

Dynamic visualization of outliers and mixtures in trade data through robust methods ¹

Domenico Perrotta

European Commission, Joint Research Centre, Institute for the Protection and Security of the Citizens, Ispra, Italy, e-mail: domenico.perrotta@ec.europa.eu

Francesca Torti

Dipartimento di Statistica, Università di Milano Bicocca and Dipartimento di Economia, Università di Parma, Italy, e-mail: f.torti1@campus.unimib.it

Abstract: This contribution is about the analysis of international trade data through a robust approach for the identification of outliers and regression mixtures called Forward Search. The focus is on interactive tools that we have developed to dynamically connect the information which comes from different robust plots and from the trade flows in the input datasets. The work originated from the need to provide the statistician with new robust exploratory data analysis tools. However the focus here is on the interpretation of the results by the end-user. We argue that with the proposed 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.

Keywords: Forward Search, outlier detection, fraud detection, statistical graphics

1. Introduction

The Forward Search (FS) is a general method for detecting unidentified subsets and masked outliers and for determining their effect on models fitted to the data. The key concepts of the method were given by Hadi (1992), while Atkinson et al (2000) and Atkinson et al (2004) introduced the idea of diagnostic monitoring and applied the method to a wide range of regression and multivariate statistical techniques. Riani et al (2008) also proposed a natural extension of the FS to the estimation of regression mixtures.

Unlike other robust approaches, the FS is a dynamic process that produces a sequence of estimates. It starts from a small, robustly chosen, subset of the data and fit subsets of increasing size, in such a way that outliers and subsets of data not following the general structure are revealed by diagnostic monitoring. In regression the FS typically monitors the evolution of residuals, parameter estimates and inferences as the subset size increases. The results are then presented as ‘forward plots’ that show the evolution of these quantities of interest as a function of sample size.

More precisely, for one univariate response and v explanatory variables, if $\hat{\beta}(m)$ is the estimate of the $(v + 1)$ -dimensional parameter vector $\beta = (\beta_0, \beta_1, \dots, \beta_v)^T$ obtained by fitting the regression hyperplane to subset $S(m)$, then the new subset $S(m + 1)$ is build with the observations with the first $m + 1$ smallest *squared regression residuals*. Then, to

¹ Authorship is given in alphabetic order. The work is 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”.

detect outliers we examine the so called *minimum deletion residual* amongst observations not in the subset. Thus, the main two forward plots in regression show the progression of the minimum deletion residual and of a scaled version of the squared regression residuals. An aspect of the FS that has never been treated is the lack of an automatic link among the great variety of forward plots which are monitored during the process. Linking of plots is especially relevant in concrete applied problems, when it is crucial to present and communicate effectively the statistical results to end-users not necessarily familiar with the statistical methods. This takes place in the Joint Research Centre (JRC) of the European Commission when the authors and other statisticians cooperate with the European Anti-Fraud Office (OLAF) to highlight in European trade data potential cases of fraud, data quality issues and other oddities related to specific international trade contexts.

In this short paper we briefly describe the interactive tools that we are developing with one of the authors of the FS (Marco Riani) to dynamically connect the information which comes from different robust FS plots. The proposed brushing and linking tools are extended to data tables that include the full set of variables present in the dataset analysed. We show with one example that this conceptually simple extension may represent a valid support to the interpretation of the results by the subject matter experts. Given the limited space for this short paper, we postpone the mathematical details of the FS to a later extended version and we give here few specific references to the literature.

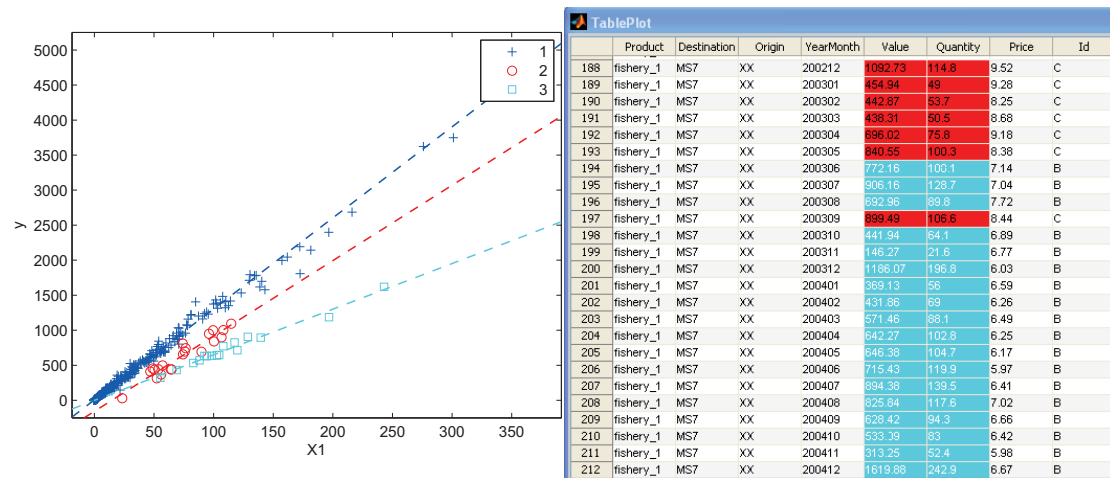
2. Dynamic FS plots applied to international trade data

The standard forward plot of the minimum deletion residual is a static graph. It is necessary to write ad hoc programs to identify the steps where selected units entered the subset or to appreciate the joint effect of such units on the regression fit or on other statistics monitored in different plots. This cumbersome off-line way of proceeding has drastically limited the possibility to propose the FS method and its results to a non statistical audience and may have even limited the use of the method in the statistical community.

We illustrate here the advantages of adopting more flexible interactive plots. We use an example taken from the thousands of international trade datasets that we analyse at the JRC. 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 traders operating at different prices, which results in mixtures of linear populations. Abnormal price transactions are made available to OLAF and its anti-fraud partners in the Member States for analyses motivated by the identification of potential fraud cases.

Figures 1 and 2 are the main output of the FS applied to such data. The scatterplot represents the quantity (x axis) and the value (y axis) of the monthly flows of a fishery product imported into the EU from a third country. The red circles and the cyan squares in the lower part of the scatterplot are imports into a specific EU Member State (say MS7) and the prices declared look suspiciously low if compared to those of the other Member States (MSs). Using the FS, Perrotta et al (2009) could identify quite precisely the group of the “fair” import flows to the other MSs, in the scatterplot with blue ‘+’, estimate their price (13 €/Kg) and finally divide the group of abnormal flows in other two linear clusters of significantly different estimated price, 9.17 (red circles) vs 6.55 (cyan squares) €/Kg. However there is a relevant fact that becomes clear only by inspecting in the input table

Figure 1: Fishery trade data. The scatterplot of values and quantities and the data table with other relevant variables. Selected units in the scatterplot are automatically highlighted in the table and vice-versa.



the transactions corresponding to the circles and squares, by comparing their prices and the dates when the transactions were recorded. In fact the two abnormal clusters include flows that took place respectively in the first 14 consecutive months of the period analysed (circles) and in the following consecutive 21 months (squares). A natural conclusion of clear operational value is that in the period analysed the traders in MS7 gradually under-declared their import price, up to half of the price reported by the other MSs, so that to reduce import duty payments.

In Perrotta et al (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 realise that near the origin of the axes there is a large number of linear groups of data, visible only by selecting that region and zooming on it.

Our solution for the end user consists in linking 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. Linking is so that the points selected in the scatterplot or the cells/row selected in the table are both and simultaneously highlighted with common colors. The selection of relevant points/cells can be done manually by the analyst at any time, but a first selection is done automatically by the procedure based on the outcome of the FS, e.g. on the set of outliers or on the group of the estimated mixture.

Linking is extended automatically to any other plot, such as those of figure 2 that are of great interest for the statisticians: the FS trajectories of the residuals and the FS minimum deletion residual curve. The statistician would easily find which observations correspond to those residuals which keep constantly high absolute values throughout the search (flows labelled with n. 200 and 212), or those entering the subset only in the last steps of the procedure (cyan and red part of the minimum deletion residual curve).

