Adv Data Anal Classif (2009) 3:263–279
DOI 10.1007/s11634-009-0050-y

REGULAR ARTICLE

New robust dynamic plots for regression mixture detection

Domenico Perrotta · Marco Riani · Francesca Torti

Received: 12 May 2009 / Revised: 23 October 2009 / Accepted: 27 October 2009 /
Published online: 12 November 2009

Abstract The forward search is a powerful general method for detecting multiple masked outliers and for determining their effect on inferences about models fitted to data. From the monitoring of a series of statistics based on subsets of data of increasing size we obtain multiple views of any hidden structure. One of the problems of the forward search has always been the lack of an automatic link among the great variety of plots which are monitored. Usually it happens that a lot of interesting features emerge unexpectedly during the progression of the forward search only when a specific combination of forward plots is inspected at the same time. Thus, the analyst should be able to interact with the plots and redefine or refine the links among them. In the absence of dynamic linking and interaction tools, the analyst risks to miss relevant hidden information. In this paper we fill this gap and provide the user with a set of new robust graphical tools whose power will be demonstrated on several regression problems. Through the analysis of real and simulated data we give a series of examples where dynamic interaction with different "robust plots" is used to highlight the

D. Perrotta
Global Security and Crisis Management Unit, Institute for the Protection and Security of the Citizens,
Joint Research Centre, European Commission, 21027 Ispra, Italy
e-mail: domenico.perrotta@ec.europa.eu

M. Riani (✉)
Dipartimento di Economia, Division of Statistics and Computing,
Università di Parma, Via Kennedy 6, 43100 Parma, Italy
e-mail: mriani@unipr.it

F. Torti
Dipartimento di Statistica, Università di Milano Bicocca, Via Arcimboldi 8, 20126 Milan, Italy
e-mail: f.torti1@campus.unimib.it

F. Torti
Dipartimento di Economia, Università di Parma, Via Kennedy 6, 43100 Parma, Italy

presence of groups of outliers and regression mixtures and appraise the effect that these hidden groups exert on the fitted model.

Keywords Forward search · Robustness · Exploratory data analysis · Data visualization · Statistical graphics · Brushing and linking

Mathematics Subject Classification (2000) 62F35 · 62-07 · 62J05 · 62J20

1 Introduction

The problem of representing data or any quantitative information in a graphical form suitable to human interpretation and exploration has deep roots and has been addressed in statistics and in many other scientific disciplines (Friendly 2005). One of the biggest challenges statisticians face when working on applied problems with non-statisticians is to be able to effectively present and communicate statistical results (Spence 2001; Tufte 1983). In this paper we develop new interactive tools which can dynamically connect the information which comes from different "robust plots". We show that the use of these tools can be very important not only to present the results in a more effective way, but also to explore the (multivariate) data in all their dimensions at once and find hidden structures such as complex regression mixtures.

The paper fits into the topic of data visualization (see Chen et al. 2008 for a comprehensive review on this topic) and data projections (Buja et al. 2009) and aims at extending the well known paradigms of linking and brushing (Wilhelm 2008) to the literature of robust statistics. More specifically, in this paper we focus on the forward search which is a powerful general method for detecting multiple masked outliers and for determining their effect on inferences about models fitted to data. Essentially, the forward search approach starts by considering only a subset of the observations and looks for the evolution of parameters, residuals or other statistics of interest when this subset is increased by choosing observations in the next subset with ordering criteria that in general are very simple to implement and quick to execute. Use of the forward search is described in Atkinson and Riani (2000) for linear and non-linear regression, response transformation and in generalized linear models. Related forward techniques for multivariate data are given in Atkinson et al. (2004). Riani et al. (2008) also proposed a natural extension of the forward search to the estimation of regression mixtures.

In the forward search we monitor the evolution of residuals, parameters estimates and inferences as the subset size increases, presenting our results as "forward plots" which show the evolution of the quantities of interest as a function of subset size. Therefore, unlike other robust approaches, the forward search is a dynamic process that produces a sequence of estimates and related plots.

One of the problems of the forward search has always been the lack of an automatic link among the great variety of plots which are monitored and of tools to facilitate the extraction of the information which comes from such plots. In this paper we fill this gap and provide the user with a set of new robust graphical tools whose power will be demonstrated on several regression problems, including the detection of groups of outliers and the identification and estimation of regression mixtures.

The output is a series of linking and interaction tools which connect together different views of the data represented in several classical and robust plots. Such tools, not only help to better present the data, but also enable the researcher to highlight the presence of hidden subgroups of data. In fact, the objects of real interest for the statistician or for the end-user can emerge unexpectedly during the progression of the forward search or only when a specific combination of plots is inspected jointly. Dynamic interaction can be very useful and helpful to appraise the effect that these hidden groups exert on the fitted model.

The structure of the paper is as follows. In Sect. 2, after introducing the notation, we briefly recall the main quantities which are monitored along the forward search algorithm in regression and we stress the need of linking the information which comes from the monitoring of the different statistics. In Sect. 3 we give a series of examples from real and simulated data where, thanks to the dynamic links among different robust plots and interactive graphics, it is immediately and easily possible to reveal the real structure of the data and to appraise the effect that subgroups of units exert on particular statistics. Sect. 4 concludes and provides food for thought for future research.

Given that the role of colours is often crucial in representing visual information in black/white, we recommend the reader to have a look also at the colored version of the paper available on the ADAC website of the publisher.

2 From static to dynamic linked robust forward plots

To introduce the notation let us say that in the regression model $y = X\beta + \epsilon$, y is the $n \times 1$ vector of responses, X is an $n \times p$ full-rank matrix of known constants, with ith row x_i^T, and β is a vector of p unknown parameters. The normal theory assumptions are that the errors ϵ_i are i.i.d. $N(0, \sigma^2)$.

The least squares estimator of β is $\hat{\beta}$. Then the vector of n least squares residuals is $e = y - \hat{y} = y - X\hat{\beta} = (I - H)y = Ay$, where $H = X(X^TX)^{-1}X^T$ is the 'hat' matrix, with diagonal elements h_i and off-diagonal elements h_{ij}. The residual mean square estimator of σ^2 is $s^2 = e^Te/(n-p) = \sum_{i=1}^n e_i^2/(n-p)$.

The forward search in regression typically starts by fitting a small subset of m_0 of the n observations in the data. The subset is chosen robustly, using Least median of squares (LMS) or Least trimmed squares (LTS) estimators (Rousseeuw 1984). For general m, with $m_0 \leq m \leq n$, let $S_*^{(m)}$ be the subset of size m, for which the matrix of regressors is $X(m^*)$ and the response is $y(m^*)$.[1] Least squares on this subset of observations yield parameter estimates $\hat{\beta}(m^*)$ and $s^2(m^*)$, the mean square estimate of σ^2 on $m - p$ degrees of freedom. Residuals can be calculated for all observations

[1] The notation $X(m^*)$ means that we consider the m observations from the subset $S_*^{(m)}$, and similarly for $y(m^*)$, etc.

including those not in $S_*^{(m)}$. The n resulting least squares residuals are

$$e_i(m^*) = y_i - x_i^T \hat{\beta}(m^*), \quad i = 1, \ldots, n. \tag{1}$$

The estimates of the parameters are based on only those observations giving the central m residuals. The search moves forward with the augmented subset $S_*^{(m+1)}$ consisting of the observations with the $m+1$ smallest absolute values of $e_i(m^*)$. Notice that we use the plural form '$m+1$ smallest values' because from step m to step $m+1$ some units may leave the subset.

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

$$\hat{\beta}_{\text{FS}} = \left(\hat{\beta}(p^*), \ldots, \hat{\beta}(n)\right), \tag{2}$$

where $\hat{\beta}(p^*) = \hat{\beta}_{\text{LMS}}$ or $\hat{\beta}(p^*) = \hat{\beta}_{\text{LTS}}$. During the search we can monitor any relevant statistic. Typically, we monitor quantities indicative of model quality or inadequacy, such as residuals or the score test for transformation. In regression we can also monitor the evolution of $s^2_{S^{(m)}}$, the estimate of the error variance. Because the search orders the observations by the magnitude of their residuals from the fitted subsets, the value of $s^2_{S^{(m)}}$ increases during the search, although not necessarily monotonically. As a consequence, even in the absence of outliers and model inadequacies, the values of the t tests for the parameters in the model decrease during the search and are hard to interpret. In Atkinson and Riani (2002) the method of added variables is used to provide plots of t tests which are orthogonal to the search and have the required t-distribution. In order to judge the importance of a particular variable (say w), it is customary to use the so called added variable plot, that is the plot showing the residuals of y against all predictors except w, versus the residuals of w on all other predictors. More specifically, given the model

$$y = X\beta + w\gamma + \epsilon,$$

where γ is a scalar, the added variable plot is the plot of $(I_n - H)y = Ay = e$ (residual response) against Aw (residual added variable). Given that residuals are subject to the well known masking and swamping problems, the information which comes out form this plot can be highly misleading. Clearly, it is very cumbersome to monitor during the search for each subset size m the plot of $(I_m - H(m^*))y(m^*)$, where $H(m^*) = X(m^*)\left(X(m^*)^T X(m^*)\right)^{-1} X^T(m^*)$ against $(I_m - H(m^*))w(m^*)$. On the other hand, once the monitoring of deletion t tests or any other forward plot has highlighted the importance of particular units at step, say, m^*, it is useful to project the remaining $n - m^*$ units on the added variable plot based on m^* units in order to better appraise the importance they exert on the significance of a particular explanatory variable. Up to now this could be done only off line in a very ad hoc and cumbersome way. Thanks to our new dynamic interactive tools, once a set of trajectories are highlighted, it is possible to project them in the added variable plot and immediately appraise their importance. In Sect. 3.2 we will see an example of this technique.

In the context of response transformation, when the null hypothesis is that $\lambda = \lambda_0$, the added t tests based on constructed variables are known in the statistical literature as "score tests for transformation". A powerful robust tool to understand the percentage of observations which are in accordance with the different values of the transformation parameters is the forward plot of the score test statistic for transformation of the set of constructed variables for different values λ_0, using a separate search for each λ_0. These trajectories of the score tests can be combined in a single picture named the "fan plot" (Atkinson and Riani 2000). Up to now it was very cumbersome to connect relevant features in the fan plot to the position of the units in the scatter plot matrix. This limited considerably the possibility of exploring all the patterns which came out unexpectedly from the fan plot.

One of the most important plots monitors all residuals at each step of the forward search. Large values of the residuals among cases not in the subset indicate the presence of outliers, as do nonsmooth changes in the value of the residual sum of squares. Because of the strong dependence of $s^2_{S^{(m)}}$ on m, we standardize all residuals by the final root mean square estimate s^2. One of the drawbacks of this plot is that when n is large there are too many lines and therefore this plot tends to become confused and misleading.

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

$$r_i^*(m^*) = \frac{y_i - x_i^T \hat{\beta}(m^*)}{\sqrt{s^2(m^*)\{1 + h_i(m^*)\}}} = \frac{\mathrm{e}_i(m^*)}{\sqrt{s^2(m^*)\{1 + h_i(m^*)\}}}, \tag{3}$$

where $h_i(m^*) = x_i^T \left\{X(m^*)^T X(m^*)\right\}^{-1} x_i$, the leverage of each observation, depends on $S_*^{(m)}$. Let the observation nearest to those constituting $S_*^{(m)}$ be $i_{\min}$ where

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

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

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

$$r^*_{\mathrm{imin}}(m^*) = \frac{\mathrm{e}_{\mathrm{imin}}(m^*)}{\sqrt{s^2(m^*)\{1 + h_{\mathrm{imin}}(m^*)\}}}. \tag{4}$$

Riani and Atkinson (2007), using the properties of order statistics from folded t distributions, found good approximations to the forward envelopes of the minimum deletion residual. In order to keep into account the problem of multiple testing during the various steps of the search and to control the overall size and power of the test Riani et al. (2009) suggested a procedure for automatic outlier detection during the forward search.

Up to now the forward plot of residuals has always been static and unlinked with that of the minimum deletion residual. Only manually and in a rather cumbersome off line way it was possible to link the different features which came out from the monitoring of these two plots. Some of these features, such as outliers entering at the end of the search, were easy to identify and link among the various plots. However, groups of points of potential interest and relevant events occurring at specific steps of the process, which sometimes emerge unexpectedly during the progression of the forward search, can only be appraised by inspecting these two plots simultaneously, possibly in combination with other plots and by interacting with the graphics. Often, it is crucial to find the right trade-off between detail and scope of the data region inspected, by means of zoom and unzoom tools.

As concerns the order of entry of the units, in most moves from m to $m+1$ just one new unit joins the subset. However it may also happen that two or more units join $S_*^{(m)}$ as one or more leave. Such an event is quite unusual, only occurring when the search includes one unit that belongs to a cluster of outliers. When this happens, at the next step the remaining outliers in the cluster seem less outlying and so several may be included at once. In this case several other units then have to leave the subset. Up to now the analysis of the steps in which a certain set of units entered the subset was done off line. Similarly, off-line was the analysis of the joint effect on the regression fit of a subgroup of units. The information in the different forward plots was retrieved with ad hoc routines written to appraise the effect that a particular subgroup had exerted on another statistic monitored in another plot. Therefore, in practice it was almost impossible to detect and analyze complex structures such as regression mixtures.

All these limitations used to cause difficulties to effectively communicate statistical results, especially for an audience with limited statistical experience, and may have even limited the use of the forward search method in the statistical community.

3 Examples

In this section we compare the output of the "traditional" forward search plots with that obtained with the new dynamic linking and selection tools. The emphasis is on the information gained from the adoption of such tools.

3.1 Hawkins data

The set of simulated data analyzed in this section were intended by Hawkins to be misleading for standard regression methods. An analysis is given by Atkinson and Riani (2000, §3.1). There are 128 observations and nine explanatory variables.

The traditional forward plot of scaled residuals clearly shows that there is a lot of interchange in the order of magnitude of the residuals during the forward search. Similarly, the minimum deletion residual among observations not in the subset, that is the outlier test statistic (4), shows three clear peaks. These plots however, are static in the sense that do not enable for example to understand from which units the three peaks, which clearly appear in the minimum deletion residual plot, are formed. Similarly, it

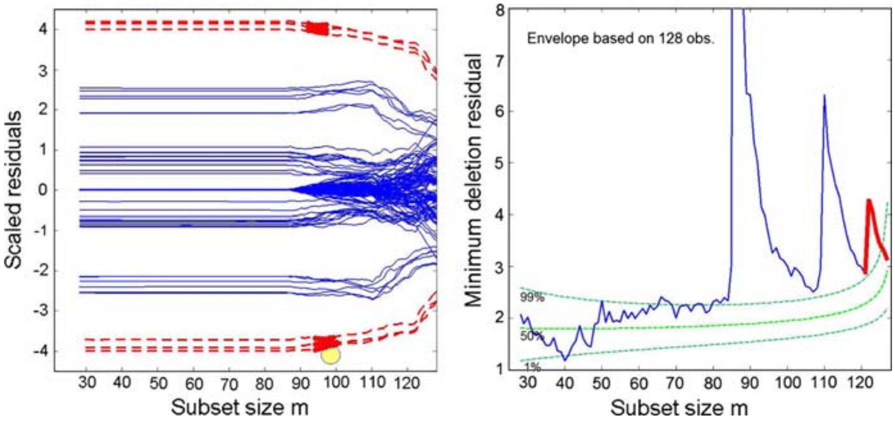


Fig. 1 Hawkins data. Dynamic forward plots of scaled residuals (*left panel*) and minimum deletion residual outside subset with 1, 50 and 99% confidence envelopes (*right panel*). The set of trajectories brushed in the *left panel* is automatically shown in *bold face* in the *right panel*. This set of brushed units corresponds to the last peak in the monitoring of minimum deletion residual

is not clear which trajectories of the residuals correspond to the three peaks. So far all this information had to be checked manually analyzing the units which entered the search. At present, thanks to the interactive connection which we have created among the different plots it is possible to select a set of trajectories in the monitoring residuals plot (left panel of Fig. 1) and at the same time to see them highlighted in the monitoring of the minimum deletion residual (right panel of Fig. 1). For example, Fig. 1 shows that the most extreme (negative and positive) residuals enter the last 6 steps of the search and are associated with the final peak. Similarly, if we do a second brushing, we can see that the set of trajectories selected in the left hand panel of Fig. 2 correspond to units which enter consecutively and are associated with the second peak in the right panel of Fig. 2.

Figure 3 shows the output of a third additional brushing which is connected with the first peak in the minimum deletion residual plot. Every time a brushing action is performed on the monitoring residuals plot, it is possible to display in an automatic way also the information about the position of the brushed units in the scatter diagram of y against the required explanatory variable (s) (see for example the bottom panel of Fig. 3). Finally, it is worthwhile to stress that even if in this example we started brushing from the monitoring of scaled residuals, it is also possible to start brushing from the minimum deletion residual plot and see the corresponding trajectories highlighted in the residuals plot.

3.2 Multiple regression data

This dataset (introduced by Atkinson and Riani 2000) contains 60 observations on a response y with the values of three explanatory variables. The scatterplot matrix of the data and the traditional diagnostic plots show no apparent outliers. On the other hand, the monitoring of scaled squared residuals clearly reveals that there are 6 trajectories

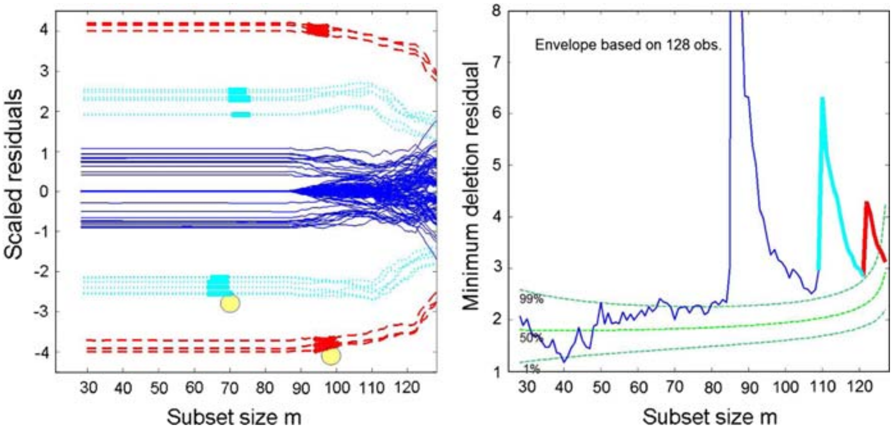


Fig. 2 Hawkins data. Dynamic forward plots of scaled residuals (*left panel*) and minimum deletion residual outside subset with 1, 50 and 99% confidence envelopes (*right panel*). The additional set of trajectories brushed in the *left panel* is automatically shown in *bold face* in the *right panel*. This last set of brushed units corresponds to the second peak in the monitoring of minimum deletion residual

whose residual seems to be different. From these traditional static plots it is not clear which effect this group of 6 units exerts for example in the minimum deletion residual and or in other statistics. Finally, it is not clear in which steps the entry into the subset of the six potential outliers took place and or if the entry of the outliers caused some units already in the subset to leave it.

At present all these questions can be answered simply brushing the set of required units or particular steps in a plot. When we brush the trajectories of the 6 potential outliers in the monitoring residuals plot (left panel of Fig. 4), it is possible to notice (centre panel of Fig. 4) that, when the first of this group of units enters the subset, the value of minimum deletion residual lies just above the 99% envelope. On the other hand, due to the well known masking effect when all the 6 potential outliers enter the subset, the value of the minimum deletion residual is well below the 1% envelope. The effect of the 6 units on the significance of variable $X1$ is shown egregiously in the right panel of the Figure which gives the added variable plot at step $m = n - 6 = 54$ together with the regression line coming from the addition of variable $X1$ when these 6 potential outliers are included or excluded. When they are excluded the slope is positive and the strength of the association of variable $X1$ measured by its added t statistic is equal to 2.25 (its p-value is equal to 0.029). On the other hand, when they are included, the slope turns out to be negative and the added t statistic becomes equal to -1.26 (p-value $= 0.21$).

The left panel of Fig. 4 shows that there is a unit (observation 43) whose residual increases in the final steps of the search. When this trajectory is brushed (top left panel of Fig. 5), we can immediately see that it corresponds to a unit which is included for the first time into the subset in the step prior to the inclusion of the six potential outliers (top right panel of Fig. 5). When all the group of potential outliers is included, unit 43 goes out of the subset and reenters in the final step. The final brushing is associated to the set of trajectories which in the monitoring of scaled squared residuals seems to

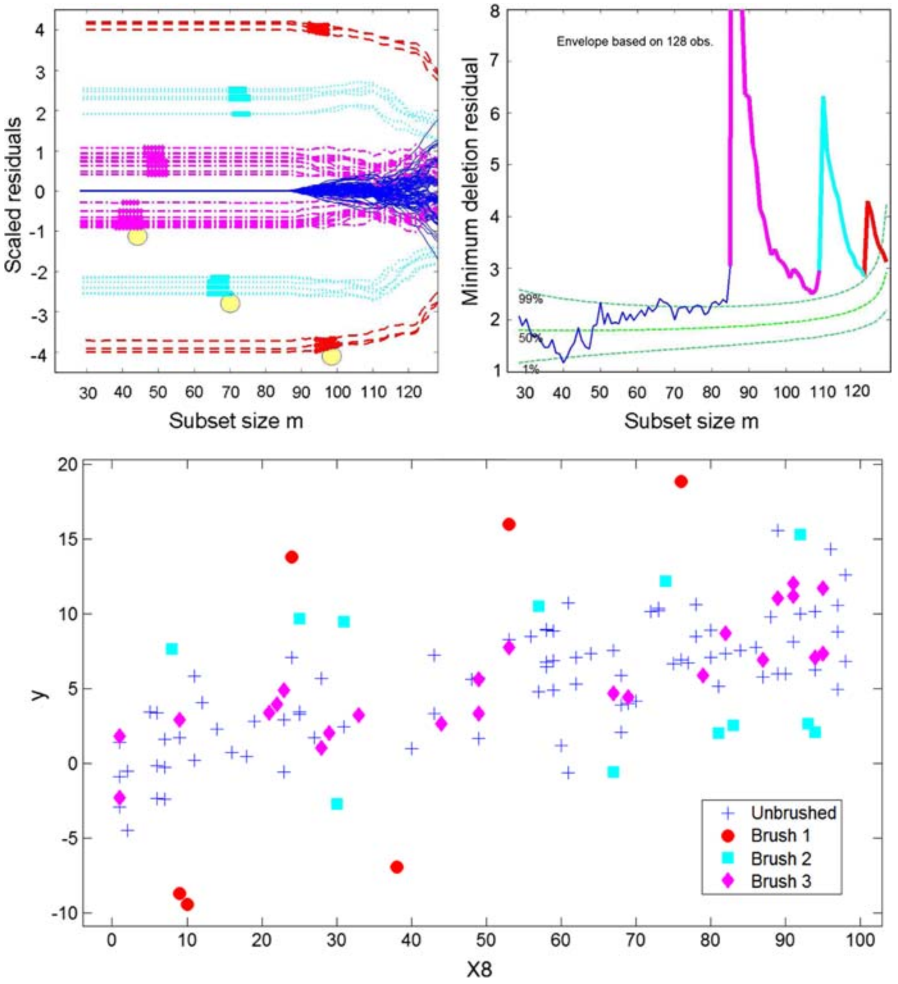


Fig. 3 Hawkins data. Dynamic forward plots of scaled residuals (*left panel*) and minimum deletion residual outside subset (*right panel*) and scatter of y versus $X8$ (*bottom panel*). The three sets of trajectories brushed in the *left panel* is automatically shown in *bold face* (*right panel*). It is also possible to see in an automatic way the position on the required scatter plot of the brushed units (*bottom panel*)

lie between 4 and 10 (bottom left panel of Fig. 5). The right bottom panel of Fig. 5 shows that this set of trajectories enters in steps 43–50 and is associated with the small peak out of the upper 99% envelope in the central part of the search of the minimum deletion residual.

3.3 Loyalty cards data

In this section we consider 509 observations on the behaviour of customers with loyalty cards from a supermarket chain in Northern Italy. The response (y) is the amount, in

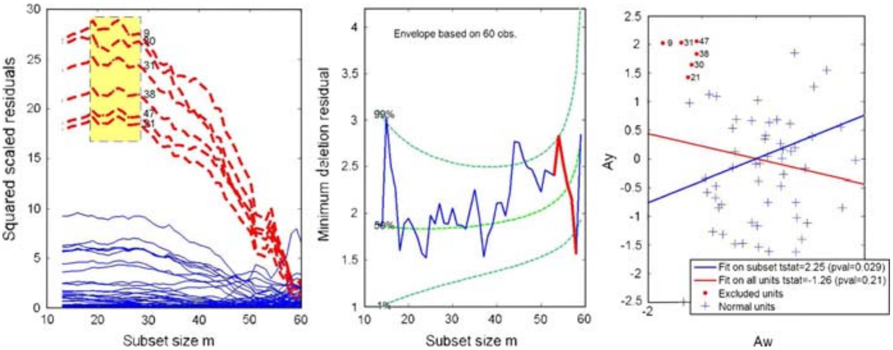


Fig. 4 Multiple regression data. Dynamic forward plots of scaled squared residuals (*left panel*), minimum deletion residual outside subset with superimposed 1, 50 and 99% envelopes (*centre panel*) and added variable plot at step $m = 54$ (*right panel*), for the addition of variable $X1$ with the six outlier projected. The effect of the group of 6 potential outliers on the judgement of the importance of variable $X1$ is clear

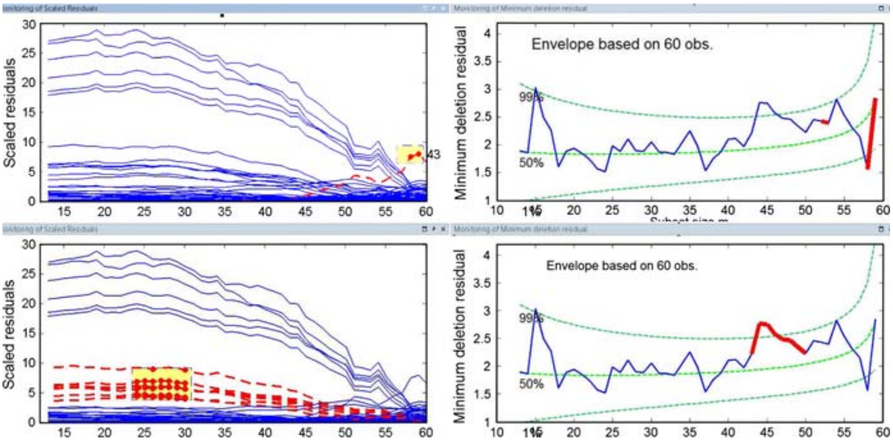

Scaled residuals
43
Minimum deletion residual
Envelope based on 60 obs.
99%
50%
Scaled residuals
Minimum deletion residual
Envelope based on 60 obs.
99%
50%

Fig. 5 Multiple regression data. Dynamic forward plots of squared scaled residuals (*left panels*), minimum deletion residual outside subset with superimposed 1, 50 and 99% envelopes (*right panels*). In the *top left panel* the trajectory corresponding to unit 43 is brushed. In the *bottom left panel* the set of residuals associated with the central peak in the minimum deletion residual is brushed

euros, spent at the shop over 6 months and the explanatory variables are: $X1$, the number of visits to the supermarket in the six month period; $X2$, the age of the customer and, $X3$, the number of members of the customer's family. Given that the response is likely not to have a homogeneous variance, it may need to be transformed. Often the analysis of the data can be strongly improved if the response variable is transformed. If the sample size is small (i.e. $n \leq 100$), generally it is enough to consider the five most common values of the transformation parameter (that is $\lambda = (-1, -0.5, 0, 0.5, 1)$). On the other hand, when the sample size increases, a finer grid of values of λ must be considered. In both cases, it is necessary to understand which are and what is the effect of the potential outliers on the different scales. If there are indications that the

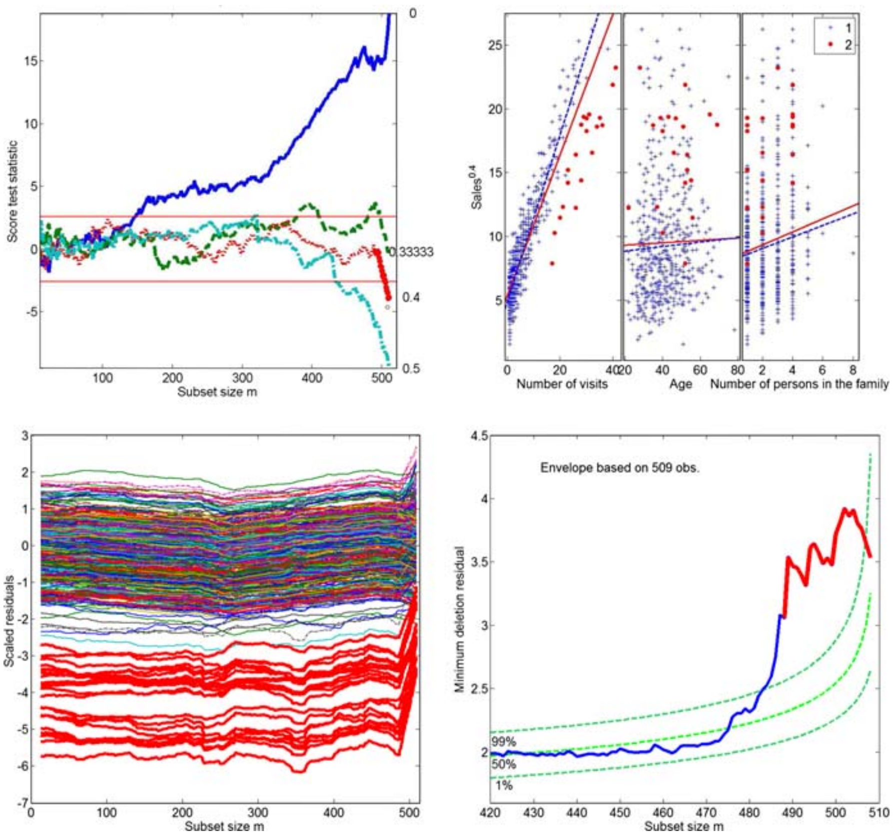


Fig. 6 Loyalty cards data. Fan plot for $\lambda = (0, 1/3, 0.4, 0.5)$ (*top left panel*) with the last steps of the search for $\lambda = 0.4$ brushed. Scatterplot of $y^{0.4}$ (transformed "Sales") versus all the three explanatory variables with the brushed units highlighted (*top right panel*). Monitoring residuals plot with highlighted the trajectories corresponding to the brushed units in the fan plot (*bottom left panel*). Monitoring minimum deletion residual plot with highlighted the steps corresponding to the brushed units in the fan plot (*bottom right panel*)

regression data should be transformed, it is important to remember that outliers in one transformed scale may not be outliers in another scale.

The top left panel of Fig. 6 shows the fan plot for the values of $\lambda = (0, 1/3, 0.4, 0.5)$. This plot clearly shows that both the log and square root transformation are not diffused throughout the data. The value of the third root tends to lie close to the 99% rejection line and at the end lies completely inside the confidence bands. The curve associated with $\lambda = 0.4$ is always inside the bands and at the end goes outside. Both for $\lambda = 1/3$ and $\lambda = 0.4$ there seems to be a considerable decrease in the trajectory of the score test in about the last 20 steps. From this static fan plot it is not clear whether the units which enter the final part of the search for these two values of λ are the same. Finally, it is necessary to understand where this group lies in terms of the scatter plot matrix, what is its effect on the significance of the explanatory variables and whether these units must be considered atypical. The dynamic brushing which we have implemented

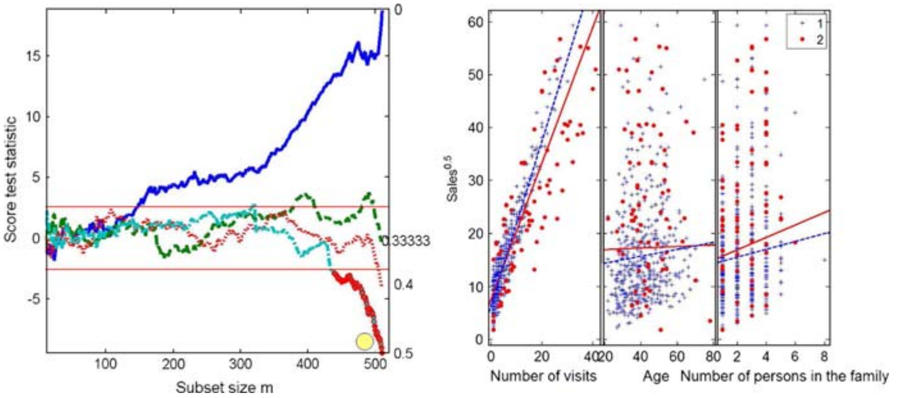


Fig. 7 Loyalty cards data. Fan plot for λ = (0, 1/3, 0.4, 0.5) (*left panel*) with the steps of the search below the 1% confidence band for λ = 0.4 brushed. Scatterplot of $\sqrt{y}$ (transformed "Sales") versus all the three explanatory variables with the brushed units highlighted (*right panel*). The brushed units are are much more spread out in the three scatterplots

enables easily to answer all these questions. The units associated with the last steps of the search for λ = 0.4, which are brushed on the top left panel of Fig. 6, immediately appear highlighted on the scatter plot (top right hand panel). This enables us to understand that the brushed steps are associated with consumers who behave in a strikingly different way from the majority of the population. More precisely, the brushed group is characterized by consumers of varying age and size of the family who spend much less than it would be expected given their high number of visits to the supermarket of the chain. The top right panel also shows the projected regression line when this brushed group is included (solid lines) or excluded (dashed lines) from the analysis. Clearly, the effect of the brushed units is to decrease considerably the slope of the coefficient associated with the first explanatory variable. The bottom left panel clearly shows that the brushed units have a very large negative residual throughout all the central part of the search and that at the end their trajectories are completely confused with those of the rest of the data. Often, when the sample size is large, the monitoring residuals plot due to the high number of curves becomes difficult to interpret. Clearly, as it happens in this case, the highlighting of a set of units helps a lot its readability. Finally, if the user is interested in outlier detection, the right bottom panel of Fig. 6 shows that the curve of the minimum deletion residual goes out of the 99% threshold in the final part of the search. If the curve of minimum deletion residual is redrawn without the set of brushed units (the figure is not given here due to lack of space) it is possible to see that it stays completely inside the new resuperimposed envelope during all the search.

One may wonder how crucial is the choice of the transformation value. We therefore check the impact of the units that, in the fan plot for the slightly upper value λ = 0.5, cause the value of the score test statistic to fall below the 1% rejection line (left panel of Fig. 7). This happens earlier than in λ = 0.4 case. In the square root case, even if the selected units have about the same impact in decreasing the slope of the coefficient associated with the first explanatory variable (see right panel of Fig. 7), they cause the slope associated with the explanatory variable “age” to become almost parallel to the x axis and they increase considerably the coefficient for the regressor “Number of persons in the family”. As concerns the position of the new brushed units they are more spread out in the three scatterplots and, thus, they cannot be associated to a typical consumers behavior. In other words, while the rejection of the null hypothesis $\lambda = 0.4$ seems to be due to the presence of a cluster of outliers, the rejection of the null hypothesis of square root transformation is not due to the presence of particular units, but seems to be diffused throughout the data.

3.4 International trade data

In this subsection we illustrate the advantages of adopting our flexible interactive plots in a context in which the patterns of interest are complex (regression mixtures) and the focus is on the interpretation of the results by the end-user. We argue that with our interactive graphical tools the end-user can combine effectively subject matter knowledge with information provided by the statistical method and draw conclusions of relevant operational value.

We use an example taken from the thousands of international trade datasets that we analyze in the context of a collaboration between two services of the European Commission: the Joint Research Centre (JRC) and the European Anti-Fraud Office (OLAF), with the purpose of highlighting potential cases of fraud, data quality issues and other oddities related to specific international trade contexts. Every country keeps record of the trade of its companies with other countries and such huge amount of data is archived in many commercial and public datasets. Trade data are sometimes aggregated, depending on the needs and the mission of specific international organizations such as the European Commission’s statistical office, EUROSTAT. The example in this subsection refers to monthly data aggregated at country level.

For each product, country of origin and destination (POD), we use the values and quantities traded to estimate the slope of the regression line, that is a sort of “fair price” for that particular combination, and we detect low and high price outliers. The situation is complicated by the fact that, for a given POD, there might be flows referring to different trade prices, which result in mixtures of linear populations. Abnormal price transactions are made available to OLAF and to its partners in the Member States for analyzes motivated by the identification of potential fraud cases. In order to illustrate the benefits which come from dynamic brushing we consider the values (in thousands of Euros) and the quantities in tons of a fishery product imported into the European Union (EU) from a third country.

Given that the relationship between value and quantity is linear, trade data should not require transformation. The need of transformation may be due to the presence of multiple populations: that is fair trade declarations together with price under-declarations to reduce import duties and over-declarations to carry out money laundering activities.

The left panel of Fig. 8 shows that the inclusion of the last observations which enter the forward search, causes strong rejection of the hypothesis of no transformation. After a simple brushing action on the trajectory of the fan plot for $\lambda = 1$, it is

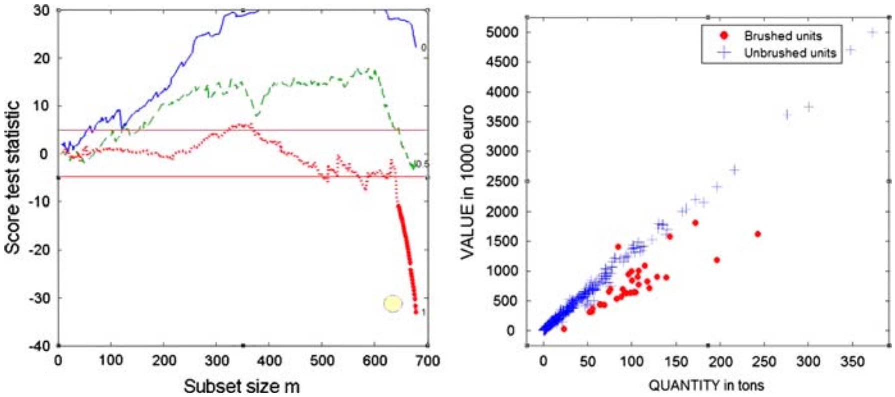


Fig. 8 Fishery trade data. Fan plot for $\lambda = (0, 0.5, 1)$ (*left panel*) and the scatter of values (in thousands of euros) against quantities (in tons). The brushed units in the fan plot are automatically highlighted in the scatter plot. The brushed units form a separate cluster from the main bulk of the data

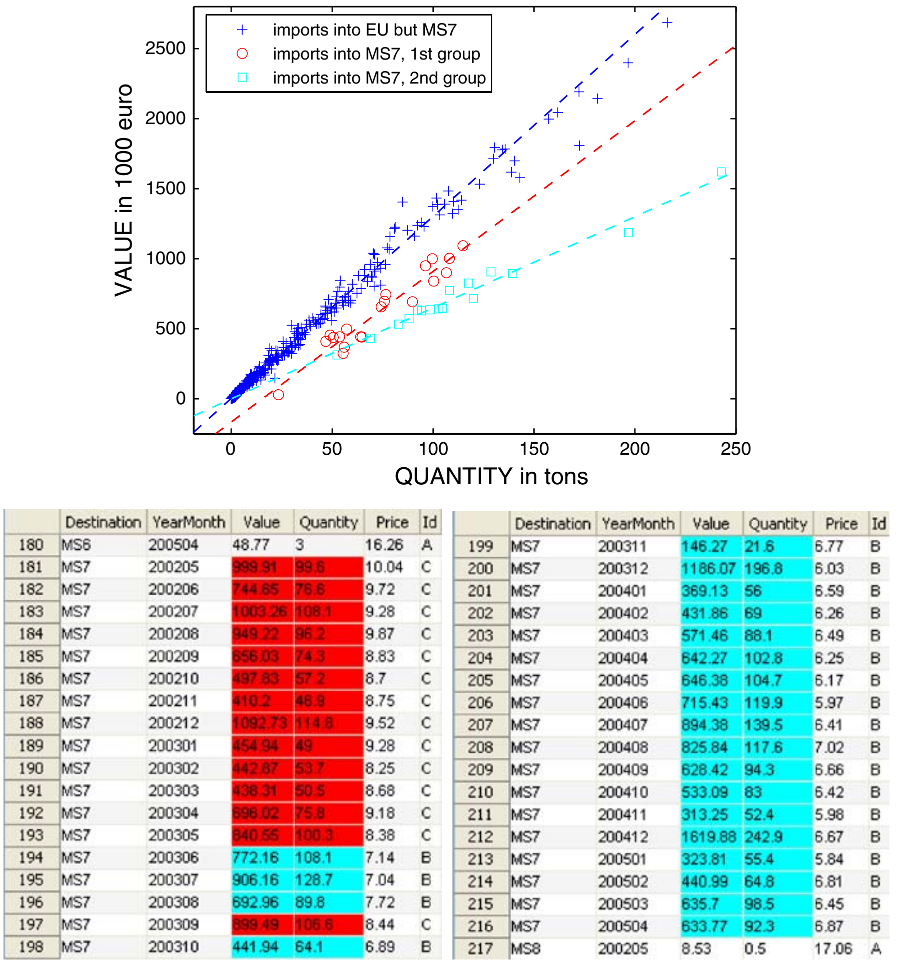
immediate to see that the brushed units form a well defined cluster below the main cloud in the scatter plot (right panel of Fig. 8) which shows the quantity (x axis) and the value (y axis) of the monthly flows.

Perrotta and Torti (2009), using repeated application of the forward search, could identify the group of the “fair” import flows (which in the scatterplot are represented with symbol ‘+’) to the other Member States, estimate their price (13 €/kg) and finally divide the group of abnormal flows in other two linear clusters (which in Fig. 9 are represented respectively with circles and squares) of significantly different estimated price, 9.17 vs. 6.55 €/kg. These two groups correspond to imports into a specific EU Member State (say MS7). In this case the prices declared look suspiciously low if compared to those of the other Member States.

However, there is a relevant fact that becomes clear only by inspecting in the input table (see bottom panel of Fig. 9) the transactions corresponding to the circles and squares, by comparing their prices and the time periods when the transactions were recorded. In fact, the two abnormal clusters include flows that took place respectively in the first 14 consecutive months of the period analyzed (circles) and in the following consecutive 21 months (squares). A natural conclusion of clear operational value is that in the period analyzed the traders in MS7 gradually lowered their import price declarations, up to half of the reported price by the other Member States.

In Perrotta and Torti (2009) such unexpected pattern was found almost incidentally, by off-line inspection of the data table. Clearly this way of proceeding is not efficient: it requires big efforts and the chance of finding unexpected patterns of this type is limited. For example, without interactive tools it would be difficult to realize that near the origin of the axes there is a large number of linear groups of data, visible only by zooming that region. We omit here to discuss this and other interesting patterns that can be found by inspecting other variables for other groups of observations.

Our solution for the end user consists in linking the familiar tools such as the scatterplot of the import values and quantities and the table of the original data, where all relevant trade variables can be found. In this paper we have illustrated the



	Destination	YearMonth	Value	Quantity	Price	Id
180	MS6	200504	48.77	3	16.26	A
181	MS7	200205	999.91	99.6	10.04	C
182	MS7	200206	744.65	76.6	9.72	C
183	MS7	200207	1003.26	108.1	9.28	C
184	MS7	200208	949.22	96.2	9.87	C
185	MS7	200209	656.03	74.3	8.83	C
186	MS7	200210	497.83	57.2	8.7	C
187	MS7	200211	410.2	46.9	8.75	C
188	MS7	200212	1092.73	114.8	9.52	C
189	MS7	200301	454.94	49	9.28	C
190	MS7	200302	442.87	53.7	8.25	C
191	MS7	200303	438.31	50.5	8.68	C
192	MS7	200304	696.02	75.8	9.18	C
193	MS7	200305	840.55	100.3	8.38	C
194	MS7	200306	772.16	108.1	7.14	B
195	MS7	200307	906.16	128.7	7.04	B
196	MS7	200308	692.96	89.8	7.72	B
197	MS7	200309	899.49	106.6	8.44	C
198	MS7	200310	441.94	64.1	6.89	B

	Destination	YearMonth	Value	Quantity	Price	Id
199	MS7	200311	146.27	21.6	6.77	B
200	MS7	200312	1186.07	196.8	6.03	B
201	MS7	200401	369.13	56	6.59	B
202	MS7	200402	431.86	69	6.26	B
203	MS7	200403	571.46	88.1	6.49	B
204	MS7	200404	642.27	102.8	6.25	B
205	MS7	200405	646.38	104.7	6.17	B
206	MS7	200406	715.43	119.9	5.97	B
207	MS7	200407	894.38	139.5	6.41	B
208	MS7	200408	825.84	117.6	7.02	B
209	MS7	200409	628.42	94.3	6.66	B
210	MS7	200410	533.09	83	6.42	B
211	MS7	200411	313.25	52.4	5.98	B
212	MS7	200412	1619.88	242.9	6.67	B
213	MS7	200501	323.81	55.4	5.84	B
214	MS7	200502	440.99	64.8	6.81	B
215	MS7	200503	635.7	98.5	6.45	B
216	MS7	200504	633.77	92.3	6.87	B
217	MS8	200205	8.53	0.5	17.06	A

Fig. 9 Fishery trade data. The scatterplot of values and quantities (*top panel*) and the data table with other relevant variables (*bottom panels*). Selected units in the scatterplot are automatically highlighted in the table and vice-versa. The 'Id' variable in the table indicates how the three groups of flows were identified by the forward search: a + in the scatterplot has id 'A' in the table, the *squares* correspond to a 'B' and the *circles* to a 'C'

brushing and linking tools starting from the forward plot of scaled residuals, minimum deletion residual, or fan plot. However, we have also implemented the possibility of brushing directly from the scatter plot matrix. For example, if in Fig. 9 we brush the units corresponding to the squares symbols, we can see that they have the largest negative residuals throughout the search (left panel of Fig. 10) and enter the search in the final steps (right panel of Fig. 10). Inside this group, the observations with the highest absolute scaled residuals value throughout the search correspond to flows 200 and 212. These two observations enter the subset only in the last two steps of the procedure. Similarly, the brushing of the circles in Fig. 9 reveals that this group is characterized by trajectories with negative residuals intermediate to those of the other

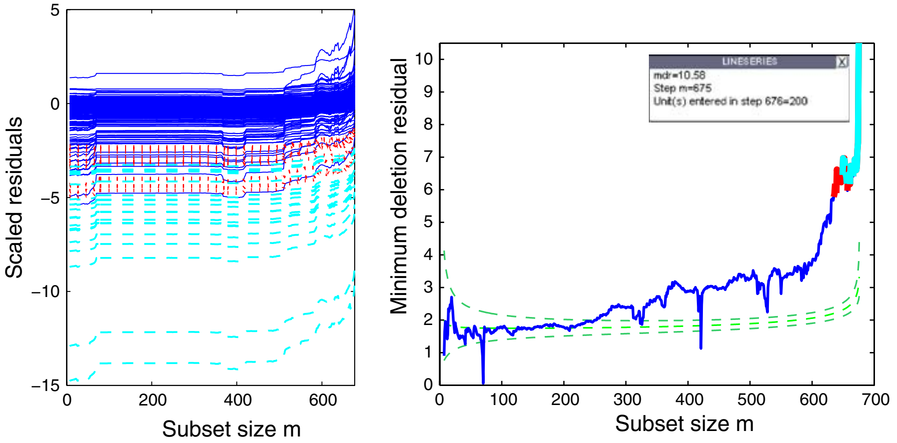


Fig. 10 Fishery trade data. The scaled residuals trajectories (*left*) and the minimum deletion residual curve at each step of the FS (*right*). The units highlighted in the two plots refer to those selected in the scatterplot. The link is automatic also in this case. The units denoted with *squares* in Fig. 9 are the last to enter the search and are those with the largest negative residuals (*dashed lines* in the *left panel*). Inside this group our data tooltip (*right panel*), which appears automatically when the user clicks on a point in the curve, shows detailed information about the associated observation. The units denoted with circles in Fig. 9 (*dotted lines* in the *left panel* of scaled residuals) enter the search just before those shown with a *square* (*dashed lines* in the *left panel*)

two groups and is formed by units which enter the search just before the group of the squares.

4 Conclusions

All the dynamic plots and the interactive tools discussed in this paper have been developed in MATLAB and are part of a toolbox which can be used to experiment the forward search on regression problems. The toolbox contains an extensive documentation integrated in the standard MATLAB help system, for each function and also for the “forward search philosophy of data analysis”. The toolbox is freely available at http://www.riani.it/matlab.htm, where the user can also watch or download videos which show in a dynamic way the brushing actions which have been illustrated in this paper. A book on practical visualization and exploratory data analysis in MATLAB that can be used as a reference and baseline for our paper is Martinez and Martinez (2004).

The work discussed covers about a year of software development and experimentation, but the project is far from being complete. Currently, we are working on the integration of the forward search to the multivariate context, so that to extend the applicability of the method. This will force us to address new plots which are more typical of multivariate visualization. We will progressively include various robust techniques, so that to enable comparisons among the main different robust approaches and definition of rigorous benchmark experiments. Another objective is to provide a stand alone run-time version of our toolbox so that also researchers, without the MATLAB licence, can use it.

Our project also targets end-users who are unsatisfied with standard static output or with the way traditional statistical output is presented or with the standard interpretation of the statistical results and their translation into subject matter terminology. We consider the linking and brushing extension to data tables proposed in the international trade example only as a first step. In the future it will be necessary to couple the new brushing and linking tools with the possibility of interactively incorporate subject matter knowledge as new data variables. For example, the antifraud expert might want to mark the outliers which are not fraud (false signals, from his perspective), define a model for the new data and run again the outlier detection tool. We are seeking at interactivity and visualization approaches sufficiently general and flexible to target a wide user community.

Acknowledgments The authorship is given in alphabetic order. The work was jointly supported by the University of Parma and the Joint Research Centre of the European Commission, under the institutional work-programme 2007–2013 of the research action “Statistics and Information Technology for Anti-Fraud and Security”. We are grateful to Professors A.C. Atkinson and A. Cerioli and to Dr. Vytis Kopustinskas for providing helpful suggestions on a preliminary version of the paper. We also wish to thank the referees and Professor Hans-Hermann Bock for their very thorough reading of our paper and constructive comments which have led to an appreciable improvement in the presentation of our work.

References

Atkinson AC, Riani M (2000) Robust diagnostic regression analysis. Springer, New York

Atkinson AC, Riani M (2002) Forward search added variable *t* tests and the effect of masked outliers on model selection. Biometrika 89:939–946

Atkinson AC, Riani M, Cerioli A (2004) Exploring multivariate data with the forward search. Springer, New York

Buja A, Cook D, Asimov D, Hurley C (2009) Theory of dynamic projections in high-dimensional data visualization. Electron J Stat

Chen C, Härdle W, Unwin A (eds) (2008) Handbook of data visualization, vol XIV of springer handbooks of computational statistics. Springer, Berlin

Friendly M (2005) Milestones in the history of data visualization: a case study in statistical historiography. In: Weihs C, Gaul W (eds) Classification: the ubiquitous challenge. Springer, New York, pp 34–52

Martinez WL, Martinez AR (2004) exploratory data analysis with MATLAB. Computer science and data analysis series. Chapman & Hall/CRC, London

Perrotta D, Torti F (2009) Detecting price outliers in European trade data with the forward search. In: Data analysis and classification: from exploration to confirmation, studies in classification, data analysis, and knowledge organization. Springer, Berlin (Forecoming)

Riani M, Atkinson AC (2007) Fast calibrations of the forward search for testing multiple outliers in regression. Adv Data Anal Classif 1:123–141. doi:10.1007/s11634-007-0007-y

Riani M, Atkinson AC, Cerioli A (2009) Finding an unknown number of multivariate outliers. J Royal Stat Soc Ser B 71:201–221

Riani M, Cerioli A, Atkinson A, Perrotta D, Torti F (2008) Fitting mixtures of regression lines with the forward search. In: Fogelman-Soulie F, Perrotta D, Piskorski J, Steinberger R (eds) Mining massive data sets for security. IOS Press, Amsterdam, pp 271–286

Rousseeuw PJ (1984) Least median of squares regression. J Am Stat Assoc 79:871–880

Spence R (2001) Information visualization. Addison Wesley, California

Tufte ER (1983) The visual display of quantitative information. Graphics Press, Cheshire

Wilhelm A (2008) Linked views for visual exploration, vol XIV. Chen, Härdle, and Unwin, pp 199–215