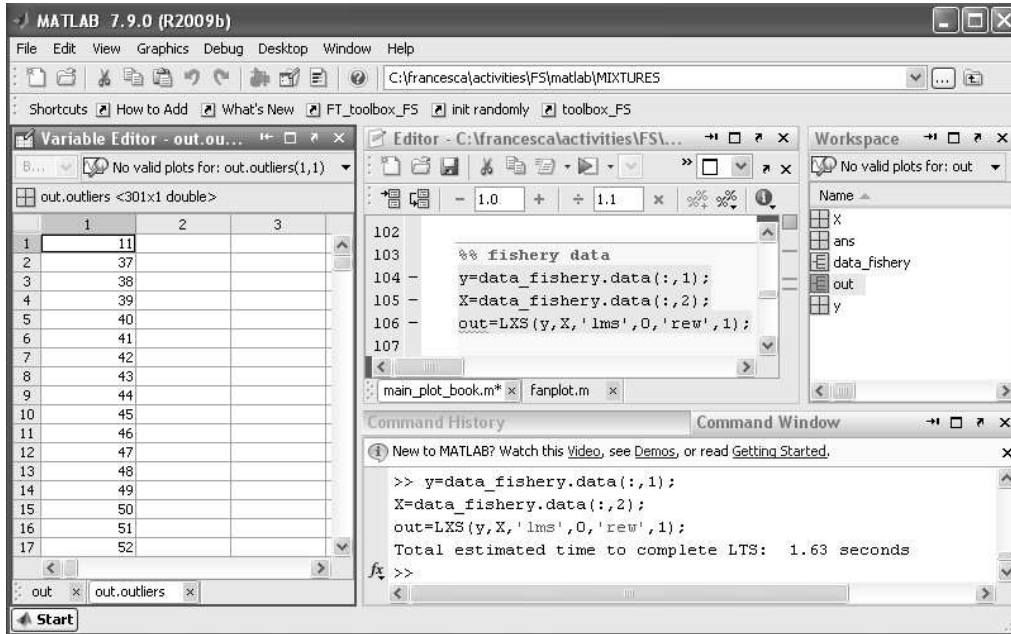


Figure B: LXS function. The reweighted LTS procedure, based on all $\binom{n}{p}$ possible subsets, has been executed and has produced, among the other output, the outliers list (table on the left).

- *FSReda* function implements the Forward Search with exploratory purposes. In fact it enables to store a series of quantities along the Forward Search evolution, e.g. residuals, leverage, minimum deletion residual outside subset (example in Figure D);
- *FSRbsb* function implements the Forward Search with the purpose of monitoring the units entering the subset in each step of the Forward Search (example in Figure E);
- *FSRmdr* implements the Forward Search with the purpose of monitoring the minimum deletion residual and other basic linear regression quantities, such as the regression parameters and the linear correlation coefficient, in each step of the search (example in Figure F).

B.2 Transformations

It can happen that the response variable of some dataset has to be transformed for improving the regression analyses of the datasets. At this goal we have implemented in the FSDA toolbox the parametric family of power

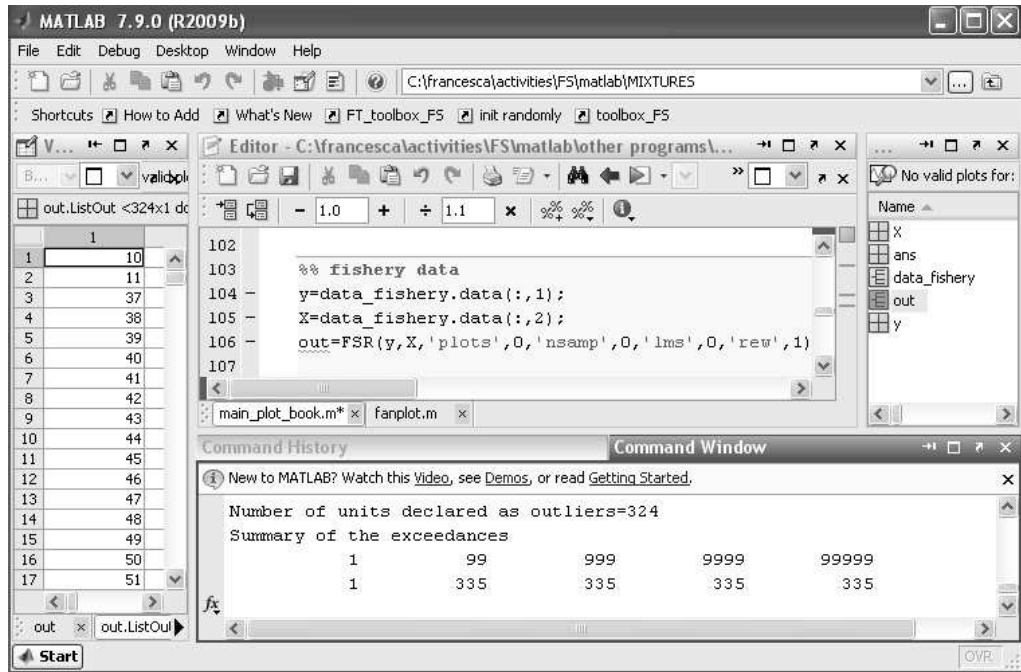


Figure C: FSR function. The reweighted LTS procedure, based on all $\binom{n}{p}$ possible subsets, has been executed in order to compute the initial robust subset. The Forward Search has been run and has produced, among the other output, the outliers list (table on the left).

transformations of [BC64], the details of which are discussed by [AR00]. Given that the estimated transformation and related test statistic may be sensitive to the presence of outliers, we use the Forward Search to see how the estimates and statistics evolve as we move through the ordered data. Influential observations may only be evident for some transformations of the data. Since observations that appear as outlying in untransformed data may not be outlying once the data have been transformed, and vice versa, we employ the Forward Search on data subject to various transformations, as well as on untransformed data. At this goal we have built the following functions:

- *Score test* function approximates score test statistic for testing the transformation $\lambda = \lambda_0$ (example in Figure G);
- *FSRfan* implements the Forward Search with the purpose of monitoring the values of the score test statistic for several values of the transformation parameter λ in each step of the Forward Search. If the size

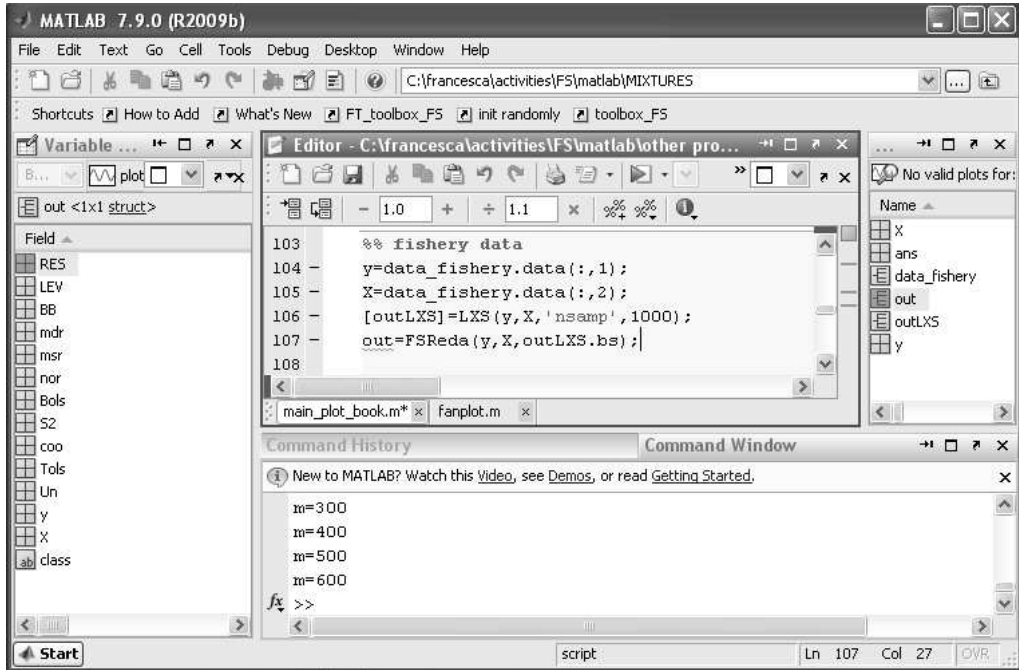


Figure D: FSReda function. The raw LTS procedure, based on 10000 possible subsets, has been executed in order to compute the initial robust subset. The Forward Search has been run and has produced the monitoring of many statistics (on the left), such as the scaled residuals (RES) and the leverage (LEV).

of the data set is small (i.e. smaller than 100) generally it is enough to consider the five most common value of $\lambda = -1, -0.5, 0, 0.5, 1$. The simultaneous forward plot of the score test statistic for several values of the transformation parameter λ is known in the literature as “fan plot”. Each search is separate, so that the observations may enter in different orders in the searches. This plot enables us to appraise in a quantitative way the percentage of observations in agreement with the different values of the transformation parameter (example in Figure H).

B.3 Variable selection

The Forward Search sorts the observations by the magnitude of their residuals from the fitted subsets. Therefore the value of the residual mean square estimate s^2 of equation 3.4 increases during the search. As a consequence, even in the absence of outliers and model inadequacies, the values of the

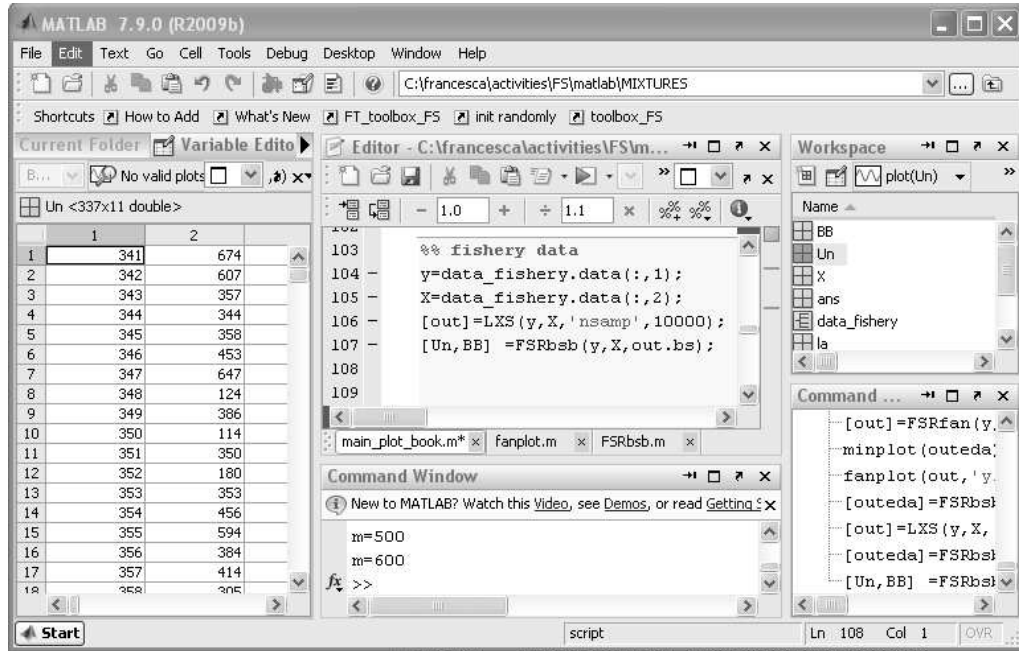


Figure E: FSRbsb function. The raw LTS procedure, based on 10000 possible subsets, has been executed in order to compute the initial robust subset. The Forward Search has been run and has produced the monitoring of the units entering the subset in each step of the Forward Search (table on the left).

t-statistics for the parameters in the model decrease during the search and are hard to interpret. [AR02] in order to overcome this problem use the method of added variables to provide plots of t-tests which are orthogonal to the search. We have therefore implemented in our toolbox the function *FSRaddt* which produces the forward deletion t-statistics for each explanatory variable (example in Figure I).

C Data visualization functions

The toolbox extension to data visualization concerns a series of functions which produce robust and dynamic statistics plots. The description of the main of these functions follows:

- *resplot* function plots the scaled residuals of all units in each step of the Forward Search. It is a tool to identify outliers and, more in general, groups of observations which behave differently from the

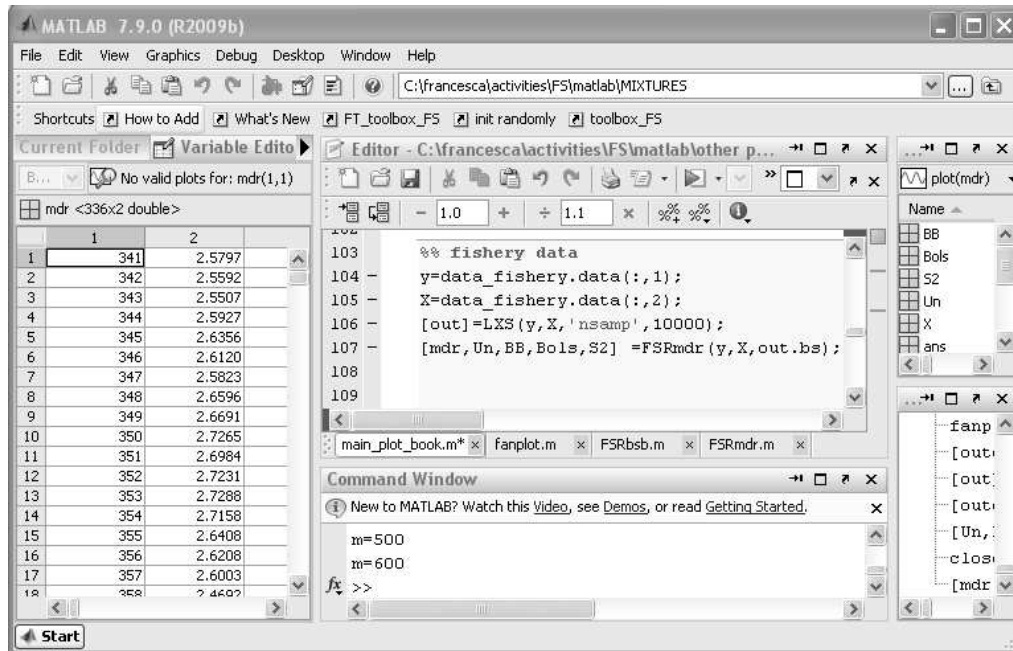


Figure F: FSRmdr function. The raw LTS procedure, based on 10000 possible subsets, has been executed in order to compute the initial robust subset. The Forward Search has been run and has produced the monitoring of the minimum deletion residual (table on the left) and other basic linear regression quantities in each step of the Forward Search.

rest of the data. In this plot the option “datatooltip” can be used to provide information about the unit selected, the step in which the unit enters the search and the associated label. On the other hand, option “databrush” enables the user to select a set of residual trajectories and to see them highlighted in the matrix of scatter plots of y against each column of X , which is automatically created. Brushed units are also automatically highlighted in other forward plots (e.g. minimum deletion residual plot) if they are already open (example in Figure J);

- *minplot* function plots the minimum deletion residual in each step of the Forward Search. If one or more atypical observations are present in the data, the plot of minimum deletion residual will show a peak in the step prior to the inclusion of the first outlier. The plot may show a subsequent decrease, due to the effect of masking, as further outliers enter the subset. In this plot the option “datatooltip” can be used to provide information about the units which enter the search in

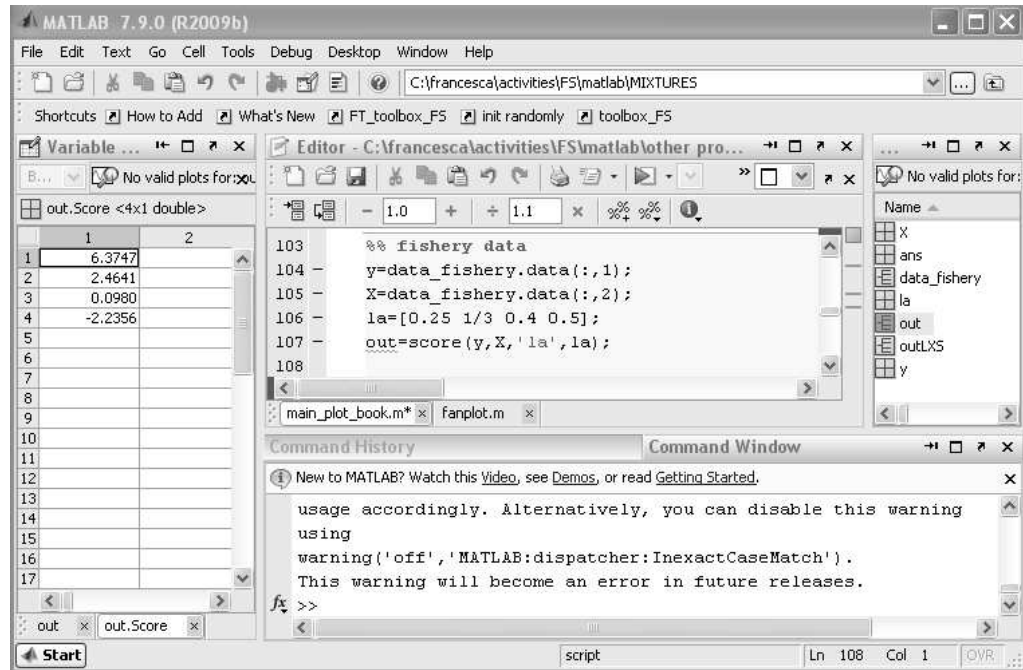


Figure G: Score function. The score procedure has been run and has produced, for each value of the vector of transformation λ , the values of the score test (table on the left).

the selected step. On the other hand, option “databrush” enables the user to select a part of the curve of minimum deletion residual and to see the corresponding units highlighted in the scatter plot matrix and in other forward plots (example in Figure K);

- *yXplot* function plots the dependent variable against the columns of the independent variables in the input dataset. The function, based on the MATLAB “gplotmatrix” function, allows to select units in one of the scatter plots and to produce automatically the monitoring residuals plot.

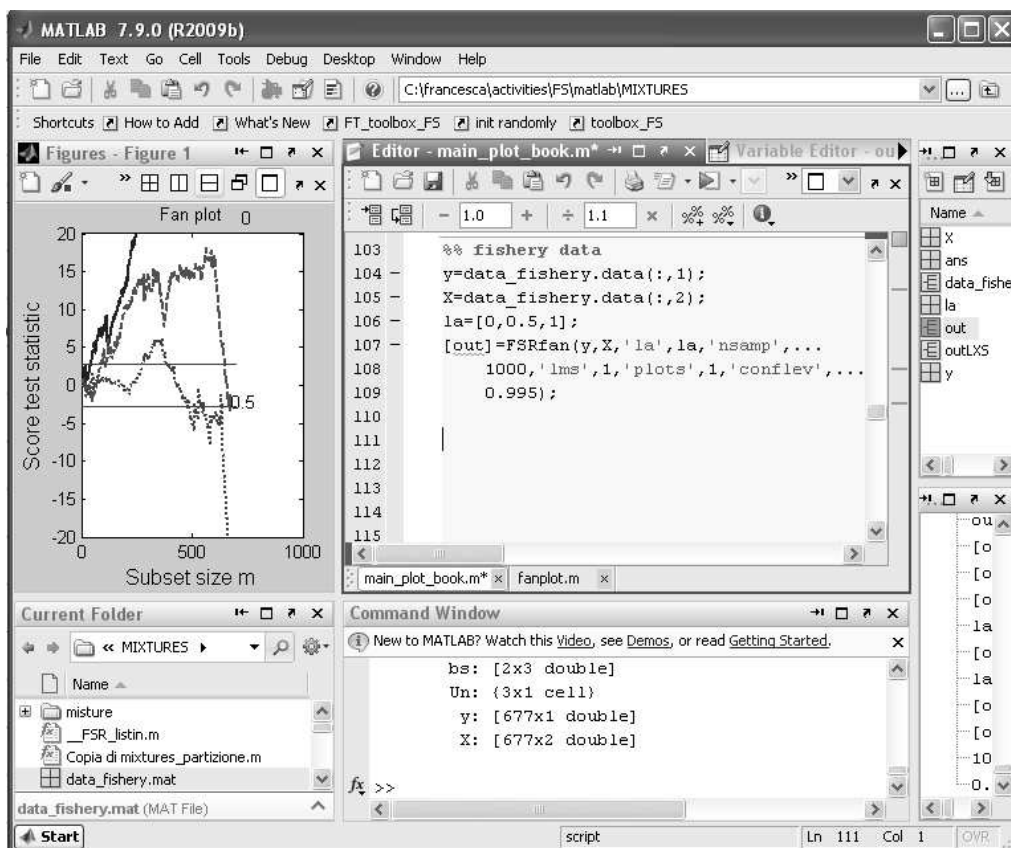


Figure H: FSRfan function. The Forward Search has been run and has produced the plot on the left which monitors the values of the score test statistic for the transformation parameters $\lambda = 0, 0.5, 1$ (see Section 6.2).

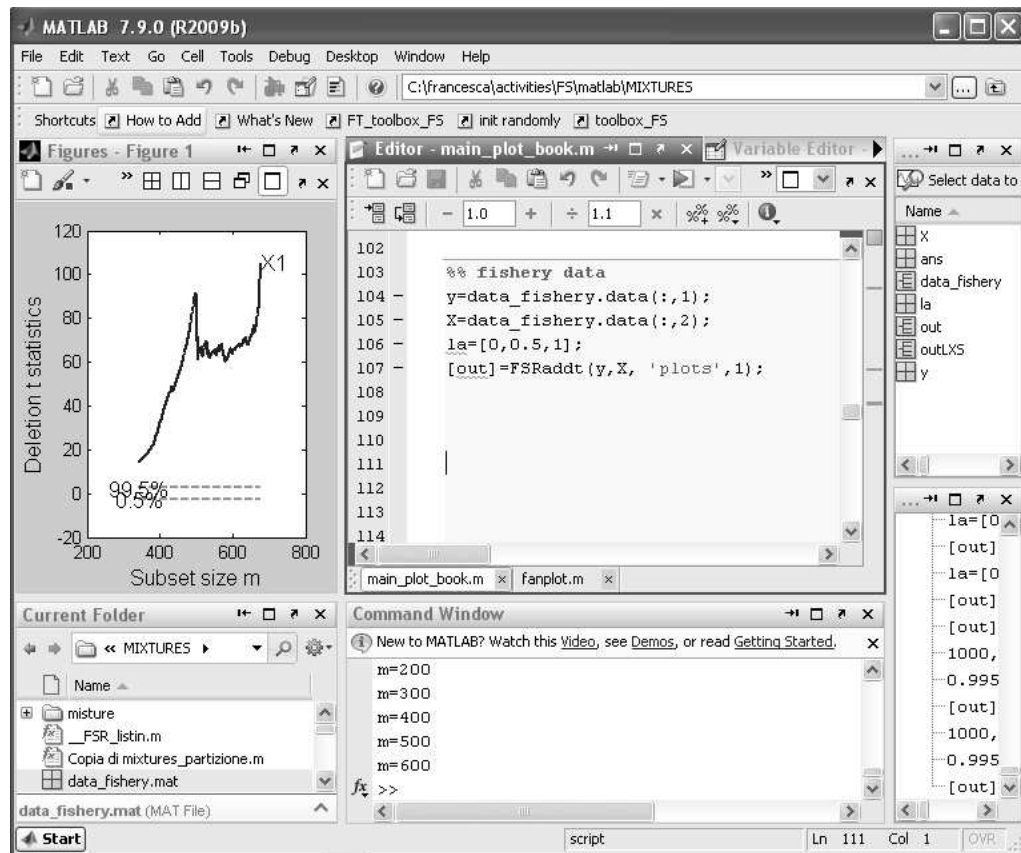


Figure I: The FSRaddt function has produced the plot on the left which monitors the deletion t-statistics.

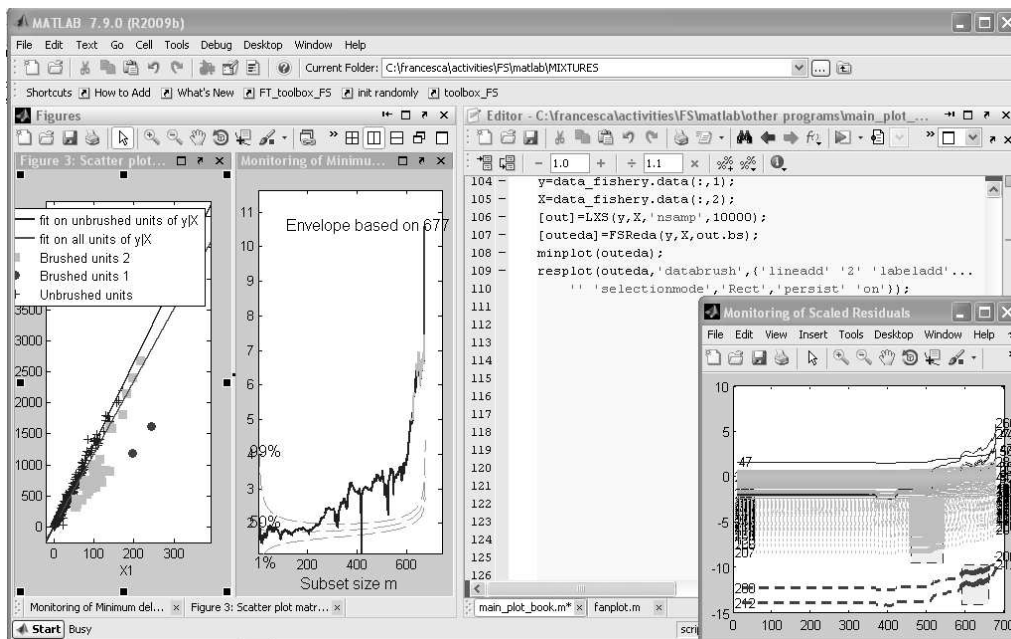


Figure J: The resplot function has produced the plot of the scaled residuals (on the right). Thanks to the options “databrush” and “persist=on” we could select two groups of trajectories, the red and the cyan ones, which are automatically highlighted in the plot of the minimum deletion residuals and in the scatterplot (on the left). The option “lineadd=2” fits in the scatterplot an ols line based on “simple regression” to all points and another one to the unselected points.

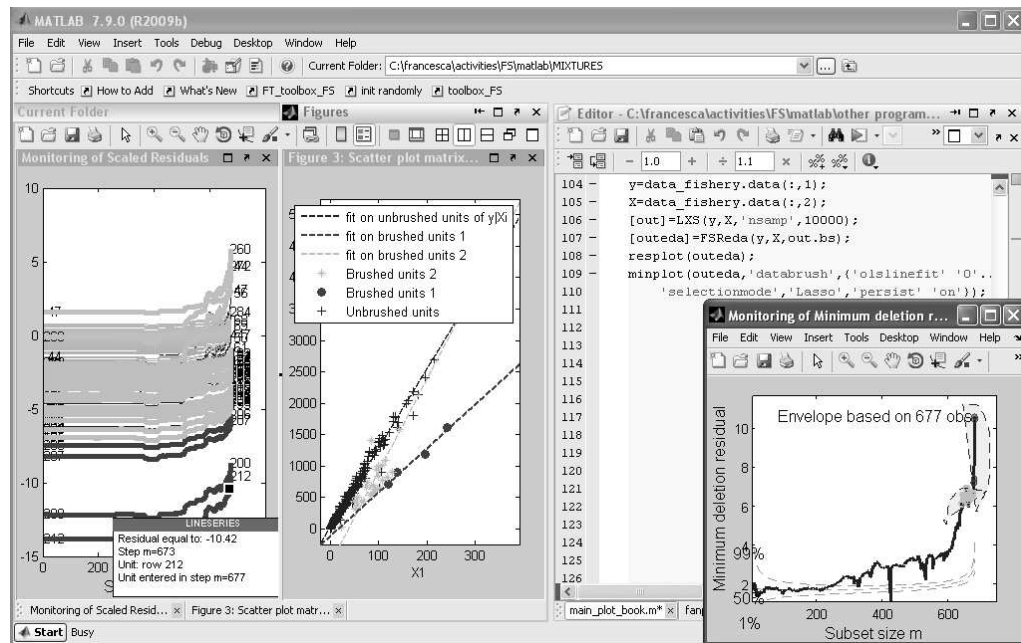


Figure K: The minplot function has produced the plot of the minimum deletion residuals among observations not in the subset (on the right). Thanks to the options “databrush” and “persist=on” we could select two parts of the minimum deletion residual curve, the red and the cyan ones, which are automatically highlighted in the plot of the scaled residuals and in the scatterplot (on the left). The option `olslinefit=0` fits an ols line based on “multiple regression” for each selected group.

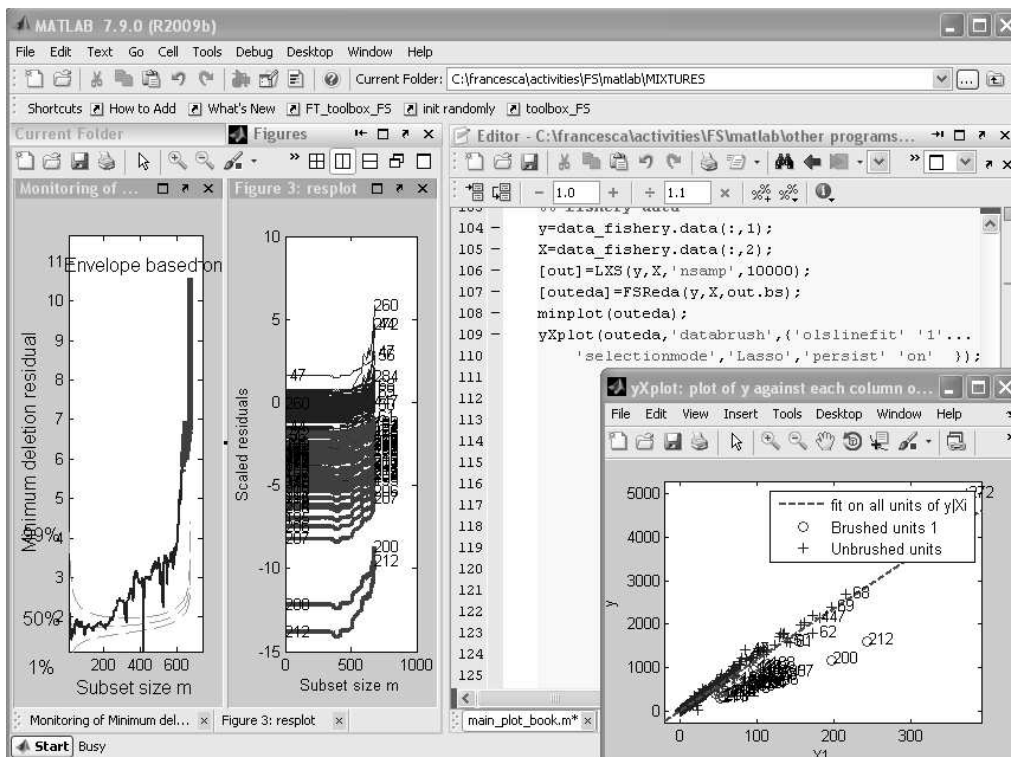


Figure L: The yXplot function has produced the scatterplot of the data (on the right). Thanks to the options “databrush” we could select the points in the lower part of the scatterplot, which are automatically highlighted in the plot of the scaled residuals and in the minimum deletion residual plot (on the left). The option “olslinefit=1” fits an ols line based on “multiple regression” to all the observations in the scatterplot.

Bibliography

- [ABD08] M. Ackerman and S. Ben-David. Which data sets are clusterable? - a theoretical study of clusterability. Submitted, 2008.
- [AR00] A. C. Atkinson and M. Riani. *Robust Diagnostic Regression Analysis*. Springer Series in Statistics. Springer-Verlag, New-York, 2000.
- [AR02] A. C. Atkinson and M. Riani. Forward search added-variable t-tests and the effect of masked outliers on model selection. *Biometrika*, 89(4):939–946, 2002.
- [AR06] A. C. Atkinson and M. Riani. Distribution theory and simulations for tests of outliers in regression. *Journal of Computational and Graphical Statistics*, 15:460–476, 2006.
- [AR07] A. C. Atkinson and M. Riani. Exploratory tools for clustering multivariate data. *Computational Statistics & Data Analysis*, 52:272–285, 2007.
- [ARC04] A. C. Atkinson, M. Riani, and A. Cerioli. *Exploring Multivariate Data With the Forward Search*. Springer Series in Statistics. Springer-Verlag, New-York, 2004.
- [Atk94] A. C. Atkinson. Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association*, 89(428):1329–1339, 1994.
- [BC64] G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252, 1964.
- [BDJ93] R. W. Butler, P. L. Davies, and M. Jhun. Asymptotics for the minimum covariance determinant estimator. *The Annals of Statistics*, 21(3):1385–1400, 1993.

- [BDVLP06] S. Ben-David, U. Von Luxburg, and D. Pal. A sober look at clustering stability. In *Lecture Notes in Computer Science*, pages 5–19. Springer, 2006.
- [BF81] R. C. Bolles and M. A. Fischler. A ransac-based approach to model fitting and its application to finding cylinders in range data. In *IJCAI'81: Proceedings of the 7th international joint conference on Artificial intelligence*, pages 637–643, San Francisco, CA, USA, 1981. Morgan Kaufmann Publishers Inc.
- [BG99] C. Becker and U. Gather. The masking breakdown point of multivariate outlier identification rules. *Journal of the American Statistical Association*, 94(447):947–955, 1999.
- [BH95] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.
- [BKW80] D. A. Belsley, E. Kuh, and R. E. Welsch. *Identifying influential data and sources of collinearity*. Wiley Series In Probability And Statistics. John Wiley And Sons Ltd, 1980.
- [BR93] J.D. Banfield and A.E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, 1993.
- [CF09] P. L. Conti and A. Farcomeni. Distributional convergence of robust mahalanobis distances obtained from the forward search. In M. Riani, A. Cerioli, C. Agostinelli, and D. Perrotta, editors, *ICORS 2009 International Conference of Robust Statistics, Book of Abstracts*, pages 46–47, Parma, June 2009. Libero Libri - Claudio Agostinelli.
- [CR99] A. Cerioli and M. Riani. The ordering of spatial data and the detection of multiple outliers. *Journal of Computational and Graphical Statistics*, 8(2):239–258, 1999.
- [CR02] A. Cerioli and M. Riani. Robust methods for the analysis of spatially autocorrelated data. *Statistical Methods and Applications*, 11(3):335–358, October 2002.
- [CRA09] A. Cerioli, M. Riani, and A. C. Atkinson. Controlling the size of multivariate outlier tests with the mcd estimator of scatter. *Statistical Computing*, 19:341 – 353, 2009.

- [CV07] F. Crosilla and D. F. S. Visentini. An automatic classification and robust segmentation procedure of spatial objects. *Statistical Methods and Applications*, 15:329–341, 2007.
- [DC88] W. DeSarbo and W. Cron. A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, 5(2):249–282, September 1988.
- [GEG05] L. A. Garcia-Escudero and A. Gordaliza. Generalized radius processes for elliptically contoured distributions. *Journal of the American Statistical Association*, 100:1036–1045, September 2005.
- [Had92] A. S. Hadi. Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(3):761–771, 1992.
- [Ham74] F. R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.
- [Hen98] C. Hennig. Clustering and outlier identification: Fixed point cluster analysis, 1998.
- [Hen02] C. Hennig. Fixed point clusters for linear regression: Computation and comparison. *Journal of Classification*, 19(2):249–276, 2002.
- [Hen03] C. Hennig. Clusters, outliers, and regression: Fixed point clusters. *Journal of Multivariate Analysis*, 86(1):183–212, 2003.
- [Hen05] C. Hennig. A method for visual cluster validation. In C. Weihs and W. Gaul, editors, *Classification - The Ubiquitous Challenge*, pages 153–160. Springer, Heidelberg, 2005.
- [Hen07] C. Hennig. Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis*, 52(1):258 – 271, 2007.
- [HR05] J. Hardin and D. Roche. The distribution of robust distances. *Journal of Computational and Graphical Statistics*, 14(4):928–946, 2005.
- [HS93] A. S. Hadi and J. S. Simonoff. Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association*, 88(424):1264–1272, 1993.

- [HT87] Y. Hochberg and A. C. Tamhane. *Multiple comparison procedures*. Wiley Series In Probability & Mathematical Statistics: Applied Probability & Statistics. John Wiley And Sons Ltd, 1987.
- [Lei04] F. Leisch. Flexmix: A general framework for finite mixture models and latent class regression in r. *Journal of Statistical Software*, 11:1–18, 2004. The FlexMix manual is available at: <http://cran.R-project.org/doc/packages/flexmix.pdf>.
- [Mar85] M. G. Marasinghe. A multistage procedure for detecting several outliers in linear regression. *Technometrics*, 27(4):395–399, 1985.
- [Mil66] R.G. Miller. *Simultaneous Statistical Inference*. New York: McGraw Hill, 1966.
- [MMss] D. Mavridis and I. Moustaki. The forward search algorithm for detecting aberrant response patterns in factor analysis for binary data. *Journal of Computational and Graphical Statistics*, in press.
- [MMY06] R. A. Maronna, R.D. Martin, and V. J. Yohai. *Robust Statistics*. Wiley, Chichester, 2006.
- [MP00] G.J. McLachlan and G.J. Peel. *Finite Mixture Models*. Wiley, New York, 2000.
- [NFDN07] N. Neykov, P. Filzmoser, R. Dimova, and P. Neytchev. Robust fitting of mixtures using the trimmed likelihood estimator. *Computational Statistics & Data Analysis*, 52(1):299 – 308, 2007.
- [NH02] M. Ng and J. Huang. M-fastmap: A modified fastmap algorithm for visual cluster validation in data mining. In *Advances in Knowledge Discovery and Data Mining*, volume 2336/2002 of *Lecture Notes in Computer Science*, pages 224–236. Springer, Berlin / Heidelberg, 2002.
- [NM02] N. Neykov and C. H. Muller. Breakdown points and computation of trimmed likelihood estimators in generalized linear models. In R. Dutter, P. Filzmoser, U. Gather, and P. J. Rousseeuw, editors, *Developments in Robust Statistics*, pages 277–286. Physica Verlag, Heidelberg, 2002.

- [PR09] T. Proietti and M. Riani. Transformations and seasonal adjustment. *Journal of Time Series Analysis*, 30(1):47–69, January 2009.
- [PRT09] D. Perrotta, M. Riani, and F. Torti. New robust dynamic plots for regression mixture detection. *Advanced Data Analysis and Classification*, pages 263–279, 2009.
- [PT09] D. Perrotta and F. Torti. Detecting price outliers in european trade data with the forward search. *Data Analysis and Classification: From Exploration to Confirmation*, 2009.
- [RA00] M. Riani and A. C. Atkinson. Robust diagnostic data analysis: Transformations in regression. *Technometrics*, 42(4):384–394, 2000.
- [RA01] M. Riani and A. C. Atkinson. A unified approach to outliers, influence, and transformations in discriminant analysis. *Journal of Computational and Graphical Statistics*, 10(3):513–544, 2001.
- [RA07] M. Riani and A. C. Atkinson. Fast calibrations of the forward search for testing multiple outliers in regression. *Advances in data analysis and classification*, 1:123–141, 2007.
- [RAC07] M. Riani, A. C. Atkinson, and A. Cerioli. Results in finding an unknown number of multivariate outliers in large data sets. Technical report, London School of Economics, Department of Statistics, 2007.
- [RAC09] M. Riani, A. C. Atkinson, and A. Cerioli. Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society B*, 71:1–21, 2009.
- [Ria04] M. Riani. Extensions of the forward search to time series. *Studies in Nonlinear Dynamics & Econometrics*, 8(2):1–23, 2004.
- [Ria09] M. Riani. Robust transformations in univariate and multivariate time series. *Econometric Reviews*, 28:262–278, 2009.
- [RL87] P. J. Rousseeuw and A. M. Leroy. *Robust regression and outlier detection*. Wiley-Interscience, New York, 1987.
- [Rou84] P. J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79:871–880, 1984.

- [RVD06] P. J. Rousseeuw and K. Van Driessen. Computing lts regression for large data sets. *Data Mining and Knowledge Discovery*, 12:29–45, 2006.
- [SBVA08] M. Salibian-Barrera and S. Van Aelst. Robust model selection using fast and robust bootstrap. *Computational Statistics & Data Analysis*, 52(12):5121 – 5135, 2008.
- [SP07] N. Solaro and M. Pagani. The forward search for metric multi-dimensional scaling. In *Classification and Data Analysis 2007*, pages 381–384. EUM Editore, 2007.
- [Spe01] R. Spence. *Information Visualization*. Addison Wesley, California, 2001.
- [Tuf83] E. R. Tufte. *The Visual Display of Quantitative Information*. CT:Graphics Press, Cheshire, 1983.
- [VDV98] A.W. Van Der Vaart. *Asymptotic statistics*. Cambridge Series in statistical and probabilistic mathematics, 1998.