



European
Commission

JRC TECHNICAL REPORT

Tools for Monitoring Robust Regression in SAS IML Studio

*S, MM, LTS, LMS and Especially
the Forward Search*

Francesca Torti, Domenico Perrotta,
Anthony C. Atkinson, Aldo Corbellini, Marco Riani

2020

This publication is a Technical report by the Joint Research Centre (JRC), the European Commission's science and knowledge service. It aims to provide evidence-based scientific support to the European policymaking process. The scientific output expressed does not imply a policy position of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of this publication. For information on the methodology and quality underlying the data used in this publication for which the source is neither Eurostat nor other Commission services, users should contact the referenced source. The designations employed and the presentation of material on the maps do not imply the expression of any opinion whatsoever on the part of the European Union concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

Contact Information

Name: Francesca Torti
Address: Joint Research Centre, Via Enrico Fermi 2749, 21027 Ispra (VA), Italy
E-mail: francesca.torti@ec.europa.eu
Tel.: +39 0332 786209

EU Science Hub

<https://ec.europa.eu/jrc>

JRC121650

EUR 30341 EN

PDF ISBN 978-92-76-21438-0 ISSN 1831-9424 doi:10.2760/35922

Luxembourg: Publications Office of the European Union, 2020

© European Union, 2020



The reuse policy of the European Commission is implemented by the Commission Decision 2011/833/EU of 12 December 2011 on the reuse of Commission documents (OJ L 330, 14.12.2011, p. 39). Except otherwise noted, the reuse of this document is authorised under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence (<https://creativecommons.org/licenses/by/4.0/>). This means that reuse is allowed provided appropriate credit is given and any changes are indicated. For any use or reproduction of photos or other material that is not owned by the EU, permission must be sought directly from the copyright holders.

All content © European Union, 2020

How to cite this report: Torti, F., Perrotta, D., Atkinson, A.C., Corbellini, A. and Riani, M., *Tools for Monitoring Robust Regression in SAS IML Studio: S, MM, LTS, LMS and Especially the Forward Search*, EUR 30341 EN, Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-21438-0, doi:10.2760/35922, JRC121650

Table of contents

Abstract	5
Acknowledgments	6
1 Introduction	7
2 Three Classes of Estimator for Robust Regression	7
3 Algebra for the Forward Search	8
4 Testing for Outliers	9
5 Regression Outlier Detection in the FS	10
6 FS Analysis of the Transformed Loyalty Card Data.....	11
7 Why SAS?.....	12
8 The FS batch procedure	15
9 Timing comparisons.....	16
10 Transformation of the Response	17
10.1 Analysis of Loyalty Card Data with the FS on the Original Scale.....	17
10.2 The Fanplot	19
10.3 Loyalty Card Data	20
11 Monitoring Other Forms of Robust Regression.....	21
11.1 Soft Trimming.....	21
11.2 M and S Estimation.....	23
12 Data Analyses with S, LTS and LMS Routines	24
13 Discussion and Extensions.....	25
A Annex: code to replicate the results and the figures in the report.....	30
B Annex: use of the monitoring tools in WebAriadne.....	31

List of figures

1	Example of SAS IML Studio code which uploads the Loyalty card data in SAS IML.	11
2	Loyalty card data: monitoring plots on for transformed data with $\lambda = 0.4$. The top panel shows the absolute values of minimum deletion residuals among observations not in the subset; the last part of the curve, corresponding to the 18 identified outliers, is automatically highlighted in red (in the on-line .pdf version). The bottom panel shows the scaled residuals, with the trajectories corresponding to the 18 detected outliers automatically represented in red (in the on-line .pdf version). The box under each panel contains the SAS code used to generate the plot.	12
3	Loyalty card data: scatterplots of transformed data when $\lambda = 0.4$, with the 18 outliers detected plotted as red crosses (in the on-line .pdf version). The box under the figure contains the SAS code used to generate the plots.	13
4	Loyalty card data: monitoring of estimated beta coefficients on transformed data when $\lambda = 0.4$, with the part of the trajectory corresponding to the 18 detected outliers, highlighted in red (in the on-line .pdf version). The box under the figure contains the SAS code used to generate the plots.	13
5	Two artificial datasets generated with MixSim for the assessment.	15
6	Top panel. Boxplots showing, for different values of k (the <code>fs_step</code> parameter, on the x-axis), the bias and dispersion of the estimated slopes and intercepts (respectively from left to right for each k). The estimates are obtained from 500 simulated datasets of 5,150 observations. Bottom panel: percentage of estimated values lying outside the boxplot whiskers for slope (blue asterisks) and intercept (black circles).	16
7	Execution time of our SAS (R9.4) and MATLAB (R2018b) implementations of the FSR function; for SAS, the comparison is also with the batch version of FSR (with $k = 10$). The assessment covers data with one explanatory variable and size ranging from 30 to 100,000. Results are split into three panels for small, medium and large data sizes. The last, bottom-right, panel gives the ratio between the time required by the MATLAB implementation and the two SAS ones. The associated table reports the time in seconds for selected sample sizes.	17
8	Loyalty card data: monitoring plots on untransformed data. The top panel shows the absolute minimum deletion residuals among observations not in the subset, with the last part of the trajectory corresponding to the 82 detected outliers, automatically highlighted in red (in the on-line .pdf version). The bottom panel shows the scaled residuals with the trajectories corresponding to the 82 detected outliers, automatically represented in red (in the pdf version). The box under each panel contains the SAS code used to generate the plot.	18
9	Loyalty card data: scatterplots of untransformed data, with the 82 detected outliers automatically plotted as red crosses (in the on-line .pdf version). The box under the figure contains the SAS code used to generate the plots.	19
10	Loyalty card data: monitoring of estimated beta coefficients from untransformed data, with the last part of the trajectory, corresponding to the 82 detected outliers, automatically highlighted in red (in the on-line .pdf version). The box under the figure contains the SAS code used to generate the plots.	19
11	Loyalty card data: fanplot for seven values of λ . The box under the figure contains the code used to generate the plot.	20
12	Loyalty card data: monitoring plots for log-transformed data ($\lambda = 0$). The top panel shows the absolute values of minimum deletion residuals among observations not in the subset, with the last part of the trajectory corresponding to the 14 detected outliers, automatically highlighted in red (in the on-line .pdf version). The bottom panel shows the scaled residuals with the trajectories corresponding to the 14 detected outliers automatically represented in red (in the on-line .pdf version). The box under each panel contains the SAS code used to generate the plot.	21
13	Loyalty card data: scatterplots of log-transformed data ($\lambda = 0$), with the 14 detected outliers automatically plotted as red crosses (in the on-line .pdf version). At the figure foot, the code used to generate the plots.	22
14	Loyalty card data: monitoring of estimated beta coefficients for log-transformed data ($\lambda = 0$), with the last part of the trajectory, corresponding to the 14 detected outliers, automatically highlighted in red (in the on-line .pdf version). At the figure foot, the code used to generate the plots.	22

15	A rather complex trade dataset. Following Perrotta and Torti (2018), we analyze the subset of retained units: "Printed books, brochures, leaflets and similar printed matter" (723 units).	24
16	Books data; S, LTS and LMS estimators: monitoring the scaled residuals as the breakdown point varies. At the foot of each panel, the code used to generate the plot.	25
17	The login page of the WebARIADNE and THESEUS applications.	31
18	WebARIADNE. Left panel: selection of dataset and statistical application of interest. Right panel: wizard for importing a new dataset.	32
19	WebARIADNE. Selection of an existing dataset: example of data preview.	32
20	WebARIADNE. Selection of input parameters for the execution of a given method.	32
21	WebARIADNE. Selection of an existing set of results obtained (in the example) with the Forward Search and the LTS.	32
22	THESEUS. Left panel: list of identified outliers. Right panel: scatterplot of an homogeneous group of data plus a strong outlier.	33

Abstract

This report focuses on robust regression tools that are at the core of a JRC system for the routine generation and dissemination of EU import prices and the detection of patterns of anti-fraud relevance in large volumes of trade. These tools have been implemented in SAS in the context of a project supported by the Hercule III program of the European Commission. Although the development framework is very specific to anti-fraud, the applicability of the SAS package is much wider and the underlying models (previously conceived by the academic co-authors of the report) are very general.

The forward search (FS) is a general method of robust data fitting that moves smoothly from very robust to maximum likelihood estimation. The regression procedures are already included in a MATLAB toolbox, FSDA, developed by the same authors of this report. The work on a SAS version of the FS originates from the need for the analysis of large data sets expressed by law enforcement services operating in the European Union that can use our SAS software for detecting data anomalies that may point to fraudulent customs returns. The series of fits provided by the FS leads to the adaptive data-dependent choice of highly efficient robust estimates. It also allows monitoring of residuals and parameter estimates for fits of differing robustness. Our SAS package applies the idea of monitoring to several robust estimators for regression for a range of values of breakdown point or nominal efficiency, leading to adaptive values for these parameters. Examples in the report are for S estimation and (not yet included in FSDA) for Least Median of Squares (LMS) and Least Trimmed Squares (LTS) regression.

Specific to our SAS implementation, we describe the approximations used to provide fast analyses of large datasets using a FS with batches. We also present examples of robust transformations of the response in regression. Further, our package provides the SAS community with methods of monitoring robust estimators for multivariate data, including multivariate data transformations.

Acknowledgments

This work was mainly supported by Administrative Arrangements of the “Automated Monitoring Tool” project (OLAF-JRC SI2.601156 and SI2.710969), funded under the Hercule III Programme. In OLAF, we would like to thank in particular Juergen Marke, Nicholas Shaw and Winfried Kleinegriss who, since the early days of AMT, have firmly believed in the idea of estimating trade prices using the Forward Search and similar data monitoring tools.

A first preliminary version of the SAS macros for the Forward Search were developed in 2016 by Dr Jos Polifet under the guidance of the authors of this report, in the framework of a consultancy activity requested by the JRC to SAS Belgium (Hertenbergstraat 6, 3080 Tervuren, Belgium). Compared to the earlier SAS macros, which are available at <https://github.com/JosPolfliet/FSDA-SAS>, our new package contains many enrichments and bug fixes.

The package is now used in the AMT for the routine generation and dissemination of the EU import prices in THESEUS, and in Web-Ariadne for facilitating the use of robust regression methods by data analysts and practitioners: the deployment of the SAS macros in the two JRC resources is respectively by Giuseppe Sgarlata and Emmanuele Sordini.

Authors

Francesca Torti and Domenico Perrotta, European Commission, Joint Research Centre (JRC)
Anthony C. Atkinson, London School of Economics, UK
Aldo Corbellini and Marco Riani, University of Parma, Italy

1. Introduction

Automated monitoring of external trade data is a tactical operational activity by the Anti Fraud Office (OLAF) of the European Commission (EC) in support of its own investigation Directorate and its partners in the EU Member States Customs. The Automated Monitoring Tool (AMT) is a sequence of projects financed by administrative agreements between OLAF and the EC Joint Research Centre in support of that activity. A key objective of AMT is the systematic estimation of trade prices and statistical detection of patterns of anti-fraud relevance in large volumes of trade and other relevant data. The JRC and its academic partners, represented in this report, work together on the developed of instruments for this purpose. This report focuses on robust regression tools that are at the core of the AMT.

The forward search (FS) is a general method of robust data fitting that moves smoothly from very robust to maximum likelihood estimation. The FS procedures are included in the MATLAB toolbox FSDA. The work on a SAS version of the FS, presented in this paper, is in the framework of an European Union program supporting the Customs Union and Anti-Fraud policies. It originates from the need for the analysis of large data sets expressed by law enforcement services operating in the European Union (the EU anti-fraud office in particular) that are already using our SAS software for detecting data anomalies that may point to fraudulent customs returns. For them, the library is also accessible through a restricted web platform called Web-Ariadne: <https://webariadne.jrc.ec.europa.eu>.

The series of fits provided by the FS is combined with an automatic procedure for outlier detection that leads to the adaptive data-dependent choice of highly efficient robust estimates. It also allows monitoring of residuals and parameter estimates for fits of differing robustness. Linking plots of such quantities, combined with brushing, provides a set of powerful tools for understanding the properties of data including anomalous structures and data points. Our SAS package extends this ideas of monitoring to several traditional robust estimators of regression for a range of values of their key parameters (maximum possible breakdown or nominal efficiency). We again obtain data adaptive values for these parameters and provide a variety of plots linked through brushing. Examples in the paper are for S estimation and for Least Median of Squares (LMS) and Least Trimmed Squares (LTS) regression.

In the next section we define three classes of robust estimators (downweighting, hard trimming and adaptive hard trimming) of all of which occur in the numerical examples. Algebra for the FS is in Section 3. Sections 4 and 5 describe general procedures for outlier detection and the rule to control the statistical size of the procedure to allow for testing at each step of the FS. These procedures are illustrated in §6 by analysis of data on 509 bank customers. The next two sections are specific to our SAS implementation: in §7 we describe the properties of the language that make it suitable for handling large datasets and list the procedures that we have implemented; §8 describes the approximations used to provide fast analyses of large datasets. As Figure 7 shows, there is a considerable advantage in using SAS instead of MATLAB functions for analysing large datasets.

The data analysed in §6 have been transformed to approximate normality by the Box-Cox transformation (Box and Cox, 1964). In §10 we illustrate the use of our SAS routines, including the "fan plot", to monitor a series of robust fits and establish the value of the transformation parameter in the presence of data contamination. Section 11 provides background for soft-trimming estimation, in our case M- and S-estimators. Examples of the use of our SAS routines for S, LMS and LTS regression are in §12 for a trade dataset that has a more complicated structure than the loyalty card data. Section 13 concludes. The supplementary material reported in Annex B illustrates the use of our SAS software by European Union services, for anti-fraud purposes.

2. Three Classes of Estimator for Robust Regression

We work with the customary null regression model in which the n univariate response variables y_i are related to the values of a set of p explanatory variables x by the relationship

$$y_i = \beta^\top x_i + \epsilon_i \quad i = 1, \dots, n, \quad (1)$$

including an intercept term. The independent errors ϵ_i have constant variance σ^2 . The purpose of robust estimation is to find good estimators of the parameters when there are departures from this model, typically caused by outliers, which should also be identified.

It is helpful to divide methods of robust regression into three classes.

1. Hard (0,1) Trimming. In Least Trimmed Squares (LTS: Hampel, 1975, Rousseeuw, 1984) the amount of trimming is determined by the choice of the trimming parameter h , $[n/2] + [(p+1)/2] \leq h \leq n$,

which is specified in advance. The LTS estimate is intended to minimize the sum of squares of the residuals of h observations. For LS, $h = n$. In the generalization of Least Median of Squares (LMS, Rousseeuw, 1984) that we monitor, the estimate minimizes the median of h squared residuals.

2. Adaptive Hard Trimming. In the Forward Search (FS), the observations are again hard trimmed, but the value of h is determined by the data, being found adaptively by the search. Data analysis starts from a very robust fit to a few, carefully selected, observations found by LMS or LTS with the minimum value of h . The number of observations used in fitting then increases until all are included. (See Atkinson and Riani, 2000 and Riani *et al.*, 2014c for regression, Atkinson *et al.*, 2010 for a general survey of the FS, with discussion, and Cerioli *et al.*, 2014 for results on consistency).
3. Soft trimming (downweighting). M estimation and derived methods. The intention is that observations near the centre of the distribution retain their value, but the ρ function ensures that increasingly remote observations have a weight that decreases with distance from the centre.

We shall consider all three classes of estimator. The FS by its nature provides a series of decreasingly robust fits which we monitor for outliers in order to determine how to increment the subset of observations used in fitting. For LTS and LMS we fit the regression model to increasing sized subsets h . For S estimation, which we use as our example of soft trimming, we look at fits as the breakdown point varies. Here our focus is on SAS programs.

3. Algebra for the Forward Search

Examples and a discussion of monitoring using the MATLAB version of FSDA are in Riani *et al.* (2014a). To describe the SAS procedures which are the subject of this paper, needs fuller details than are given there.

It is convenient to rewrite the regression model Eq. (1) in matrix form as $y = X\beta + \epsilon$, where 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 least squares estimator of β is $\hat{\beta}$. Then the vector of n least squares residuals is $e = y - \hat{y} = y - X\hat{\beta} = (I - H)y$, 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 search fits subsets of observations of size m to the data, with $m_0 \leq m \leq n$. Let $S^*(m)$ be the subset of size m found by the forward search, for which the matrix of regressors is $X(m)$. Least squares on this subset of observations yields 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

$$e_i(m) = y_i - x_i^T \hat{\beta}(m). \quad (2)$$

The search moves forward with the augmented subset $S^*(m + 1)$ consisting of the observations with the $m + 1$ smallest absolute values of $e_i(m)$. In the batch algorithm of §8 we explore the properties of a faster algorithm in which we move forward by including $k > 1$ observations.

To start we take $m_0 = p$ and search over subsets of p observations to find the subset that yields the LMS estimate of β . However, this initial estimator is not important, provided masking is broken. Our computational experience for regression is that randomly selected starting subsets also yield indistinguishable results over the last one third of the search, unless there is a large number of structured outliers.

To test for outliers the deletion residual is calculated for the $n - m$ observations not in $S^*(m)$. These residuals, which form the maximum likelihood tests for the outlyingness of individual observations, are

$$r_i(m) = \frac{y_i - x_i^T \hat{\beta}(m)}{\sqrt{s^2(m)\{1 + h_i(m)\}}} = \frac{e_i(m)}{\sqrt{s^2(m)\{1 + h_i(m)\}}}, \quad (3)$$

where the leverage $h_i(m) = x_i^T \{X(m)^T X(m)\}^{-1} x_i$. Let the observation nearest to those forming $S^*(m)$ be i_{\min} where

$$i_{\min} = \arg \min_{i \notin S^*(m)} |r_i(m)|.$$

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

$$r_{\min}(m) = \frac{e_{i_{\min}}(m)}{\sqrt{s^2(m)\{1 + h_{i_{\min}}(m)\}}}, \quad (4)$$

as a test statistic. If the absolute value of (4) is too large, the observation i_{\min} is considered to be an outlier, as well as all other observations not in $S^*(m)$.

4. Testing for Outliers

The test statistic (4) is the $(m+1)$ st ordered value of the absolute deletion residuals. We can therefore use distributional results to obtain envelopes for our plots. The argument parallels that of Riani *et al.* (2009) where envelopes were required for the Mahalanobis distances arising in applying the FS to multivariate data.

Let $Y_{[m+1]}$ be the $(m+1)$ st order statistic from a sample of size n from a univariate distribution with c.d.f. $G(y)$. Then the c.d.f. of $Y_{[m+1]}$ is given exactly by

$$P\{Y_{[m+1]} \leq y\} = \sum_{j=m+1}^n \binom{n}{j} \{G(y)\}^j \{1 - G(y)\}^{n-j}. \quad (5)$$

See, for example, Lehmann (1991, p. 353). We then apply properties of the beta distribution to the RHS of (5) to obtain

$$P\{Y_{[m+1]} \leq y\} = I_{G(y)}(m+1, n-m), \quad (6)$$

where $I_p(A, B)$ is the incomplete beta integral. From the relationship between the F and the beta distribution equation (6) becomes

$$P\{Y_{[m+1]} \leq y\} = P\left\{F_{2(n-m), 2(m+1)} > \frac{1 - G(y)}{G(y)} \frac{m+1}{n-m}\right\}, \quad (7)$$

where $F_{2(n-m), 2(m+1)}$ is the F distribution with $2(n-m)$ and $2(m+1)$ degrees of freedom (Guenther, 1977). Thus, the required quantile of order γ of the distribution of $Y_{[m+1]}$, say $y_{m+1, n; \gamma}$, is obtained as

$$y_{m+1, n; \gamma} = G^{-1}(q) = G^{-1}\left(\frac{m+1}{m+1 + (n-m)x_{2(n-m), 2(m+1); 1-\gamma}}\right), \quad (8)$$

where $x_{2(n-m), 2(m+1); 1-\gamma}$ is the quantile of order $1-\gamma$ of the F distribution with $2(n-m)$ and $2(m+1)$ degrees of freedom.

In our case we are considering the absolute values of the deletion residuals. If the c.d.f. of the t distribution on ν degrees of freedom is written as $T_\nu(y)$, the absolute value has the c.d.f.

$$G(y) = 2T_\nu(y) - 1, \quad 0 \leq y < \infty. \quad (9)$$

The required quantile of $Y_{[m+1]}$ is given by

$$y_{m+1, n; \gamma} = T_{m-p}^{-1}\{0.5(1+q)\},$$

where q is defined in Eq. (8). To obtain the required quantile we call an inverse of the F and then an inverse of the t distribution.

If we had an unbiased estimator of σ^2 the envelopes would be given by $y_{m+1, n; \gamma}$ for $m = m_0, \dots, n-1$. However, the estimator $s^2(m^*)$ is based on the central m observations from a normal sample – strictly the m observations with smallest squared residuals based on the parameter estimates from $S^*(m-1)$. The variance of the truncated normal distribution containing the central m/n portion of the full distribution is.

$$\sigma_T^2(m) = 1 - \frac{2n}{m} \Phi^{-1}\left(\frac{n+m}{2n}\right) \phi\left\{\Phi^{-1}\left(\frac{n+m}{2n}\right)\right\}, \quad (10)$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are respectively the standard normal density and c.d.f. See, for example, Johnson *et al.* (1994, pp. 156-162) and Riani *et al.* (2009) for a derivation from the general method of Tallis (1963). Since the outlier tests we are monitoring are divided by an estimate of σ^2 that is too small, we need to scale up the values of the order statistics to obtain the envelopes

$$y_{m+1, n; \gamma}^* = y_{m+1, n; \gamma} / \sigma_T(m).$$

To be specific, in the case of the 99% envelope $\gamma = 0.99$, corresponds to a nominal pointwise size $\alpha = 1 - \gamma$ which is equal to 1%. We expect, for the particular step m which is considered, to find exceedances of the quantile in a fraction 1% of the samples under the null normal distribution. We however require a samplewise probability of 1% of the false detection of outliers, that is over all values of m considered in the search. The algorithm in the next section is accordingly designed to have a size of 1%.

5. Regression Outlier Detection in the FS

We have to find appropriate bounds for the outlier test Eq. (4). For efficient parameter estimation we want to use as many observations as possible. However, we wish to avoid biased estimation due to the inclusion of outliers. We therefore need to control the size of the test. Because we are testing for the existence of an outlier at each step of the search we have to allow for the effect of simultaneous testing. Atkinson and Riani (2006) adapt and extend a sophisticated simulation method of Buja and Rolke (2003) to show how severe this problem can be. For example, for a nominal pointwise significance level of 5%, the probability of observing at least one outlier in the null case of no outliers is 55.2% when $n = 100$, $p = 3$ and outliers are only sought in the last half of the search. Even if an outlier is only declared when 3 successive values lie above the pointwise boundary, the size of the test is 23.2%.

If there are a few large outliers they will enter at the end of the search, as in Figure 2, and their detection is not a problem. However, even relatively small numbers of outliers can be more difficult to identify and may cause a peak in the centre of the search. Masking may then cause the plot to return inside the envelopes at the end of the search. Methods of using the FS for the formal detection of outliers have to be sensitive to these two patterns: a few "obvious" outliers at the end and a peak earlier in the search caused by a cluster of outliers.

To use the envelopes in the forward search for outlier detection we accordingly propose a two-stage process. In the first stage we run a search on the data, monitoring the bounds for all n observations until we obtain a "signal" indicating that observation m^\dagger , and therefore succeeding observations, may be outliers, because the value of the statistic lies beyond our threshold. In the second part we superimpose envelopes for values of n from this point until the first time we introduce an observation we recognize as an outlier. The conventional envelopes shown, for example, in the top panel of Figure 2, consist roughly of two parts; a flat "central" part and a steeply curving "final" part. Our procedure FS for the detection of a "signal" takes account of these two parts and is similar to the rule used by Riani *et al.* (2009) for the detection of multivariate outliers. In our definition of the detection rule we use the nomenclature $r_{\min}(m, n^*)$ to denote that we are comparing the value of $r_{\min}(m)$ with envelopes from a sample of size n^* .

1. Detection of a Signal

There are four conditions, the fulfillment of any one of which leads to the detection of a signal.

- In the central part of the search we require 3 consecutive values of $r_{\min}(m, n)$ above the 99.99% envelope or 1 above 99.99%;
- In the final part of the search we need two consecutive values of $r_{\min}(m, n)$ above 99.9% and 1 above 99%;
- $r_{\min}(n - 2, n) > 99.9\%$ envelope;
- $r_{\min}(n - 1, n) > 99\%$ envelope. In this case a single outlier is detected and the procedure terminates.

The final part of the search is defined as:

$$m \geq n - \lceil 13(n/200)^{0.5} \rceil,$$

where here $\lceil \cdot \rceil$ stands for rounded integer. For $n = 200$ the value is slightly greater than 6% of the observations.

2. Confirmation of a Signal

The purpose of, in particular, the first point is to distinguish informative peaks from random fluctuations in the centre of the search. Once a signal takes place (at $m = m^\dagger$) we check whether the signal is informative about the structure of the data. If $r_{\min}(m^\dagger, m^\dagger) < 1\%$ envelope, we decide the signal is not informative, increment m and return to Step 1.

3. Identification of Outliers

With an informative signal we start superimposing 99% envelopes taking $n^* = m^\dagger - 1, m^\dagger, m^\dagger + 1, \dots$ until the final, penultimate or ante-penultimate value are above the 99% threshold or, alternatively, we have a value of $r_{\min}(m, n^*)$ for any $m > m^\dagger$ which is greater than the 99.9% threshold. Let this value be m^+ . We then obtain the best parameter estimates by using the sample of size $m^+ - 1$.

Automatic use of $m = m^+ - 1$ is programmed in our SAS routines. It is also central to the comparisons involving the batch method of §8.

```

/* SAS working library and data matrix creation */
libname lib "C:\FSDA\data\regression";
use ("lib.loyalty");
read all var {'x1' 'x2' 'x3'} into x[colname=colnx];
read all var 'y' into y[colname=colny];
close ("lib.loyalty");
/* Add constant variable to the data for model with intercept */
x = x || j(nrow(x),1,1);

```

Figure 1: Example of SAS IML Studio code which uploads the Loyalty card data in SAS IML.

6. FS Analysis of the Transformed Loyalty Card Data

The data example we use in virtually all the calculations in this paper, taken from Atkinson and Riani (2006), is of 509 observations on the behavior of customers with loyalty cards from a supermarket chain in Northern Italy. The data are themselves a random sample from a larger database. The sample of 509 observations is part of the FSDA toolbox for MATLAB. The response is the amount, in euros, spent at the shop over six months and the explanatory variables are: x_1 , the number of visits to the supermarket in the six-month period; x_2 , the age of the customer and, x_3 , the number of members of the customer's family. The data are loaded in SAS IML with the commands reported in Figure 1.

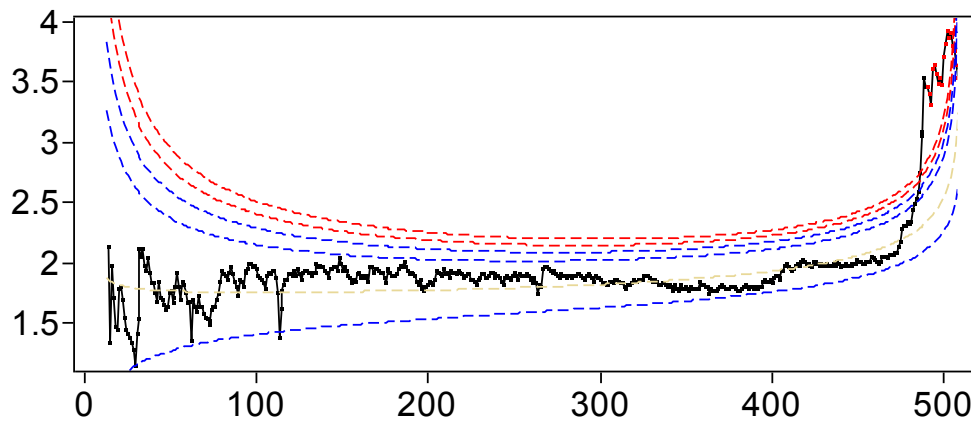
Atkinson and Riani (2006) show that the data need transformation to achieve constant variance for which purpose we use the Box-Cox power transformation. As we see in §10 a value of 0.4 is indicated, and we work with this transformation for the rest of this section.

Figure 2 shows, in the top panel, a forward plot of absolute minimum deletion residuals for observations not in the subset used in fitting. In addition to the residuals, the plot includes a series of pointwise percentage levels for the residuals (at 1%, 50%, 99%, 99.9%, 99.99% and 99.999%) found by the order statistic arguments of §4. Several large residuals occur towards the end of the search. These are identified by an automatic procedure including the resuperimposition of envelopes described at the end of §5. In all 18 outliers (plotted as red crosses in the .pdf version), are identified. These form the last observations to enter the subset in the search. The figure shows that, at the very end of the search, the trajectory of residuals returns inside the envelopes, the result of masking. As a consequence, the outliers would not be detected by the deletion of single observations from the fit to all n observations.

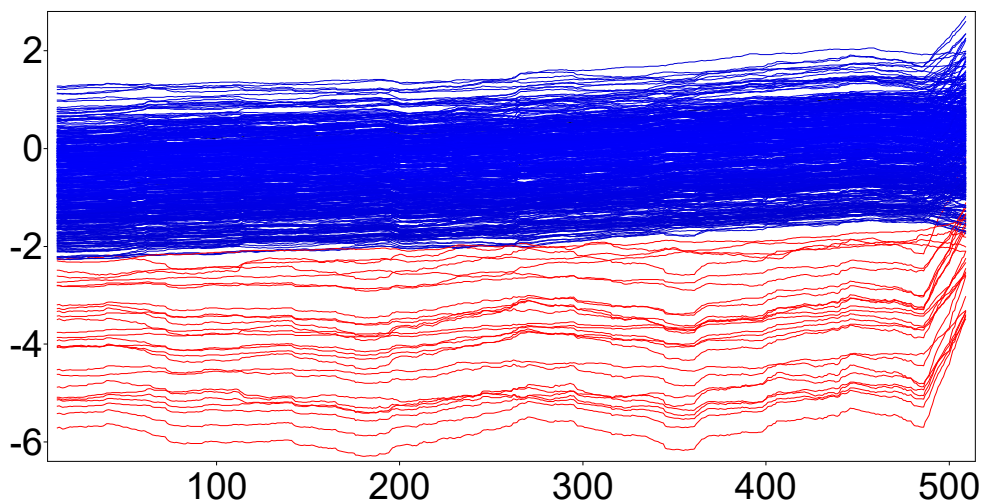
These results are very stable once a subset of non-outlying observations has been achieved. Figure 2 also shows, in the bottom panel, the monitoring of scaled residuals during the search with the 18 outliers shown in red (in the on-line .pdf version). The outliers all have negative residuals the values of which change little during the search until the end, when the outliers start to enter the subset. Then the residuals for the outliers decrease steadily in magnitude. At the same time, the residuals for some of the observations from the main body of the data begin to increase. The plots in Figures 2-4 were produced by brushing, that is selecting the observations of interest from the top panel of Figure 2 and highlighting them in all others.

The observations we have found are outlying in an interesting way, especially for the values of x_1 . Figure 3 shows the scatterplots of y against the three explanatory variables, with brushing used to highlight the outlying observations in red (in the on-line .pdf version). The first panel is of y against x_1 . The FS has identified a subset of individuals, most of whom are behaving in a strikingly different way from the majority of the population. They appear to form a group who spend less than would be expected from the frequency of their visits. The scatterplots for x_2 and x_3 , on the other hand, do not show any distinct pattern of outliers.

The effect of the 18 outliers on inference can be seen in Figure 3 which gives forward plots of the four parameter estimates, again with brushing used to plot the outlying observations in red (in the on-line .pdf version). The upper left panel, that for $\hat{\beta}_1$, is the most dramatic. As the outliers are introduced, the estimate decreases rapidly in a seemingly linear manner. This behaviour reflects the inclusion of the outliers in Figure 3, all of which lie below the general linear structure: their inclusion causes $\hat{\beta}_1$ to decrease. The outliers also have effects on the other three parameter estimates (the lower right panel is for the estimate of the intercept β_0). Although the effects for these three parameters are appreciable, they do not take any of the estimates outside the range of values which was found before the inclusion of outliers. The group of outliers who are spending less than would be expected, and which does not agree with the model for the majority of the data, will be important in any further modelling.



```
CALL FSR("lib.loyalty", {'x1' 'x2' 'x3'}, "y", "CLASSIFY mdrplot ")
transform_original_data = 0.4 ;
```



```
CALL FSR("lib.loyalty", {'x1' 'x2' 'x3'}, "y", "CLASSIFY RESFWDPLOT ")
transform_original_data = 0.4 ;
```

Figure 2: Loyalty card data: monitoring plots on for transformed data with $\lambda = 0.4$. The top panel shows the absolute values of minimum deletion residuals among observations not in the subset; the last part of the curve, corresponding to the 18 identified outliers, is automatically highlighted in red (in the on-line .pdf version). The bottom panel shows the scaled residuals, with the trajectories corresponding to the 18 detected outliers automatically represented in red (in the on-line .pdf version). The box under each panel contains the SAS code used to generate the plot.

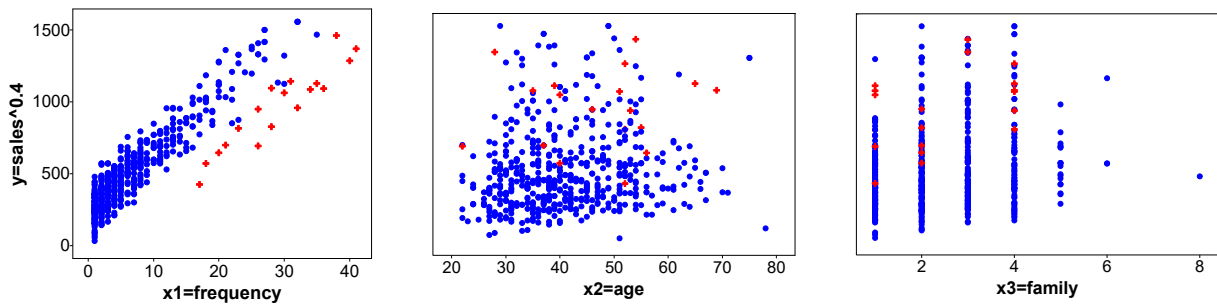
7. Why SAS?

The statistical community currently has three main environments for program development, which target rather different market segments.

The R environment¹ is the most popular among statisticians and offers many packages for robust statistics, for example `rrcov` for multivariate analysis (Todorov and Filzmoser, 2009) and `robustbase` for regression, univariate and multivariate analysis (Rousseeuw *et al.*, 2009). Recently Riani *et al.* (2017) have developed `FSDAr` for regression analysis.

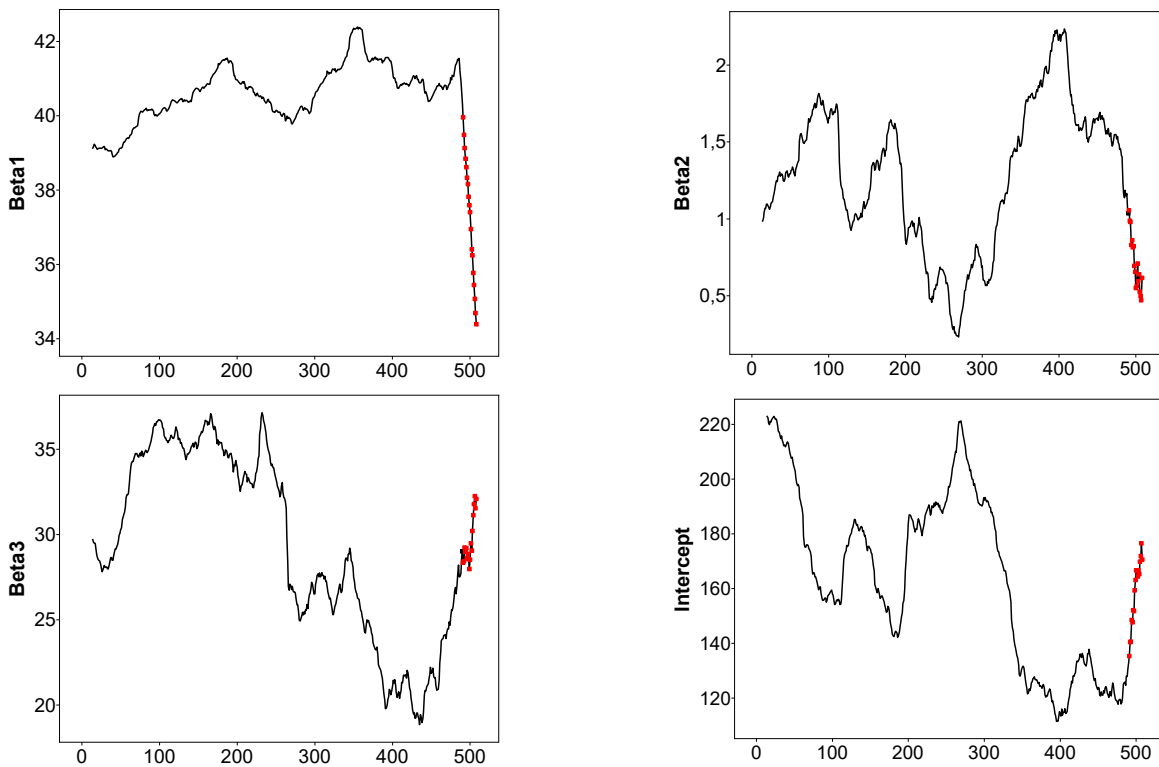
Engineers and practitioners in physics, geology, transport, bioinformatics, vision and other fields usually prefer MATLAB, but find in the default distribution only a few robust tools, such as the Minimum Covariance Determinant (MCD, Hubert and Debruyne, 2010, introduced in the 2016 release through function `robustcov`) and robust regression computed via iteratively re-weighted least squares (functions `robustfit` and `fitlm`). Many more robust procedures are provided by two open toolboxes: Library for Robust Analysis

¹R is available from the CRAN website: <https://cran.r-project.org/>



```
CALL FSR("lib.loyalty", {'x1' 'x2' 'x3'}, "y", "CLASSIFY scatter")
transform_original_data = 0.4 ;
```

Figure 3: Loyalty card data: scatterplots of transformed data when $\lambda = 0.4$, with the 18 outliers detected plotted as red crosses (in the on-line .pdf version). The box under the figure contains the SAS code used to generate the plots.



```
CALL FSR("lib.loyalty", {'x1' 'x2' 'x3'}, "y", "CLASSIFY BETA_PLOT")
transform_original_data = 0.4 ;
```

Figure 4: Loyalty card data: monitoring of estimated beta coefficients on transformed data when $\lambda = 0.4$, with the part of the trajectory corresponding to the 18 detected outliers, highlighted in red (in the on-line .pdf version). The box under the figure contains the SAS code used to generate the plots.

(LIBRA² Vanden Branden and Hubert, 2005; Verboven and Hubert, 2010) and Flexible Statistics for Data Analysis (FSDA³ Perrotta *et al.*, 2009; Riani *et al.*, 2012).

LIBRA addresses robust Principal Component Analysis, robust Partial Least Squares regression and robust Principal Component Regression (Hubert *et al.*, 2005), classification and depth-based methods. FSDA includes robust clustering (García-Escudero *et al.*, 2008, 2010), S, MM (Maronna *et al.*, 2006) and MVE (Van Aelst and Rousseeuw, 2009) estimators, and tools for monitoring a number of traditional robust multivariate and regression estimators for various choices of breakdown or efficiency, along with the FS approach (Riani

²LIBRA website: <https://wis.kuleuven.be/stat/robust/LIBRA>

³FSDA website: <http://www.riani.it/MATLAB.htm> and <http://fsda.jrc.ec.europa.eu>

et al., 2014a). Both toolboxes offer functions for Least Trimmed Squares (LTS) (Rousseeuw, 1984), MCD and M estimation Maronna *et al.* (2006).

SAS is widely used by large commercial and public organizations. The work reported in our paper originates from needs expressed by law enforcement services operating in the European Union (the EU anti-fraud office in particular) that are using our software for detecting anomalies that may point to fraud. Such services could find in the standard SAS distribution only a procedure for robust regression, PROC ROBUSTREG, which depends on IML modules for LTS, S, M and MM estimators. No equivalent code exists for multivariate methods. Our package provides the SAS community with methods of monitoring these estimators and their multivariate counterparts, as in the MATLAB FSDA, and also a full set of methods for the Forward Search.

The idea of monitoring an estimator for various values of its key parameters has shown great potential in data analysis, but the method can be time and space consuming, as the statistics of interest have to be computed and stored many times. This is particularly true for the Forward Search that, for monitoring statistics at each subset size, requires approximately n^2 elements to store regression residuals or Mahalanobis distances for a dataset of size n . This means, for example, that almost 1 Gigabyte of RAM would be necessary to store a structure for $n = 11,000$ observations (each numeric variable typically requires 8 bytes). SAS is known for its superior capacity of treating such large datasets. There are several ingredients behind this capacity improvement.

1. When the data are at the limit of the physical memory, caching strategies become crucial to avoid the deterioration of performance. Unlike other statistical environments that only run in memory and crash when a data set is too large to be loaded, SAS uses file-swapping to handle out-of-memory problems. The swapping is very efficient, as the SAS procedures are optimized to limit the number of files created within a procedure, avoiding unnecessary swapping steps.
2. File records are stored sequentially, in such a way that processing happens one record at a time. Then, the SAS data step reads through the data only one time and applies all the commands to each line of data of interest. In this way, data movements are drastically limited and processing time is reduced.
3. A data step only reads the data that it needs in the memory and leaves out the data that it does not need in the source.
4. Furthermore, data are indexed to allow for faster retrieval from datasets.
5. Finally, in regression and other predictive modeling methods, multi-threading is applied whenever this is appropriate for the analysis.

These good nominal properties seem confirmed by the computing time assessments presented in the next section, showing that our SAS implementation of robust regression tools outperforms the MATLAB counterpart for datasets with more than 1,000 units (see Figure 7).

SAS has therefore excellent scalability properties, but unfortunately the standard distribution does not provide graphical interactivity for exploratory data analysis and satisfactory graphical output. We have used a separate library based on the IML language (SAS/IML Studio) to realize in SAS a number of FSDA functions requiring advanced graphical output and interactivity:

- `FSR.sx` and `FSM.sx`, which implement the FS approach to detect outliers respectively in regression and in multivariate data,
- `FSRfan.sx` and `FSMfan.sx` for identifying the best transformation parameter for the Box-Cox transformation in regression and multivariate analysis (as shown in Section 10),
- `Monitoring.sx` for monitoring a number of traditional robust multivariate and regression estimators (S, MM, LTS and LMS), already present in SAS, for specific choices of breakdown point or efficiency. Riani *et al.* (2014a) introduced the monitoring of regression estimators detailed in Section 11, but, in the FSDA toolbox, only for S and MM estimators (and the FS) in MATLAB. The extension to monitoring of LTS and LMS is a particularly powerful new feature and a novelty in the statistical literature;

Finally, we have modified the standard LTS and LMS IML functions by introducing the small sample correction factor of Pison (Pison *et al.*, 2002) and by increasing the range of values of the trimming parameter h in LTS.

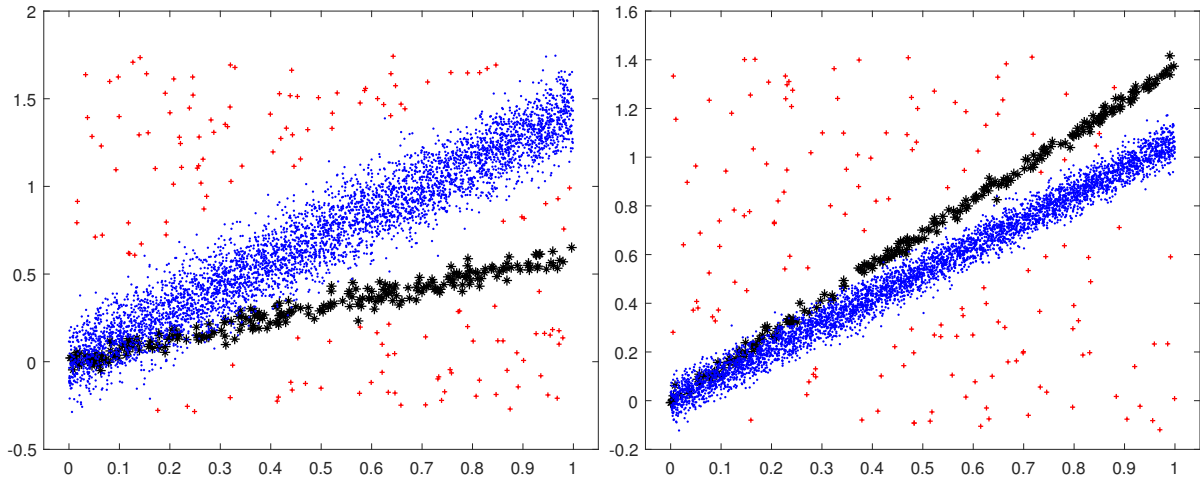


Figure 5: Two artificial datasets generated with MixSim for the assessment.

8. The FS batch procedure

Our SAS library contains a new Forward Search strategy that increases the possibility to treat large datasets. The idea is to reduce the size of the output tables and the amount of memory required through a batch updating procedure.

The standard FS algorithm in §3 produces a sequence of $n - m_0$ subsets with corresponding model parameters and relevant test statistics, used typically to test the presence of outliers. The initial subset size m_0 can be as small as p , the minimum number of observations necessary to provide a fit to the data. In the standard algorithm the subset size $m_0 \leq m \leq n$ is increased by one unit at a time and only the smallest value of the test statistic among the observations outside the subset is retained. The batch version of the algorithm fits instead only one subset every $k > 1$ steps. The value of k is set by the user through the input parameter `fs_steps`. The number of subsets to be evaluated therefore reduces to $(n - k)/k$. For each subset and set of estimated model parameters, *the k smallest values of the test statistic are retained*: they are assigned to the current step and to the preceding $k - 1$, in order to obtain the complete vector of minimum test statistics Eq. (4) to compare with the envelopes. Of course this vector is an approximation to the real one which would be found by evaluating each of the k steps individually; the approximation is the cost of reducing the number of fits to $(n - k)/k$ while still applying the signal detection, signal validation and envelope superimposition phases described in §5 at each of the $n - m_0$ FS steps.

If the data are contaminated and k is too large, this approach may not be accurate enough to detect the outliers, giving rise to biased estimates. The problem can be appraised by monitoring the statistical properties of the batch algorithm for increasing k . We have conducted such exploratory assessment using artificial data.

We generated the data using MixSim (Maitra and Melnykov, 2010) in the MATLAB implementation of the FSDA toolbox (Torti *et al.*, 2018, Section 3); the functions used are `MixSimreg.m` and `simdataset.m`. MixSim allows generation of data from a mixture of linear models on the basis of an average overlap measure $\bar{\omega}$ pre-specified by the user. We generated a dominant linear component containing 95% of the data and a 5% “contaminating” one with small average overlap ($\bar{\omega} = 0.01$). The generating regression model is without intercept, with random slopes from a Uniform distribution between $\tan(\frac{\pi}{6}) = \frac{\sqrt{3}}{3}$ and $\tan(\frac{\pi}{3}) = \sqrt{3}$, and independent variables from a Uniform distribution in the interval $[0, 1]$. Each slope is equally likely to be that of the dominant component. We took the error variances in the two components to be equal when specification of the value of $\bar{\omega}$, together with the values of the slopes, defines the error variance for each sample. We also added additional uniform contamination of 3% of the above data over the rectangle defined by the two slopes and the range of the two independent variables. The plots in Figure 5 are examples of two datasets with $4750 + 250 + 150$ units.

The boxplots of Figure 6 show the bias for the slope and intercept obtained from 500 such datasets with 5, 150 observations each, for $k \in \{1, 5, 10, 15, 20, 40, 60, 80, 100\}$. The bias here is simply the difference between the estimated and real slopes, the latter referring to the dominant generating component.

The upper panel of the figure shows that the median bias for both the slopes and intercepts are virtually zero. The dispersion of the estimates for slopes and intercept remain both stable and quite small even for values of k approaching 100 (note that the boxplot whiskers are in $[-0.01, 0.01]$). However, the variability of the estimates outside the whiskers rapidly increases for $k = 100$. The fact that the bottom and top edges

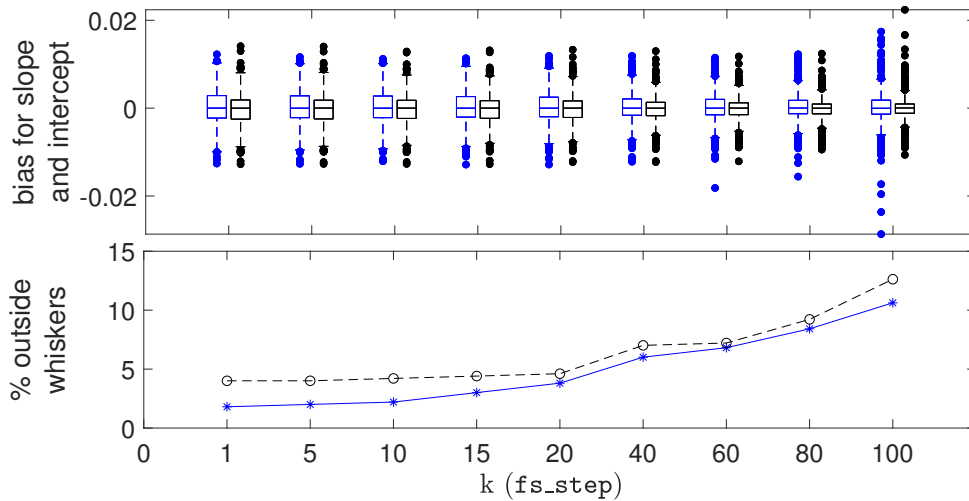


Figure 6: Top panel. Boxplots showing, for different values of k (the `fs_step` parameter, on the x-axis), the bias and dispersion of the estimated slopes and intercepts (respectively from left to right for each k). The estimates are obtained from 500 simulated datasets of 5,150 observations. Bottom panel: percentage of estimated values lying outside the boxplot whiskers for slope (blue asterisks) and intercept (black circles).

of the box seem even to become smaller for increasing k , may be interpreted as a reduced capacity of the batch FS to capture the fine grained structure of the data when k is too large.

The stability of the batch procedure can also be appreciated by looking, in the bottom panel of the figure, at the number of estimated slopes/intercepts outside the boxplot whiskers: up to $k = 10$ there is no appreciable increase with respect to the standard FS with $k = 1$; between $k = 10$ and $k = 20$ the increase is still contained to 5%; then the number of bad estimates rapidly increases exceeding 10%. Finally, there is no evidence of major failure of the batch FS to reject outliers, which would be shown by occasional large values of bias.

9. Timing comparisons

We now describe the results of an assessment of the computational benefit of the new batch Forward Search approach available only in SAS, in comparison with the standard SAS and FSDA MATLAB implementations. We tested the functions on a workstation with a CPU 2 x Xeon E5-262v4 (2.6GHz 4cores), two RAM of 32GB DDR4 2400 ECC, and a Disk SSD of 512GB, equipped with MATLAB R2018b and SAS 9.4.

Figure 7 shows the elapsed time needed for analysing simulated datasets of different sizes (from 30 to 100,000), when fitting one explanatory variable. The results are split into three panels for small ($n = 30, \dots, 1,000$), medium ($n = 2,000, \dots, 15,000$) and large data sizes ($n = 20,000, \dots, 100,000$). The bottom-right panel gives the ratio between the time required by the MATLAB implementation and the two SAS ones. For small samples, the FSDA MATLAB implementation (orange squares) is faster than the standard SAS implementation (blue diamonds), but there is a crossing point at a sample size between $n = 800$ and $n = 900$ where the latter starts to perform better. The advantage of using the SAS function increases for larger sample sizes. For example in a sample of 50,000 observations SAS is about 7 times faster. The batch option in SAS (red circles), with $k = 20$, is even faster: 12 times faster in a sample of 50,000 observations; note that in Figure 7 the batch results are reported only for $n \geq 20,000$, because the computational benefit for smaller n values would not “compensate” the loss in statistical accuracy due to the approximate batch solution.

The bottom-left panel shows that the standard SAS and FSDA MATLAB implementations crash (because of memory limits) when the sample sizes exceed 50,000 observations. Only the SAS batch algorithm seems to cope with larger datasets ($n = 100,000$ in the figure), which however, requires about 3.5 hours to terminate.

Finally, by interpolating the time values in the three cases with a quadratic curve – the time complexity for producing n statistics for n steps is expected to be $\mathcal{O}(n^2)$ – we found the following approximate coefficients for the quadratic terms: $1.23 \cdot 10^{-5}$ for the MATLAB implementation, $1.98 \cdot 10^{-6}$ for the SAS standard implementation, $7.17 \cdot 10^{-7}$ for the SAS batch implementation (the last one fitted on all n values, not reported in the Figure). This ranking might be used to extrapolate the computational performances of the three FSR implementations for n values not considered here, on hardware configurations that can cope with

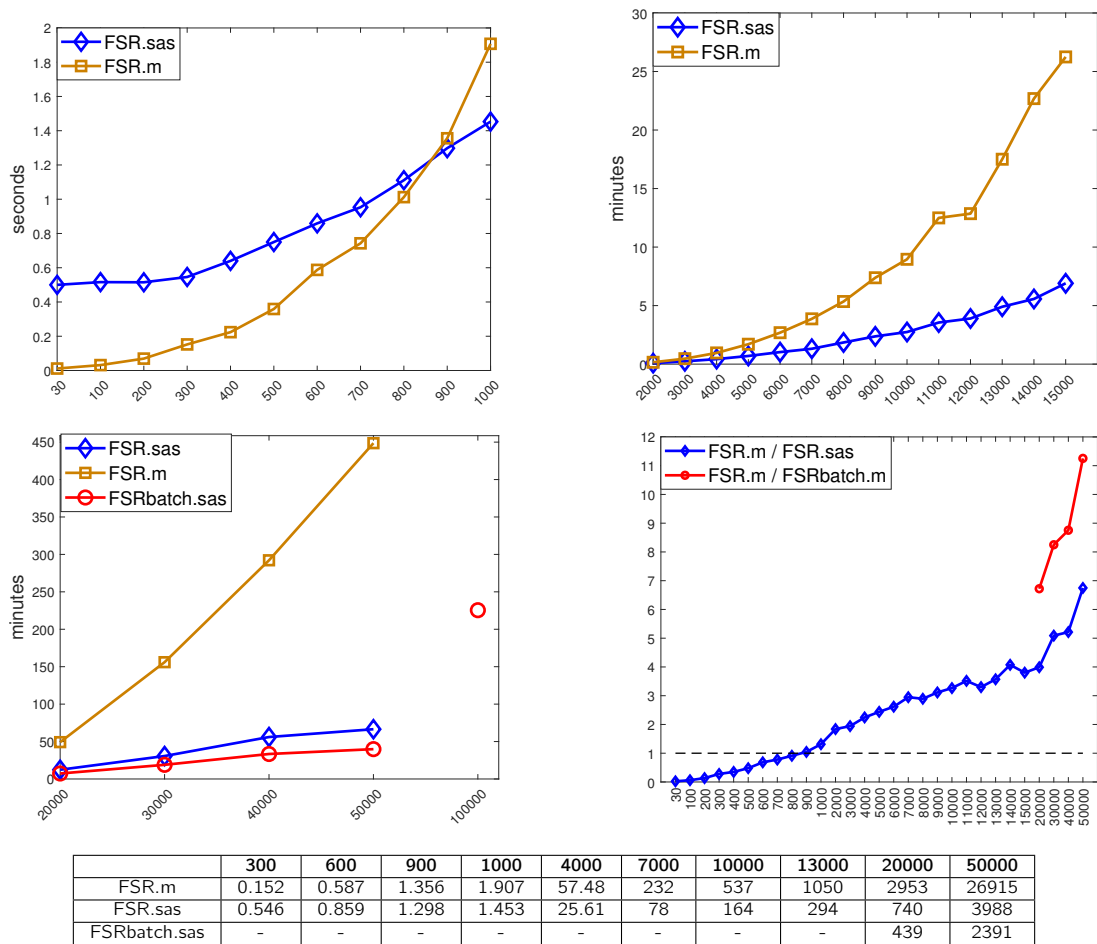


Figure 7: Execution time of our SAS (R9.4) and MATLAB (R2018b) implementations of the FSR function; for SAS, the comparison is also with the batch version of FSR (with $k = 10$). The assessment covers data with one explanatory variable and size ranging from 30 to 100,000. Results are split into three panels for small, medium and large data sizes. The last, bottom-right, panel gives the ratio between the time required by the MATLAB implementation and the two SAS ones. The associated table reports the time in seconds for selected sample sizes.

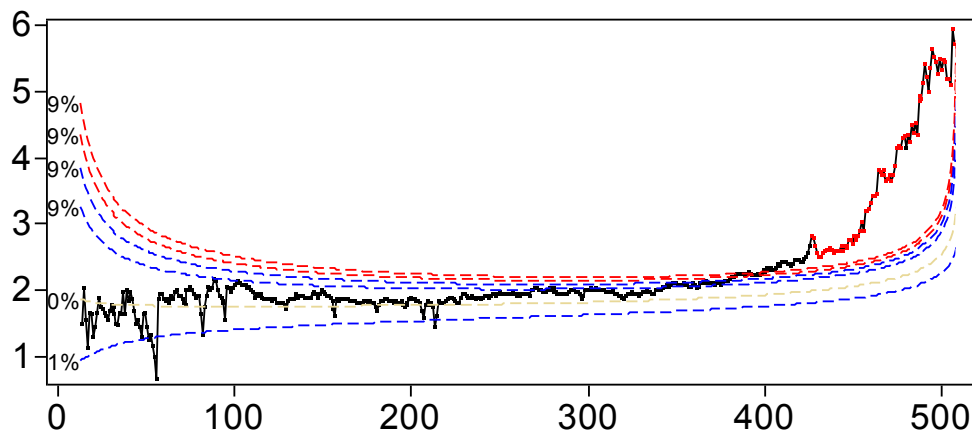
larger data structures.

10. Transformation of the Response

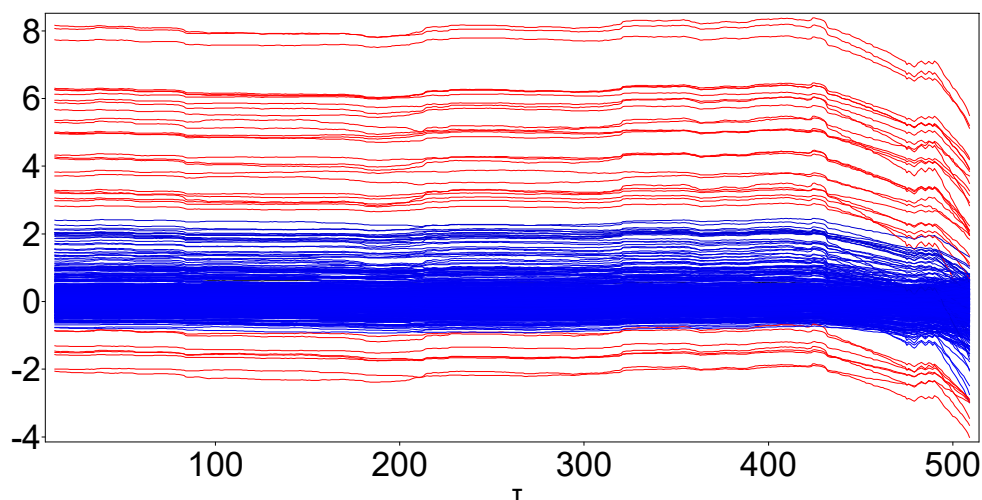
In this section we motivate transformation of the response in regression models by a reanalysis of the loyalty card data before transformation. This analysis indicates 82 outliers, a surprising increase on the number (18) found in §6. We then describe an approximate score test associated with the parametric family of power transformations introduced by Box and Cox (1964) which is suitable for monitoring data transformations during the forward search. This procedure is illustrated on the loyalty card data.

10.1. Analysis of Loyalty Card Data with the FS on the Original Scale

The top panel of Figure 8 repeats the monitoring of the minimum deletion residuals from the Forward Search that we saw in §6, but now for the untransformed data. The change is appreciable; instead of 18 outliers our automatic procedure had identified 82. Brushing this plot leads to the forward plot of scaled residuals in the bottom panel of the same figure, in which the outliers are shown in red (in the on-line .pdf version). Both plots of Figure 8 are stable until the outliers start to enter. Before then there are both positive and negative residuals, although more of the former. The initial effect of the outliers, from $m = 425$, is slightly to make all residuals less positive. This change becomes more marked at the end of the search as the outliers identified for the transformed data start to enter.



```
CALL FSR("lib.loyalty", {'x1' 'x2' 'x3'}, "y", "CLASSIFY mdrplot")
transform_original_data = 1 ;
```



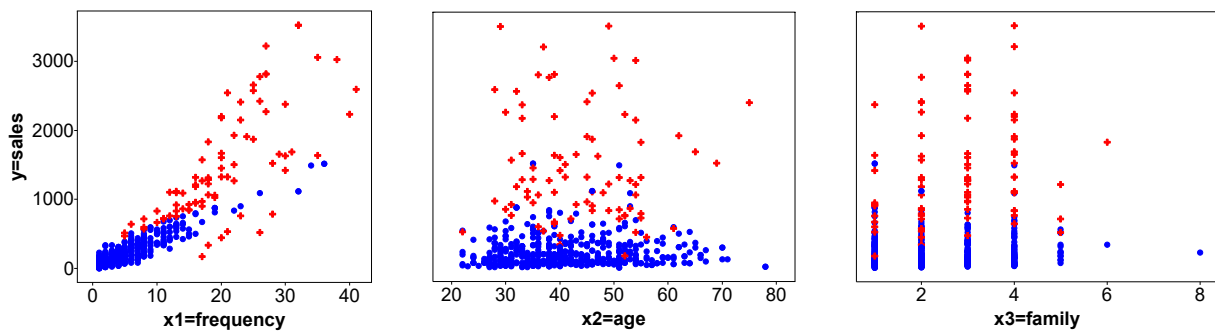
```
CALL FSR("lib.loyalty", {'x1' 'x2' 'x3'}, "y", "CLASSIFY RESFWDPLOT")
transform_original_data = 1 ;
```

Figure 8: Loyalty card data: monitoring plots on untransformed data. The top panel shows the absolute minimum deletion residuals among observations not in the subset, with the last part of the trajectory corresponding to the 82 detected outliers, automatically highlighted in red (in the on-line .pdf version). The bottom panel shows the scaled residuals with the trajectories corresponding to the 82 detected outliers, automatically represented in red (in the pdf version). The box under each panel contains the SAS code used to generate the plot.

Scatterplots of the data are in Figure 9. All, of course, show many more outliers than the panels of Figure 3. The top panel is the most informative. There is now a curved structure to the data, with variance increasing with x_1 . These are both aspects of data that require transformation for a homoscedastic linear model to be appropriate. The improvement from transformation is clear when comparing Figure 9 with Figure 3 for transformed data.

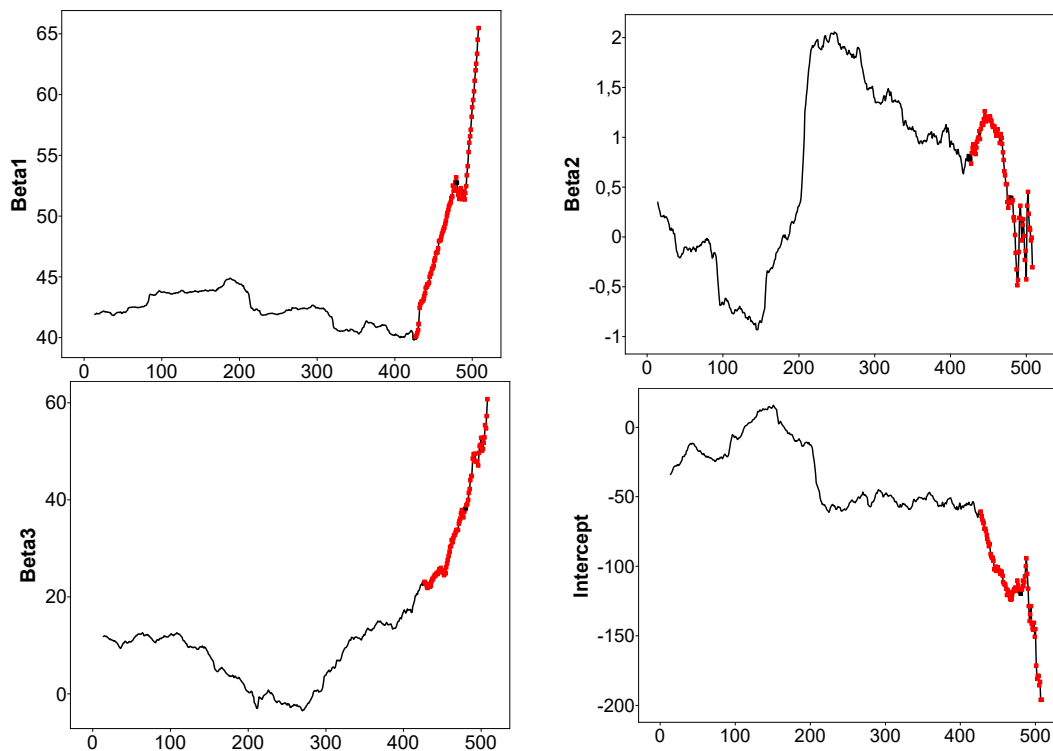
That the data need transformation is shown by the properties of the approximate score statistic (Atkinson, 1973). The forward plot of this statistic, described in §10.2, shows that the hypothesis of no transformation ($\lambda = 1$) is rejected for m around 250.

The forward plots of the parameter estimates in Figure 10 show that on the untransformed scale all estimates except $\hat{\beta}_2$ are susceptible to the presence of the observations now identified as outliers; they move outside the range of earlier variation by the end of the search. The uppermost panel of Figure 9 shows that in the absence of transformation the effect of inclusion of the outliers is to increase $\hat{\beta}_1$ as observations lying above the fitted robust line are introduced into the fit.



```
CALL FSR("lib.loyalty", {'x1' 'x2' 'x3'}, "y", "CLASSIFY scatter")
transform_original_data = 1 ;
```

Figure 9: Loyalty card data: scatterplots of untransformed data, with the 82 detected outliers automatically plotted as red crosses (in the on-line .pdf version). The box under the figure contains the SAS code used to generate the plots.



```
CALL FSR("lib.loyalty", {'x1' 'x2' 'x3'}, "y", "CLASSIFY BETA_PLOT")
transform_original_data = 1 ;
```

Figure 10: Loyalty card data: monitoring of estimated beta coefficients from untransformed data, with the last part of the trajectory, corresponding to the 82 detected outliers, automatically highlighted in red (in the on-line .pdf version). The box under the figure contains the SAS code used to generate the plots.

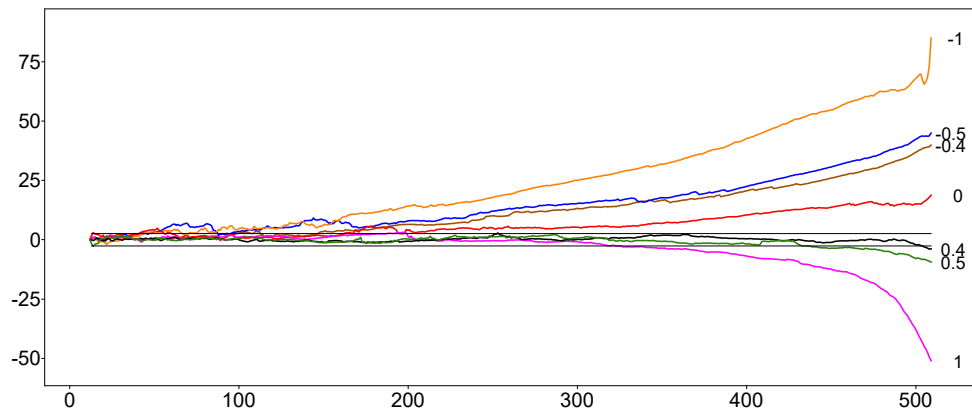
10.2. The Fanplot

For the linear regression model Eq. (1), Box and Cox (1964) analyse the normalized power transformation

$$z(\lambda) = \begin{cases} (y^\lambda - 1)/(\lambda \dot{y}^{\lambda-1}) & \lambda \neq 0 \\ \dot{y} \log y & \lambda = 0 \end{cases} \quad (11)$$

of the observations y , where the geometric mean of the observations is written as $\dot{y} = \exp(\sum \log y_i/n)$. The purpose of the transformation is to achieve a simple linear model with additive terms and homoskedastic normal errors.

If the observations are normally distributed with $R(\lambda)$ the residual sum of squares of the $z(\lambda)$, the profile



```
CALL FSRfan("lib.loyalty",{ 'x1' 'x2' 'x3'}, 'y', {-1 -0.5 -0.4 0 0.4 0.5 1});
```

Figure 11: Loyalty card data: fanplot for seven values of λ . The box under the figure contains the code used to generate the plot.

loglikelihood of the observations, maximized over β and σ^2 , is a function of $R(\lambda)$. The maximum likelihood estimate $\hat{\lambda}$ minimizes $R(\lambda)$. For inference about the transformation parameter λ , Box and Cox suggest the likelihood ratio test statistic. To test that $\lambda = \lambda_0$ this test is a function of $R(\hat{\lambda})$ and $R(\lambda_0)$. A disadvantage of the test is that a numerical maximization is required to find the value of $\hat{\lambda}$. For regression models a computationally simpler alternative test is the approximate score statistic derived by Taylor series expansion of (11) as

$$z(\lambda) \doteq z(\lambda_0) + (\lambda - \lambda_0)w(\lambda_0), \quad \text{where} \quad w(\lambda_0) = \left. \frac{\partial z(\lambda)}{\partial \lambda} \right|_{\lambda=\lambda_0}. \quad (12)$$

The combination of (12) and the regression model $y = x^T \beta + \epsilon$ yields the model

$$z(\lambda_0) = x^T \beta - (\lambda - \lambda_0)w(\lambda_0) + \epsilon = x^T \beta + \gamma w(\lambda_0) + \epsilon. \quad (13)$$

Because (13) is again a regression model with an extra variable $w(\lambda_0)$ derived from the transformation, the new variable is called the constructed variable for the transformation. The approximate score statistic $T_p(\lambda_0)$ for testing the transformation $\lambda = \lambda_0$ is the t statistic for regression on $w(\lambda_0)$ in (13). This can be calculated either directly from the regression in (13), or from the formulae in Atkinson and Riani (2000, Chapter 4) in which multiple regression on x is adjusted for the inclusion of the constructed variable.

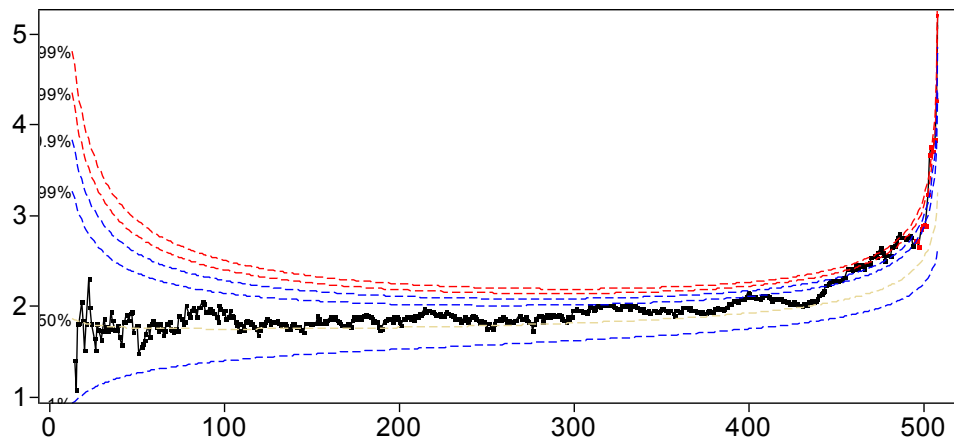
As our analysis of the loyalty card data illustrates, outliers in one transformed scale may not be outliers in another scale. If the data are analysed using the wrong transformation, too many outliers may be indicated. We therefore need to analyse the data using several values of λ . We have given the name "fan plot" to the simultaneous forward plot of the score statistic $T_p(\lambda)$ for several values of λ . For further details of the procedure see Atkinson and Riani (2000, §4.3), with distributional results in Atkinson and Riani (2002).

10.3. Loyalty Card Data

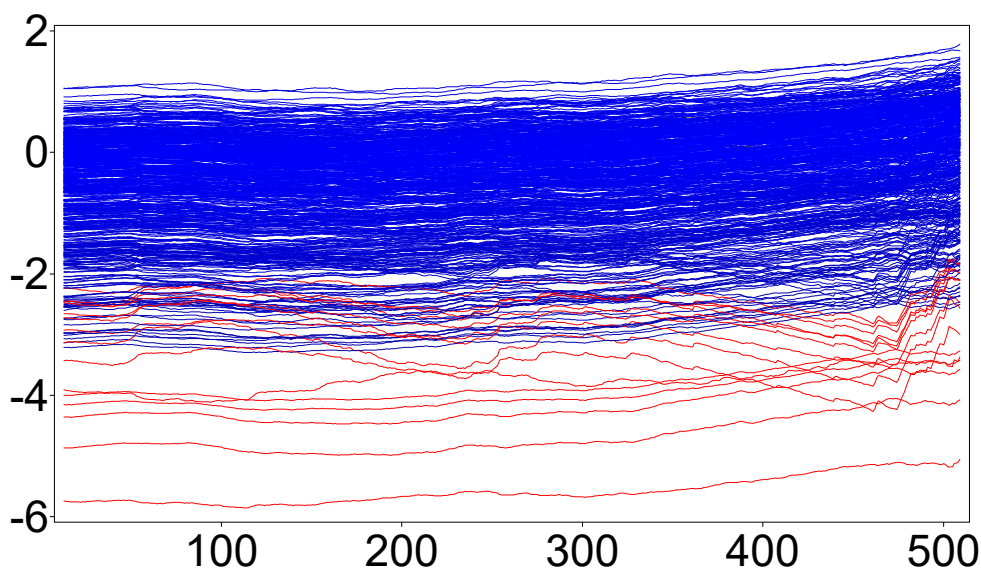
Figure 11 is the fanplot of transformation score tests for the loyalty card data calculated for seven values of λ . A separate search is used for each value of λ . The horizontal lines at the centre of the plot are at ± 2.58 , corresponding to 99.9% of the asymptotic null standard normal distribution of the statistic. It is clear from the plot that none of these values of λ is acceptable over the whole search. Box and Cox suggest that values of λ that are ratios of small integers are to be preferred as they often have a physical interpretation. Their largest set of data had 48 observations. However, for larger numbers of observations, as here, a finer grid of values is often necessary. The value we have used of 0.4 produces a trajectory that lies within the bounds until the 18 outliers start to enter the subset.

We have already considered analyses for $\lambda = 0.4$ and 1. We now show brief output from the analysis when $\lambda = 0$, that is the natural logarithmic transformation. Figures 12, 13 and 14 respectively show the forward plots of minimum deletion residuals and of residuals, the scatterplots with outliers and the forward plots of the regression coefficients. Although few outliers are identified when $\lambda = 0$, the logarithmic transformation produces a trajectory in the fan plot (Figure 11) that starts to lie outside the bounds for m -values slightly below 200.

There is a clear progression in the scatterplots as λ goes from 0 (Figure 13) to 0.4 (Figure 3) to 1 (Figure 9). In these figures we have used direct transformations of y . However, in the plots of the estimated



```
CALL FSR("lib.loyalty", {'x1' 'x2' 'x3'}, "y", "CLASSIFY mdrplot ")
transform_original_data = 0 ;
```



```
CALL FSR("lib.loyalty", {'x1' 'x2' 'x3'}, "y", "CLASSIFY RESFWDPLOT ")
transform_original_data = 0 ;
```

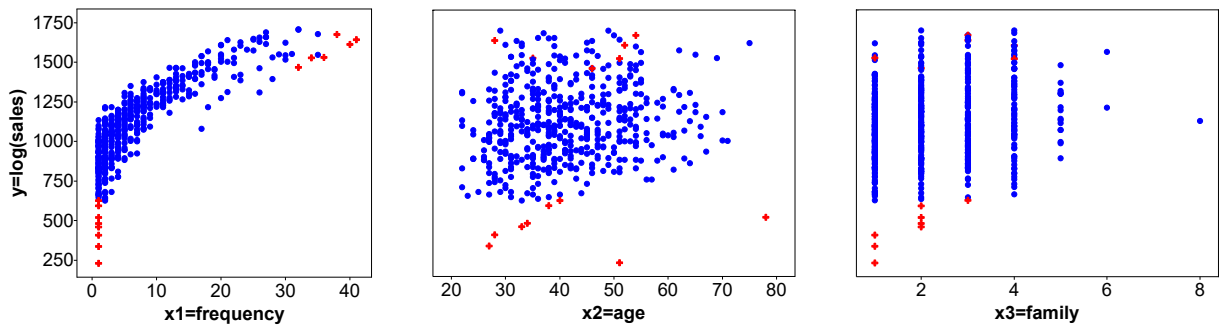
Figure 12: Loyalty card data: monitoring plots for log-transformed data ($\lambda = 0$). The top panel shows the absolute values of minimum deletion residuals among observations not in the subset, with the last part of the trajectory corresponding to the 14 detected outliers, automatically highlighted in red (in the on-line .pdf version). The bottom panel shows the scaled residuals with the trajectories corresponding to the 14 detected outliers automatically represented in red (in the on-line .pdf version). The box under each panel contains the SAS code used to generate the plot.

beta coefficients (Figures 14, Figures 4 and 10) we have calculated the coefficients using $z(\lambda)$ as the response so that the numerical values of the coefficients are comparable across λ values.

11. Monitoring Other Forms of Robust Regression

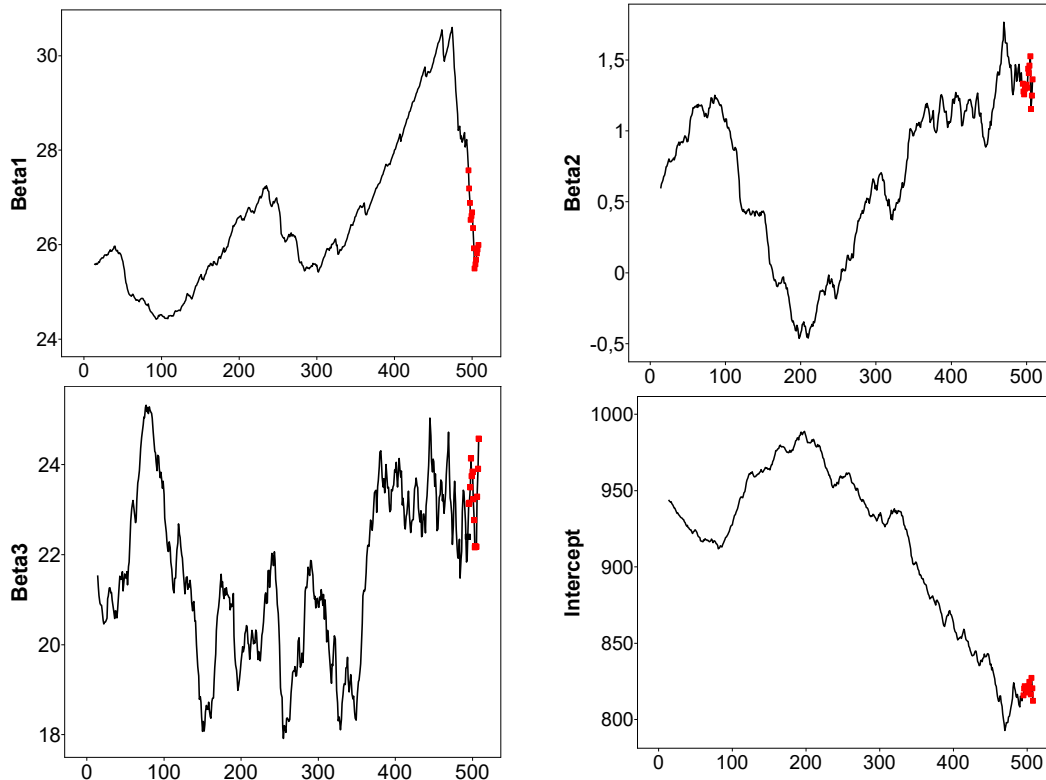
11.1. Soft Trimming

In addition to the extension of the SAS version of the FS to include the batch procedure of §8 we have provided a SAS version of the monitoring of S-estimation, which is also available in FSDA. Further we have introduced to SAS two new monitoring possibilities, those for LTS and LMS regression, which are not present in the FSDA toolbox. In this section we provide brief comments on robust regression with both soft and hard trimming. Our SAS examples are in §12. A much fuller discussion of monitoring robust regression is Riani *et al.* (2014a), including examples of S, MM, LMS and LTS analyses using the FSDA. We monitor



```
CALL FSR("lib.loyalty", {'x1' 'x2' 'x3'}, "y", "CLASSIFY scatter")
transform_original_data = 0 ;
```

Figure 13: Loyalty card data: scatterplots of log-transformed data ($\lambda = 0$), with the 14 detected outliers automatically plotted as red crosses (in the on-line .pdf version). At the figure foot, the code used to generate the plots.



```
CALL FSR("lib.loyalty", {'x1' 'x2' 'x3'}, "y", "CLASSIFY BETA_PLOT")
transform_original_data = 0 ;
```

Figure 14: Loyalty card data: monitoring of estimated beta coefficients for log-transformed data ($\lambda = 0$), with the last part of the trajectory, corresponding to the 14 detected outliers, automatically highlighted in red (in the on-line .pdf version). At the figure foot, the code used to generate the plots.

using one of two properties of the estimators

1. *Breakdown point, bdp*; the asymptotic proportion of observations that can go to ∞ without affecting the parameter estimates. We stress that this definition requires both that $n \rightarrow \infty$ and that the contaminating observations also become increasingly remote. Riani *et al.* (2014c) illustrate the degradation in performance that occurs as the contaminating observations become decreasingly remote.

As a result of monitoring we observe an *empirical breakdown point*, the point at which the fit switches from being robust to non-robust least squares. This important property depends both on the nominal

properties of the estimator and on the particular data set being analysed.

2. The *efficiency of estimation*. For normally distributed responses with explanatory variables x let the robust estimator of the parameter β_j of the linear model be $\tilde{\beta}_j$, with $\hat{\beta}_j$ the least squares estimator. The efficiency of estimation of β_j is then $\text{Eff} = \text{var}\hat{\beta}_j/\text{var}\tilde{\beta}_j$.

For S estimation we monitor performance as the theoretical bdp varies from 0.5 to 1. For hard trimming the theoretical $\text{bdp} = 1 - h/n$, with the asymptotic value of $\text{Eff} = h/n$. Thus it is not possible to have an estimator which simultaneously has high bdp and high efficiency. We now outline results showing that the same restriction applies to soft trimming estimators.

11.2. M and S Estimation

In least squares estimation, the value of $\hat{\beta}$ does not depend on the estimate of σ^2 . The same is not true in M estimation and derived procedures.

We start with M-estimation. Suppose the error variance σ^2 is known in the regression model Eq. (1) and let the residuals for some estimate b of β be $r_i = y_i - b^\top x_i$. Then the regression M-estimate of β is the value that minimizes the objective function

$$\sum_{i=1}^n \rho\{r_i(\beta)/\sigma\}, \quad (14)$$

where ρ is a function that reduces the importance of observations with large residuals. The most well known ρ functions, described for example by Riani *et al.* (2014a), are:

Tukey's Bisquare (or Biweight) function; Hampel's ρ function; the Optimal ρ function and the Hyperbolic ρ function. In our analyses we have used the Tukey's Bisquare that is the default of PROC ROBUSTREG, the SAS procedure that we used for monitoring these estimators.

For robust M estimation, σ should also be estimated robustly. The M-estimator of scale $\tilde{\sigma}_M$ is found by solution of the equation

$$\frac{1}{n} \sum_{i=1}^n \rho\left(\frac{r_i}{\sigma}\right) = \frac{1}{n} \sum_{i=1}^n \rho\left(\frac{y_i - \beta^\top x_i}{\sigma}\right) = K, \quad (15)$$

in theory solved among all $(\beta, \sigma) \in \mathbb{R}^p \times (0, \infty)$, where $0 < K < \sup \rho$. If we take the minimum value of $\tilde{\sigma}_M$ which satisfies equation Eq. (15), we obtain the S-estimate of scale ($\tilde{\sigma}_S$) and the associated estimate of the vector of regression coefficients (Rousseeuw and Yohai, 1984). The estimator of β is called an S-estimator because it is derived from a scale statistic, although in an implicit way.

For these M- and S-estimators asymptotic results on the bdp require that the distribution of x is such that $X^\top X/n$ converges to a finite limit. Let this be L . Then the asymptotic distribution of the estimators is multivariate normal with variance-covariance matrix proportional to L^{-1} , as it is for least squares. Thus the bdp of the estimators does not depend on the distribution of x .

Rousseeuw and Leroy (1987, p. 139) give conditions to be obeyed by the symmetric function ρ . One is that there should be a $c > 0$ such that ρ is strictly increasing on $[0, c]$ and constant on $[c, \infty)$. The breakdown point of the S-estimator tends to bdp when $n \rightarrow \infty$. As c increases, fewer observations are downweighted, so that the estimate of σ^2 approaches that for least squares and $\text{bdp} \rightarrow 0$. Riani *et al.* (2014b) show that choice of the value of c determines both the bdp and efficiency of the estimator, although the exact values depend upon the specific ρ function. The dependence of both *bdp* and efficiency on the value of c again means that, as it is for hard trimming, it is impossible to have an estimator with high values of both properties.

We monitor S estimators by looking over a grid of values of bdp. Riani *et al.* (2014b, §3.1) give computationally efficient calculations for finding the value of c for Tukey's bisquare once the value of bdp is specified. The calculations depend on the polynomial nature of the ρ function and require moments of truncated chi-squared random variables. For MM estimators we instead monitor efficiency. The calculations to find c again rely on expectations of truncated chi-squared variable and are given in their §3.2. The extension to the optimal loss function is given in their §7 - the calculations are similar to those for Tukey's bisquare since the ρ function is again of a polynomial form. We use numerical integration for the hyperbolic ρ function.

An important final point is that the ρ functions for the mean in Eq. (14) and Eq. (15) may be different. However, PROC ROBUSTREG, for all estimators where such a choice exists, uses the same ρ for both the mean and the scale estimators.

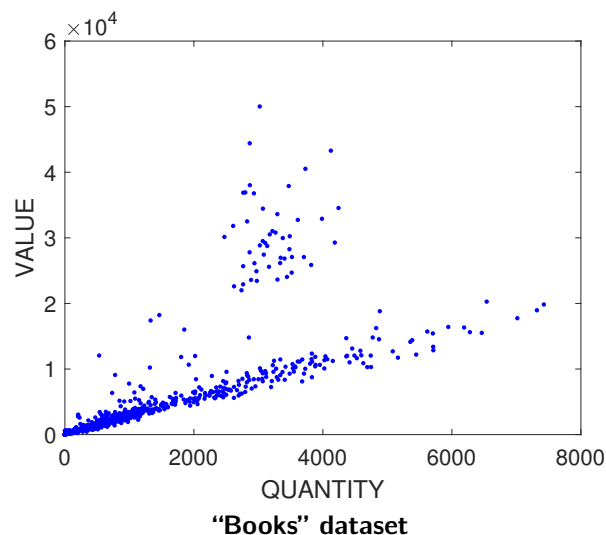


Figure 15: A rather complex trade dataset. Following Perrotta and Torti (2018), we analyze the subset of retained units: "Printed books, brochures, leaflets and similar printed matter" (723 units).

12. Data Analyses with S, LTS and LMS Routines

This section details a new SAS function that monitors a number of traditional robust multivariate and regression estimators for various choices of breakdown or efficiency. Here we provide examples of the analysis of regression data with LMS and LTS as well as using S-estimation. The full list of possibilities is presented at the end of §7.

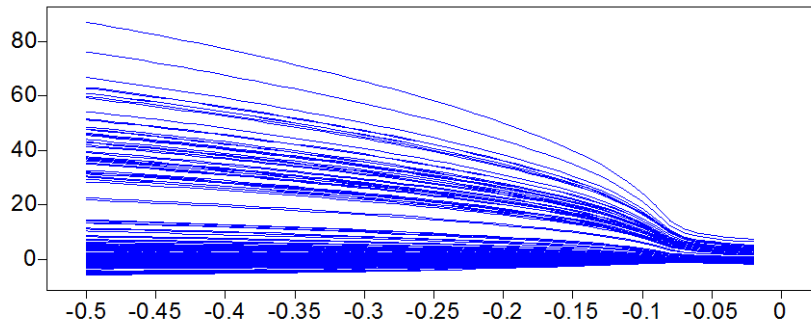
The pattern of scaled residuals in the forward plots of Figures 2, 8 and 12 (bottom panels) all show a stable horizontal pattern for the greater part of the search until outliers start to enter the fit. In introducing the idea of monitoring regression, Riani *et al.* (2014a) provided plots when monitoring S estimation as a function of bdp which had a related structure; changes in the estimate occurred at one or two values of bdp, between which the pattern of residuals remained similar, but with decreasing magnitude of the scaled residuals as bdp decreased. The structure for LTS showed the residuals decreasing more rapidly as a function of bdp with LMS similar but with appreciably more noise in the curves. Similar structures are obtained by Perrotta and Torti (2018) in the analysis of simulated data. Changes in the fit from robust to non-robust allow the determination of the empirical breakdown point and hence the provision of efficient estimates.

To illustrate the use of the SAS routines for these regressions, we use a rather more complicated data set introduced by Perrotta and Torti (2018) in the discussion of Cerioli *et al.* (2017). The dataset is an example of trade data from EU customs returns. The response is value and the single explanatory variable is quantity. If markets are working correctly, there should be a linear relationship between the two. The focus in the analysis is on the detection of observations with a different relationship between the two variables, which may be an indicator of fraudulent customs declarations. A seafood example is used by Atkinson *et al.* (2018b) in which there are two distinct linear relationships between price and quantity. An extra problem in monitoring these data sets is the presence of a large number of small transactions, which obscure any structure of the larger observations, which is where financially important fraud, if any, will occur. Perrotta and Torti (2018) analyse thinned data sets in which the number of observations is partially reduced, whilst retaining the overall structure of the data. The data set is "Books", defined as printed books, brochures, leaflets and similar printed matter. The scatterplot of the data set is in Figure 15; it contains $n = 723$ units, after thinning following the procedure of Cerioli and Perrotta (2014).

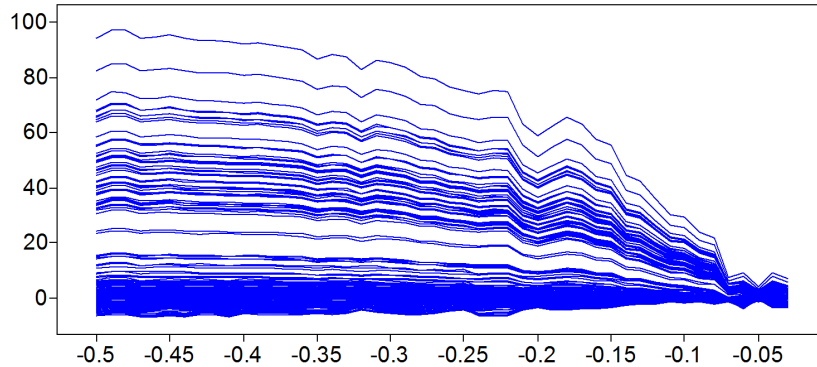
The panels of Figure 16 are monitoring plots of scaled residuals for the books data from S, LTS and LMS estimators from our IML Studio implementation. The S curve shape, which is identical to the one produced with FSDA by Perrotta and Torti (2018), as well as the new LTS and LMS curves show the presence of structure in the data. A sharp decrease in the residuals occurs below a breakdown value of approximately 8%, which corresponds to the percentage of outliers visible as two dispersed clusters in Figure 15.

The transition shown for S-estimation (top panel of Figure 16) is going from a robust analysis to non-robust least squares is smooth. The transitions for LTS (middle panel of Figure 16) happens at the same value of bdp, but is appreciably sharper. That for LMS (bottom panel of Figure 16) is less sharp and, for indication of a change point, closer to that for S-estimation.

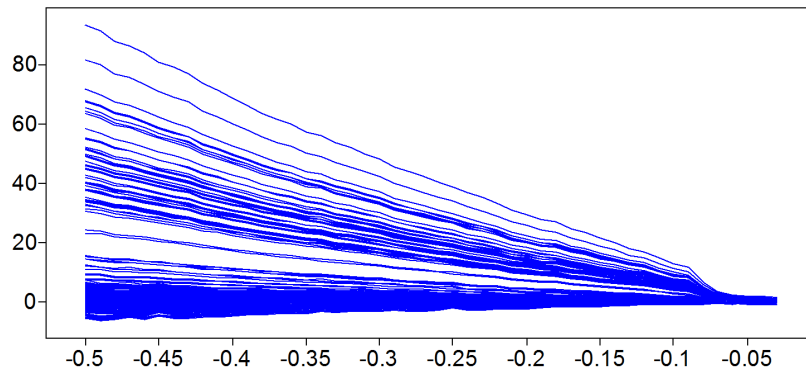
The structure of these plots can be interpreted by inspection of Figure 15. For fits with a high bdp the



```
call monitoring_S_MM_LTS_LMS("lib.books", "q", "v", "S");
```



```
call monitoring_S_MM_LTS_LMS("lib.books", "q", "v", "lts");
```



```
call monitoring_S_MM_LTS_LMS("lib.books", "q", "v", "lms");
```

Figure 16: Books data; S, LTS and LMS estimators: monitoring the scaled residuals as the breakdown point varies. At the foot of each panel, the code used to generate the plot.

clear lower linear relationship is fitted, initially to observations near the origin with a low variance. As bdp decreases fewer observations are trimmed or downweighted and those with higher variance further from the origin influence the fit, increasing the estimate of σ^2 and decreasing the values of the residuals. That most of the residuals are positive until a bdp around 8% is caused by the (basically) two clouds of observations that lie above the line that is being robustly fitted.

13. Discussion and Extensions

The main purpose of the SAS programming described in this paper is to provide a library of routines for very robust regression and to embed these in a monitoring framework. As a result adaptive values of trimming parameters or bdp can be found which, for a particular dataset, yield the most efficient robust parameter estimates. We also provide a series of innovative plots to aid interpretation of the data and also of the fitted models and their robustness. In doing so, we have introduced two new monitoring possibilities, for LTS and

LMS, which are not yet present in the FSDA toolbox. As we mention at the end of §7 we have also improved the SAS implementations of LMS and LTS.

The batch procedure described in §8 provides a computationally fast version of the FS taking advantage of the ability of SAS to handle data sets much larger than those analysable by R or in the MATLAB FSDA tool box, with little loss in statistical efficiency. Further developments which we have not space to describe here include:

1. `FSM.sx`, the multivariate counterparts of FSR, and `FSMfan.sx` for multivariate transformations.
2. `FSRms.sx` for choosing the best model in regression. This function implements the procedure of Riani and Atkinson (2010) which combines Mallows' C_p (Mallows, 1973) with the flexible trimming of the FS to yield an information rich plot "The Generalized Candlestick Plot" revealing the effect of outliers on model choice.
3. `FSRMultipleStart.sx` and `FSMmultiplestart.sx` for identifying observations that are divided into groups either of regression models or of multivariate normal clusters. The later procedure is derived from the FSDA implementation of Atkinson *et al.* (2018a).

Thus, despite the length of our paper, we have here presented only about one third of our SAS programs.

The package will be made available in github as a complement to the wider FSDA toolbox for MATLAB project (<https://github.com/UniprJRC/FSDA>) mentioned in Section 7.

References

- Atkinson, A. C. (1973). Testing transformations to normality. *Journal of the Royal Statistical Society, Series B*, **35**, 473–479.
- Atkinson, A. C. and Riani, M. (2000). *Robust Diagnostic Regression Analysis*. Springer–Verlag, New York.
- Atkinson, A. C. and Riani, M. (2002). Tests in the fan plot for robust, diagnostic transformations in regression. *Chemometrics and Intelligent Laboratory Systems*, **60**, 87–100.
- Atkinson, A. C. and Riani, M. (2006). Distribution theory and simulations for tests of outliers in regression. *Journal of Computational and Graphical Statistics*, **15**, 460–476.
- Atkinson, A. C., Riani, M., and Cerioli, A. (2010). The forward search: theory and data analysis (with discussion). *Journal of the Korean Statistical Society*, **39**, 117–134. doi:10.1016/j.jkss.2010.02.007.
- Atkinson, A. C., Riani, M., and Cerioli, A. (2018a). Cluster detection and clustering with random start forward searches. *Journal of Applied Statistics*, **45**, 777–798. doi <http://dx.doi.org/10.1080/02664763.2017.1310806>.
- Atkinson, A. C., Corbellini, A., and Riani, M. (2018b). Robust Bayesian regression with the forward search: Theory and data analysis. *International Statistical Review*, **86**, 205–218. doi:10.1111/insr.12247.
- Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B*, **26**, 211–246.
- Buja, A. and Rolke, W. (2003). Calibration for simultaneity: (re)sampling methods for simultaneous inference with applications to function estimation and functional data. Technical report, The Wharton School, University of Pennsylvania.
- Cerioli, A. and Perrotta, D. (2014). Robust clustering around regression lines with high density regions. *Advances in Data Analysis and Classification*, **8**, 5–26.
- Cerioli, A., Farcomeni, A., and Riani, M. (2014). Strong consistency and robustness of the Forward Search estimator of multivariate location and scatter. *Journal of Multivariate Analysis*, **126**, 167–183.
- Cerioli, A., Riani, M., Atkinson, A. C., and Corbellini, A. (2017). The power of monitoring: How to make the most of a contaminated multivariate sample (with discussion). *Statistical Methods and Applications*, pages <https://doi.org/10.1007/s10260-017-0409-8>.
- García-Escudero, L. A., Gordaliza, A., Matran, C., Mayo-Iscar, A., and San Martín, R. (2008). A general trimming approach to robust cluster analysis. *Annals of Statistics*, **36**, 1324–1345.
- García-Escudero, L. A., Gordaliza, A., Mayo-Iscar, A., and San Martín, R. (2010). Robust cluster-wise linear regression through trimming. *Computational Statistics and Data Analysis*, **54**, 3057–3069. doi:10.1016/j.csda.2009.07.002.
- Guenther, W. C. (1977). An easy method for obtaining percentage points of order statistics. *Technometrics*, **19**, 319–321.
- Hampel, F. R. (1975). Beyond location parameters: robust concepts and methods. *Bulletin of the International Statistical Institute*, **46**, 375–382.
- Hubert, M. and Debruyne, M. (2010). Minimum covariance determinant. *WIREs Computational Statistics*, **2**, 36–43.
- Hubert, M., Rousseeuw, P. J., and Vanden Branden, K. (2005). ROBPCA: A new approach to robust principal component analysis. *Technometrics*, **47**, 64–79.
- Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994). *Continuous Univariate Distributions - 1, 2nd edition*. Wiley, New York.
- Lehmann, E. (1991). *Point Estimation*. Wiley, New York.
- Maitra, R. and Melnykov, V. (2010). Simulating data to study performance of finite mixture modeling and clustering algorithms. *Journal of Computational and Graphical Statistics*, **19**, 354–376.

- Mallows, C. L. (1973). Some comments on C_p . *Technometrics*, **15**, 661–675.
- Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. Wiley, Chichester.
- Perrotta, D. and Torti, F. (2018). Discussion of “the power of monitoring: How to make the most of a contaminated multivariate sample”. *Statistical Methods and Applications*. <https://doi.org/10.1007/s10260-017-0420-0>.
- Perrotta, D., Riani, M., and Torti, F. (2009). New robust dynamic plots for regression mixture detection. *Advances in Data Analysis and Classification*, **3**, 263–279. doi:10.1007/s11634-009-0050-y.
- Pison, G., Van Aelst, S., and Willems, G. (2002). Small sample corrections for LTS and MCD. *Metrika*, **55**, 111–123. doi: 10.1007/s001840200191.
- Riani, M. and Atkinson, A. C. (2010). Robust model selection with flexible trimming. *Computational Statistics and Data Analysis*, **54**, 3300–3312. doi: 10.1016/j.csda.2010.03.007.
- Riani, M., Atkinson, A. C., and Cerioli, A. (2009). Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society, Series B*, **71**, 447–466.
- Riani, M., Perrotta, D., and Torti, F. (2012). FSDA: a MATLAB toolbox for robust analysis and interactive data exploration. *Chemometrics and Intelligent Laboratory Systems*, **116**, 17–32. doi:10.1016/j.chemolab.2012.03.017.
- Riani, M., Cerioli, A., Atkinson, A. C., and Perrotta, D. (2014a). Monitoring robust regression. *Electronic Journal of Statistics*, **8**, 642–673.
- Riani, M., Cerioli, A., and Torti, F. (2014b). On consistency factors and efficiency of robust S-estimators. *TEST*, **23**, 356–387.
- Riani, M., Atkinson, A. C., and Perrotta, D. (2014c). A parametric framework for the comparison of methods of very robust regression. *Statistical Science*, **29**, 128–143.
- Riani, M., Cerioli, A., Corbellini, A., Perrotta, D., Torti, F., Sordini, E., and Todorov, V. (2017). fsdar: Robust data analysis through monitoring and dynamic visualization. <https://CRAN.R-project.org/package=fsdaR>.
- Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, **79**, 871–880.
- Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. Wiley, New York.
- Rousseeuw, P. J. and Yohai, V. J. (1984). Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis: Lecture Notes in Statistics 26*, pages 256–272. Springer Verlag, New York.
- Rousseeuw, P. J., Croux, C., Todorov, V., Ruckstuhl, A., Salibián-Barrera, M., Verbeke, T., and Maechler, M. (2009). Robustbase: Basic robust statistics. r package version 0.92-7. URL <http://CRAN.R-project.org/package=robustbase>.
- Tallis, G. M. (1963). Elliptical and radial truncation in normal samples. *Annals of Mathematical Statistics*, **34**, 940–944.
- TFEU (2012). Tfeu. treaty on the functioning of the european union (consolidated version 2012), articles 310 and 325. Technical report, Official Journal of the European Union, C 326, 26.10.2012.
- Todorov, V. and Filzmoser, P. (2009). An object-oriented framework for robust multivariate analysis. *Journal of Statistical Software*, **32**, 1–47.
- Torti, F., Perrotta, D., Riani, M., and Cerioli, A. (2018). Assessing trimming methodologies for clustering linear regression data. *Advances in Data Analysis and Classification*.
- Van Aelst, S. and Rousseeuw, P. J. (2009). Minimum volume ellipsoid. *WIREs Computational Statistics*, **1**, 71–82.

- Vanden Branden, K. and Hubert, M. (2005). Robustness properties of a robust partial least squares regression method. *Analytica Chimica Acta*, **515**, 229–241.
- Verboven, S. and Hubert, M. (2010). Matlab library LIBRA. *WIREs Computational Statistics*, **2**, 509–515.
- WCO (2017). Message from the wco secretary general. Technical report, World Customs Organization.

A. Annex: code to replicate the results and the figures in the report

```
/******set path where input dataset are stored *****/
path = "D:\benchmark_FSDA\software_and_ds_to_upload\ds\ds_for_exploratory_analysis";

/*load loyalty data on original scale*/
libname lib (path);
use ("lib.loyalty");
    read all var {'x1'      'x2'      'x3'} into x[colname=colnx];
    read all var 'y' into y[colname=colny];
    x = x || j(nrow(x),1,1);      /* add intercept */
close ("lib.loyalty");

/*next command + F11 to obtain Figures 2, 3, 4, that is mdr fwd plot, res fwd plot, scatteplots,
beta fwd plots for 0.4 transformation*/
CALL FSR("lib.loyalty", {'x1' 'x2' 'x3'}, "y", "DEBUG CLASSIFY RESFWDPLOT mdrplot scatter BETA_PLOT") transform_original_data = 0.4 ;

/*next command + F11 to obtain Figures 8, 9, 10, that is mdr fwd plot, res fwd plot, scatteplots,
beta fwd plots for no transformation*/
CALL FSR("lib.loyalty", {'x1' 'x2' 'x3'}, "y", "DEBUG CLASSIFY RESFWDPLOT mdrplot scatter BETA_PLOT") transform_original_data = 1 ;

/*next command to obtain Figure 11, that is the fanplot*/
CALL FSRfan("lib.loyalty", {'x1' 'x2' 'x3'}, 'y', {-1 -0.5 -0.4 0 0.4 0.5 1});

/*next command + F11 to obtain Figures 12, 13, 14, that is mdr fwd plot, res fwd plot, scatteplots,
beta fwd plots for logarithmic transformation*/
CALL FSR("lib.loyalty", {'x1' 'x2' 'x3'}, "y", "DEBUG CLASSIFY RESFWDPLOT mdrplot scatter BETA_PLOT") transform_original_data = 0 ;

/*next commands to obtain Figure 16 on books dataset (CN 49019900), the monitoring of studentized residual fro S, LTS and LMS estimators*/
use("lib.p_49019900_thinned");
read all var {'x'} into x[colname=colnx];
read all var {'y'} into y[colname=colny];
close("lib.p_49019900_thinned");
/*S*/
CALL monitoring_S_MM_LTS_LMS("lib.p_49019900_thinned","x","y","s") ;
/*MM not in the paper*/
CALL monitoring_S_MM_LTS_LMS("lib.p_49019900_thinned","x","y","mm") ;
/*LTS*/
CALL monitoring_S_MM_LTS_LMS("lib.p_49019900_thinned","x","y","lts") ;
/*LMS*/
CALL monitoring_S_MM_LTS_LMS("lib.p_49019900_thinned","x","y","lms") ;

/*next commands to obtain Figure on jewellery dataset (CN 7117190090), the monitoring of studentized residual fro S, LTS and LMS estimators*/
*use("lib.POD41_7117190090");
*read all var {'x'} into x[colname=colnx];
*read all var {'y'} into y[colname=colny];
*close("lib.POD41_7117190090");
/*S*/
*CALL monitoring_S_MM_LTS_LMS("lib.POD41_7117190090","x","y","s",path) ;
/*MM not in the paper*/
*CALL monitoring_S_MM_LTS_LMS("lib.POD41_7117190090","x","y","mm",path) ;
/*LTS*/
*CALL monitoring_S_MM_LTS_LMS("lib.POD41_7117190090","x","y","lts",path) ;
/*LMS*/
*CALL monitoring_S_MM_LTS_LMS("lib.POD41_7117190090","x","y","lms",path) ;
```

B. Annex: use of the monitoring tools in WebAriadne

Our SAS library for robust regression is motivated by problems linked to the Customs Union and Anti-Fraud policies of the European Union (EU), which are rooted in its founding Treaties TFEU (2012). In fact, the efficient implementation of such policies, as also stated by the World Customs Organization WCO (2017), calls for a modernization of the national anti-fraud services through the adoption of tools based on state-of-the-art mathematical, statistical and computer science methods. In this framework, the Joint Research Centre (JRC) of the European Commission delivers such tools to the law-enforcement partners in the EU Institutions and Member States since decades (the roots of the activity can be dated back to 1995). The role assigned to the JRC in this policy domain comprises the modeling of fraud in pertinent statistical data, the development of the related statistical methods for fraud detection, their product software implementation, their deployment as services accessible to customers, and the routine dissemination of alerts (fraud relevant signals) to authorized users through the web. More precisely, the users access alerts related to trade-based illicit activities through the THESEUS resource (<https://theseus.jrc.ec.europa.eu>) or generate them in full autonomy, on data of their choice, using tools accessible through the web application WebARIADNE (<https://webariadne.jrc.ec.europa.eu>). Figure 17 shows their login page. The SAS library discussed in the paper is a key module of both THESEUS and WebARIADNE, illustrated below with few snapshots.

The left panel of Figure 18 shows how the user selects a dataset of interest and the statistical application to apply. On the right panel the user has selected a local dataset and is presented with a preview of its content. The user might also want to analyze a previously uploaded dataset: Figure 19 shows the preview given when the dataset is selected, with the list of the fields and a sample of records. The frame of Figure 20 is specific to the SAS-based regression module of this paper. Here, the user can set the key input parameters, namely the estimation method for fitting the data (FS, LTS, LMS, S, M, MM), the significance level for detecting the outliers, the dependent and independent regression variables, possible grouping variables for partitioning the dataset in homogeneous groups. Figure 21 shows the output of the most recent runs: there are three runs with the Forward Search and one with Least Trimmed Squares. Such output is typically presented to the user as a table with the relevant input variables and generated statistics (estimated parameters, pvalues, residuals, etc.). Same results are also disseminated to the authorized users of THESEUS in a comprehensive form: an example is shown in Figure 22.

The image shows two side-by-side screenshots of web application login pages. The left screenshot is for 'WebARIADNE - statistics for anti-fraud' by the Joint Research Centre. It features a login form with 'Username:' and 'Password:' fields and a 'LOGIN' button. Below the form is a 'Welcome to WebARIADNE!' message and a 'WebARIADNE at a glance' section with descriptive text. The right screenshot is for 'THESEUS - Statistics and Information Technologies for Anti-Fraud (SITAF)'. It features a 'Welcome to the THESEUS website' message, a navigation menu with 'Public area' and 'Restricted area' tabs, and a 'HOME' button. It also includes a 'Welcome' sidebar with links for 'Aim and Overview', 'Acknowledgements', 'Contacts', and 'Back to top'.

Figure 17: The login page of the WebARIADNE and THESEUS applications.

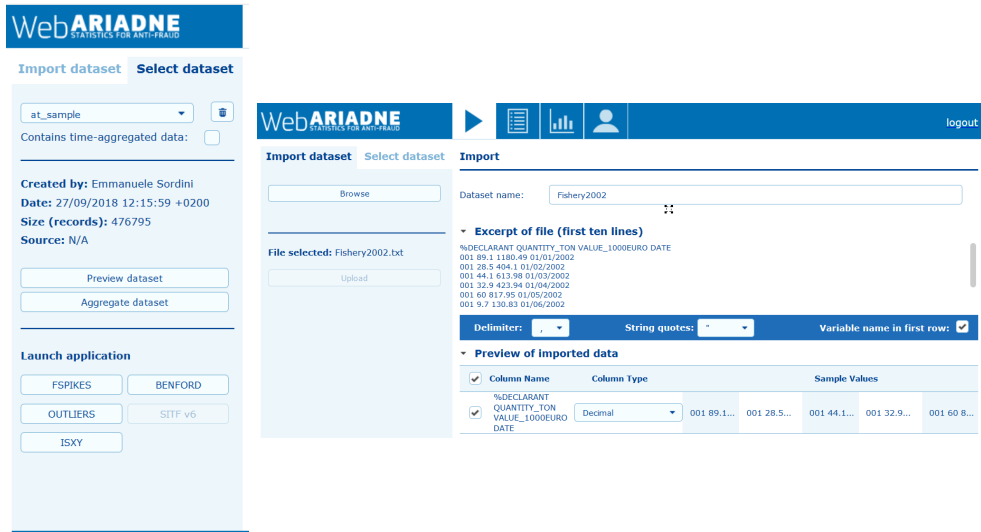


Figure 18: WebARIADNE. Left panel: selection of dataset and statistical application of interest. Right panel: wizard for importing a new dataset.

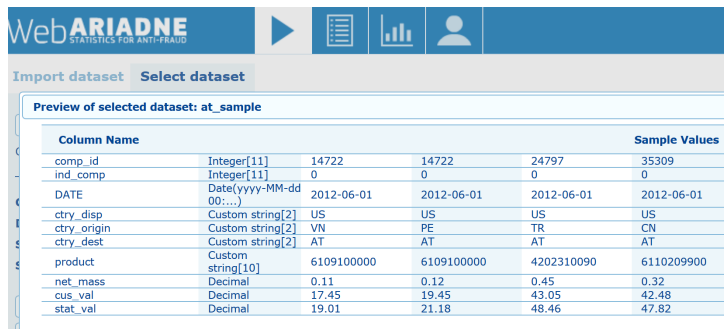


Figure 19: WebARIADNE. Selection of an existing dataset: example of data preview.

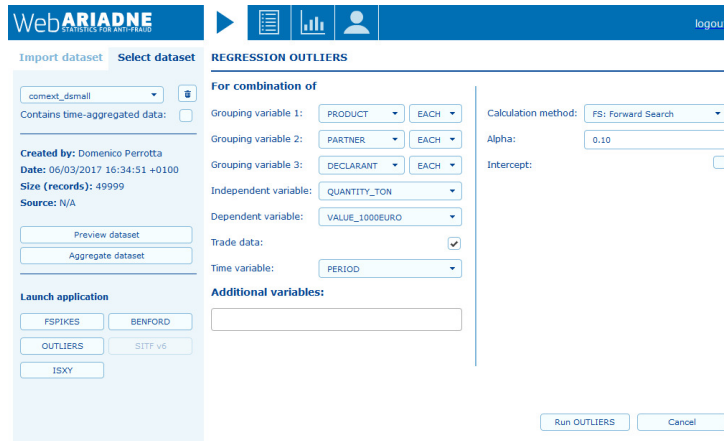



Figure 20: WebARIADNE. Selection of input parameters for the execution of a given method.

The screenshot shows the 'VIEW RESULTS' section. It displays a table with the following columns: 'Job ID', 'Date Created', 'Statistical Procedure', 'Methods', 'Dataset', and 'Published'. The table contains the following data:

Job ID	Date Created	Statistical Procedure	Methods	Dataset	Published
7318111911	26/03/2019 16:48:45 +0100	ROBUST	LTS	comext_dsmall	No
7316250995	26/03/2019 16:27:55 +0100	ROBUST	FS	cap_64_CN_small_small_smNo	No
7315840481	26/03/2019 16:13:55 +0100	ROBUST	FS	comext_dsmall	No
7311770200	26/03/2019 15:06:03 +0100	ROBUST	FS	comext_dsmall	No

Figure 21: WebARIADNE. Selection of an existing set of results obtained (in the example) with the Forward Search and the LTS.


JOINT RESEARCH CENTRE
 THESEUS - Statistics and Information Technologies for Anti-Fraud (SITAF)
 European Commission JRC | Theseus

Public area | **Restricted area** | Development area | Admin area

HOME | SIGNALS IN IMPORTS | SIGNALS IN EXPORTS | SIGNALS IN CUSTOM DATA | FAIR PRICES | MONTHLY FAIR PRICES | SYSTEMATIC OVER/UNDER PRICING | CLUSTERS OF PRICES

Signals

Pattern: Price outliers | Flow: Import | Product code: 84825000 - Cylindrical roller bearings (excl. needle roller bearings) | Origin: Canada | Destination: ALL

Pattern	Flow	POD	Period	Quantity	Value	Unit Price €/kg
Price outliers	Imp	84825000-CA-SE	Mar 2018	0.10	2.01	19.72
Price outliers	Imp	84825000-CA-SE	Jul 2016	0.05	0.34	6.24
Price outliers	Imp	84825000-CA-SE	Mar 2016	0.07	0.87	13.15
Price outliers	Imp	84825000-CA-SE	Jun 2015	0.00	23.48	7.825.33
Price outliers	Imp	84825000-CA-RG	Jan 2017	0.00	1.64	5.68.00
Price outliers	Imp	84825000-CA-PL	Oct 2015	0.09	3.27	37.11

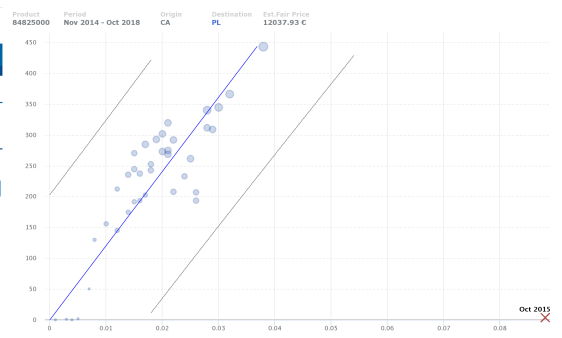


Figure 22: THESEUS. Left panel: list of identified outliers. Right panel: scatterplot of an homogeneous group of data plus a strong outlier.

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
- at the following standard number: +32 22999696, or
- by electronic mail via: https://europa.eu/european-union/contact_en

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU publications

You can download or order free and priced EU publications from EU Bookshop at: <https://publications.europa.eu/en/publications>. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

The European Commission's science and knowledge service

Joint Research Centre

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.



EU Science Hub

ec.europa.eu/jrc



@EU_ScienceHub



EU Science Hub - Joint Research Centre



EU Science, Research and Innovation



EU Science Hub



Publications Office
of the European Union

doi:10.2760/35922

ISBN 978-92-76-21438-0