Mining Massive Data Sets for Security
F. Fogelman-Soulié et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-898-4-271

Fitting Mixtures of Regression Lines with the Forward Search

Marco RIANI [a,1], Andrea CERIOLI [a], Anthony C. ATKINSON [b], Domenico PERROTTA [c] and Francesca TORTI [c]

[a] *Department of Economics, University of Parma, Italy*
[b] *Department of Statistics, London School of Economics, UK*
[c] *European Commission, Joint Research Centre, Institute for the Protection and Security of the Citizens, Support to External Security Unit, Ispra, Italy*

Abstract. The forward search is a powerful method for detecting unidentified subsets and masked outliers and for determining their effect on models fitted to the data. This paper describes a semi-automatic approach to outlier detection and clustering through the forward search. Its main contribution is the development of a novel technique for the identification of clusters of points coming from different regression models. The method was motivated by fraud detection in foreign trade data as reported by the Member States of the European Union. We also address the challenging issue of selecting the number of groups. The performance of the algorithm is shown through an application to a specific bivariate trade data set.

Keywords. Clustering, Outlier detection, Regression

Introduction

The Forward Search (FS) is a powerful general method for detecting outliers, unidentified subsets of the data and inadequate models and for determining their effect on models fitted to the data. The basic ideas started with the work of [11] and [1]. The power of the FS was considerably increased by [2] and [5] through the idea of diagnostic monitoring. They extended its applicability to a wide range of multivariate statistical techniques. Unlike most robust methods that fit to subsets of the data (see, e.g., [14] and [12]), in the FS the amount of data trimming is not fixed in advance, but is chosen conditionally on the data. Many subsets of the data of increasing size are fitted in sequence and a whole series of subsets is explored. As the subset size increases, the method of fitting moves from very robust to highly efficient likelihood methods. The FS thus provides a data dependent compromise between robustness and statistical efficiency.

In this contribution we describe a novel technique in which the FS is applied to detect clusters of observations following different regression models. Our assumptions are comparable to those underpinning latent class and model-based clustering methods [10], but our output is richer. The rationale is that if there is only one population the journey from fitting a few observations to all will be uneventful. But if we have two

[1] Corresponding Author: mriani@unipr.it or, for the anti-fraud application, domenico.perrotta@ec.europa.eu

or more groups there will be a point where the stable progression of fits is interrupted. Our tools for outlier detection and clustering are then developed from forward plots of residuals and distances computed from searches with either robust or random starting points. We also address a number of challenging issues, including selection of the number of groups and use of distributional results for precise identification of the outliers and the clusters.

Our focus is on clustering regression models. However, the ideas of clustering multivariate data are both more familiar and more easily explained, so we start in §1 with a brief review of the FS methodology for multivariate data. Two didactic examples of cluster detection for multivariate data are in §2. The new algorithm for detecting clusters of regression lines is introduced in §3. In §4 we show this algorithm in action for the purpose of detecting fraudulent transactions in trade data sets selected by the anti-fraud office of the European Commission and its partners in the Member States. The paper concludes in §5 with some remarks and suggestions for further development.

1. The Forward Search for Multivariate Observations

Outliers are observations that do not agree with the model that we are fitting to the data. Single outliers are readily detected, for example in regression by plots of residuals. However, if there are several outliers they may so affect the parameter estimates in the fitted model that they are not readily detected and are said to be "masked". Such multiple outliers may indicate that an incorrect model is being fitted.

The basic idea of the Forward Search is to start from a small subset of the data, chosen robustly to exclude outliers, and to fit subsets of increasing size, in such a way that outliers and subsets of data not following the general structure are clearly revealed by diagnostic monitoring. With multiple groups, searches from more than one starting point are often needed to reveal the clustering structure. In this section we restrict attention to data sets of multivariate continuous observations, for which outlyingness is measured through their Mahalanobis distances. The case of multivariate categorical data is addressed by [8].

The squared Mahalanobis distances for a sample $S(n) = \{y_1, \ldots, y_n\}$ of n v-dimensional observations are defined as

$$d_i^2 = \{y_i - \hat{\mu}\}^T \hat{\Sigma}^{-1} \{y_i - \hat{\mu}\}, \qquad i = 1, \ldots, n, \tag{1}$$

with $\hat{\mu} = \bar{y}$ the vector of sample means and

$$\hat{\Sigma} = \sum_{i=i}^{n} (y_i - \hat{\mu})(y_i - \hat{\mu})^T / (n - v)$$

the unbiased moment estimators of the mean and covariance matrix of the n observations. Throughout ^{T} denotes transpose.

In the FS the parameters μ and Σ are estimated from a subset $S(m) \subseteq S(n)$ of m observations, yielding estimates $\hat{\mu}(m)$ and $\hat{\Sigma}(m)$. From this subset we obtain n squared Mahalanobis distances

$$d_i^2(m) = \{y_i - \hat{\mu}(m)\}^T \hat{\Sigma}^{-1}(m) \{y_i - \hat{\mu}(m)\}, \qquad i = 1, \ldots, n. \tag{2}$$

To start the search when the observations are assumed to come from a single multivariate normal population with some outliers, [5] pick a starting subset $S(m_0)$ that excludes any two-dimensional outliers. One search is run from this unique starting point. When a subset $S(m)$ of m observations is used in fitting, we order the squared distances and take the observations corresponding to the $m+1$ smallest as the new subset $S(m+1)$.

To detect outliers we examine the minimum Mahalanobis distance amongst observations not in the subset

$$d_{\min}(m) = \min d_i(m) \quad \text{for } i \notin S(m). \tag{3}$$

If this observation is an outlier relative to the other m observations, its distance will be "large" compared to the maximum Mahalanobis distance of observations in the subset. All other observations not in the subset will, by definition, have distances greater than $d_{\min}(m)$ and will therefore also be outliers.

In order to provide sensitive inferences it is necessary to augment the graph of $d_{\min}(m)$ with envelopes of its distribution. Detailed examples of such envelopes and of their use in the FS with moderate sized data sets are presented by [6] and [4]. A synthesis is provided in the next section.

For small data sets we can use envelopes from bootstrap simulations to determine the thresholds of our statistic during the search. These are found by performing the FS on many sets of data simulated from a standard multivariate normal distribution with the same value of n and v as our sample. Note that in the simulations we can use data generated from the standard normal distribution because Mahalanobis distances are invariant to linear transformation of the data. In the example here we make 10,000 such simulations. For each FS we monitor the values of $d_{\min}(m)$ defined in (3). As a consequence, for each value of m we have the empirical distribution of $d_{\min}(m)$ under the hypothesis of normality. The envelopes we use are the quantiles of this distribution. For example, the 99% envelope is that value which is the 1% point of the empirical distribution. With 10,000 simulations, this is the 100th largest value, so that 99 of the simulated values are greater than it. We calculate these quantiles for each value of m that is of interest.

For larger data sets we can instead use polynomial approximations. Theoretical arguments not involving simulation are provided by [13], together with a formal test of multivariate outlyingness and comparisons with alternative procedures.

For cluster definition, as opposed to outlier identification, several searches are needed, the most informative being those that start in individual clusters and continue to add observations from the cluster until all observations in that cluster have been used in estimation. There is then a clear change in the Mahalanobis distances as units from other clusters enter the subset used for estimation. This strategy seemingly requires that we know the clusters, at least approximately, before running the searches. But we instead use many searches with random starting points to provide information on cluster existence and definition.

In §4 we use envelopes to determine cluster membership. Since the size of the clusters has to be established, we need envelopes for several different values of n. Simulation then becomes time consuming unless n is very small. Calculation of the envelopes via the theoretical arguments in [13] become increasingly attractive as n increases.

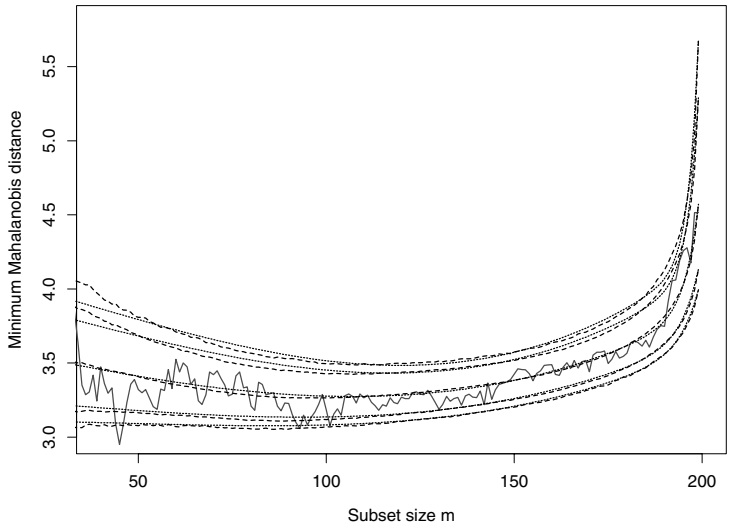

Figure 1. Swiss Heads data (n = 200): forward plot of minimum Mahalanobis distance with 1, 5, 50, 95 and 99% points: continuous lines, 10,000 simulations; dashed lines, interpolation. Simulations and approximation agree well. There is no evidence of any outliers.

2. Didactic Examples

The purpose of this paper is to provide methods for the relatively large and structured data sets that arise in the fraud detection problems illustrated in §4. However, we first look at brief analyses of two smaller examples, as a training of the eye in the interpretation of our forward plots for the detection of clusters.

2.1. Swiss Heads Data

The Swiss Heads data set was introduced by [9, p. 218]. It contains information on six variables describing the dimensions of the heads of 200 twenty year old Swiss soldiers. These data were extensively analysed, using the forward search, by [5]. The conclusion is that the observations come from a six dimensional normal distribution, from which there are no outliers. The plot of Figure 1 confirms this opinion. The envelopes in this figure were found both directly by simulation from the multivariate normal distribution and by parametric interpolation. The distances lie inside the envelopes, indicating complete agreement with the multivariate normal distribution.

2.2. Example With Three Clusters

We now look at a synthetic example with three clusters of four-dimensional correlated observations, to show how random start forward searches combined with envelope plots of forward Mahalanobis distances lead to the indication of clusters. There are 1,000 units in all, 250 of which are in the first group, 300 in the second and 450 in the third. The observations in each group are highly correlated and the third group lies between the other two, so that there is considerable overlapping. Of course, in our analysis we ignore the information about group structure, or even about the number of groups.

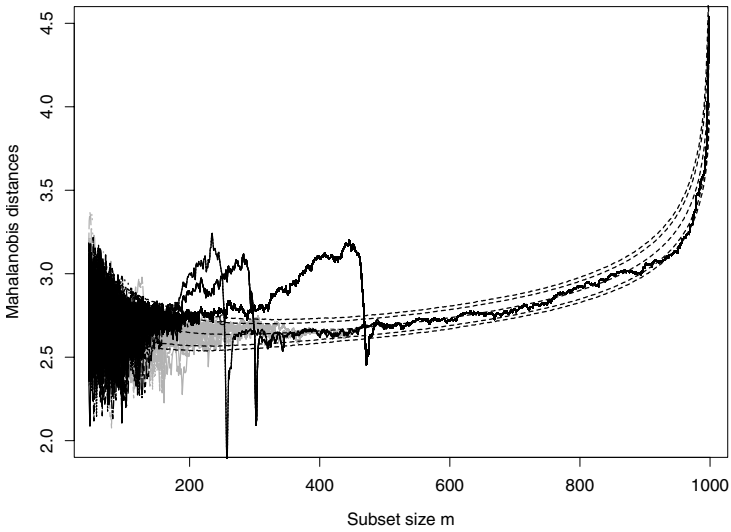

Figure 2. Three clusters of correlated normal variables: forward plot of minimum Mahalanobis distances from 200 random starts with 1%, 5%, 50%, 95% and 99% envelopes. Three clusters are evident.

We run 200 random start forward searches, each one starting with $m_0 = v + 1$, the smallest possible size and that which gives the highest probability of getting a subset consisting solely of observations from one cluster. The resulting forward plot of minimum distances is in Figure 2. The forward searches in this plot fall into four classes: those that start in each of the three groups and those that, from the beginning of the search, include observations from at least two groups. From around $m = 150$ the searches with observations from only one group start to lie outside the envelopes. These curves reach a peak and then suddenly dip below the envelopes as relatively remote observations from the other groups enter the subset used in fitting. From a little after $m = 500$ there is a single forward plot, in which a common mean and common covariance matrix are calculated from observations in more than one group, so that the group structure is no more apparent.

The approximate values of m at the three peaks are: 230, 290 and 450. Despite the overlapping nature of the groups, our method has initially indicated clusters for 97% of the observations. For precise definition of the clusters, we interrogate the subsets $S(m)$ for those trajectories where there is evidence of a cluster structure. The membership of each of the three subsets giving rise to the peaks in Figure 2 can be illustrated using the 'entry' plot of [5]. Cluster 1 includes most of the units of Group 1 and no other units. Cluster 2 contains the majority of the units in Group 2 and some borderline units from Group 3. The intermediate Group 3 is the most misclassified, as is to be expected.

One way to confirm this tentative identification is to run searches on individual clusters. If the peak for a particular cluster in the forward plot analogous to Figure 2 occurs when $m = n_c$, we include the next few units to enter and then run a search on these n_c^+ units, superimposing envelopes for a sample of size n_c^+ as we did in §2.1 for a single population. If no outliers are found, we have a homogenous cluster and increment n_c^+ to check whether we have failed to include some units that also belong to the cluster. If outliers are detected, we delete the last observation to enter, reduce the sample size by

one and superimpose envelopes for this reduced sample size. Eventually we obtain the largest group of homogenous observations containing no outliers. Examples of this procedure are given by [4], together with comparisons with other clustering methods such as k-means which completely fails.

3. Mixtures of Regression Hyperplanes

3.1. Regression Diagnostics

The cluster analysis of multivariate data is well established as is the use of the FS to determine the clusters, especially for data that are multivariate normal. However, the regression framework is different from that of multivariate analysis and there is comparatively little work on clustering regression models. Here, our interest is in clustering by regression hyperplanes. The Forward Search is easily adapted to this regression problem, keeping the same philosophy but with regression-specific ingredients. In particular, distances are replaced by regression residuals.

We now have one univariate response Y and v explanatory variables $X_1, \ldots, X_v$ satisfying

$$E(y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_v x_{iv} \tag{4}$$

with the usual assumptions of independent, additive errors of constant variance (see, for example, [15]).

Let b be a $(v+1)$ vector of constants. For any b we can define n residuals

$$e_i(b) = y_i - (b_0 + b_1 x_{i1} + \cdots + b_v x_{iv}). \tag{5}$$

The least squares estimate $\hat{\beta}$ is the value of b in (5) that minimizes the sum of squares

$$R(n, b) = \sum_{i=1}^{n} e_i^2(b). \tag{6}$$

Likewise, the estimate $\hat{\beta}(m)$, obtained by fitting the regression hyperplane to the subset $S(m)$, minimizes the sum of squares $R(m, b)$ for the m observations $\in S(m)$. From this estimate we compute n squared regression residuals

$$e_i^2(m) = [y_i - \{\hat{\beta}_0(m) + \hat{\beta}_1(m) x_{i1} + \cdots + \hat{\beta}_v(m) x_{iv}\}]^2 \qquad i = 1, \cdots, n$$

the $m+1$ smallest of which are used to define the new subset $S(m+1)$.

The search starts from an outlier-free subset of $m_0 = v+1$ observations found using the least median of squares criterion of [14]. We randomly take an appreciable number of samples of size m_0, perhaps 1,000, estimate b in (5) for the observations in each sample and take as $S(m_0)$ that sample for which

$$M(m_0, b) = \text{median}_{i \in S(M_0)} e_i^2(b) \tag{7}$$

is a minimum. In practice, [14] recommend a slight adjustment to allow for the estimation of β.

To detect outliers in §2 we used the minimum Mahalanobis distance among observations not in $S(m)$. We now instead examine the minimum deletion residual amongst observations not in the subset, which is the t test for detection of individual outliers.

Let X be the $n \times (v+1)$ matrix of the explanatory variables $X_1, \ldots X_v$ with the addition of a column of ones for the constant term in (4). The ith row of X is x_i. Then the minimum deletion residual is defined as

$$r_{\min}(m) = \min \frac{|e_i(m)|}{s(m)\sqrt{[1 + x_i^T\{X^T(m)X(m)\}^{-1}x_i]}} \quad \text{for } i \notin S(m), \tag{8}$$

where $s(m)$ is the square root of $s^2(m) = R\{m, \hat{\beta}(m)\}/\{m-(v+1)\}$, the mean square estimator of the residual variance $\sigma^2 = E\{y_i - E(y_i)\}^2$ and $X(m)$ is the block of X with rows indexed by the units in $S(m)$. The quantity $h_i = x_i^T\{X^T(m)X(m)\}^{-1}x_i$ is the "leverage" of observation i. Observations with large values of h_i are remote in the space of the explanatory variables and can, as we shall see in Figures 4 and 5, cause perturbations in forward plots when they join $S(m)$.

The FS for regression is given book-length treatment by [2]. Inferences about the existence of outliers require envelopes of the distribution of $r_{\min}(m)$, similar to those plotted in §2. Such envelopes are described by [3].

3.2. A Forward Algorithm for Clustering Observations Along Regression Hyperplanes

We now suppose that the observations come from g regressions models (2) with different and unknown parameter values. Our aim is to allocate each unit to its true model and to estimate the corresponding parameters. Also the number g of component models is not known in advance.

Clusterwise regression is the traditional technique for achieving this goal [16]. A more modern probabilistic approach is to fit the joint density of the n observations as a mixture of regressions models [7, §14.5]. However, both methods may suffer from the presence of outliers and/or strongly overlapping clusters. Another shortcoming of these methods is that they do not provide formal tests to justify the need of an additional component. Our proposal is to use the Forward Search for determining and fitting the g components of the regression mixture.

Our forward algorithm combines the strategies outlined in Sections 1 and 3.1. It consists of the following steps:

1. Let $n^*(j)$ be the size of the sample to be analysed at iteration j. At the first iteration $n^*(1) = n$;
2. The FS for regression is applied to these $n^*(j)$ observations. The search is initialised robustly through the least median of squares criterion applied to all n observations and progresses using the squared regression residuals $e_i^2(m), i = 1, \ldots, n^*(j)$;
3. At each step m of the FS, we test the null hypothesis that there are no outliers among the $n^*(j)$ observations. The test is performed using the minimum deletion residual (8);
4. If the sequence of tests performed in Step 3 does not lead to the identification of any outliers, the sample of $n^*(j)$ observations is declared to be homogeneous and the algorithm stops by fitting the regression model (4) to this sample. Otherwise go to Step 5;

5. Let m^* be the step of the FS in which the null hypothesis of no outliers is rejected by the sequence of tests of step 3. Then the observations in $S(m^*)$ identify one mixture component, i.e. one cluster of m^* observations following (4). Fit the regression model (4) to this cluster;
6. Remove the cluster identified in step 5. Return to Step 1 with a reduced sample size, by setting $n^*(j+1) = n^*(j) - m^*$.

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

4. Application to European Union Trade Data and Anti-fraud

In this Section we show how the FS can be used with European Union (EU) foreign trade data as reported by the EU Member States (MS) to detect anomalies of various kinds (e.g. recording errors), specific market price dynamics (e.g. discounts in trading big quantities of product) and cases of unfair competition or fraud. In particular fraud may be associated with anomalously low reported prices that could result in underpayment of taxes.

We use one concrete example to introduce the application context, the data and the statistical patterns of interest, i.e. *outliers* and *mixtures of linear models*. The European Commission's Joint Research Centre detects these patterns in data sets including millions of trade flows grouped in a large number of small to moderate size samples. The statistically relevant cases are presented for evaluation and feed-back to subject matter experts of the anti-fraud office of the European Commission and its partner services in the Member States.

We use the example to illustrate the regression approach introduced in Sections 3. The method of multivariate clustering of 1 is not informative when the data have a regression structure, so we do not consider it any further for this example.

4.1. European Union Trade Data

The data in Figure 3 refer to the quantity (x axis) and the value (y axis) of the monthly import flows of a fishery product into the European Union from a certain third country. The solid dots are the flows to a Member State that we call MS7, and the black circles are the flows to other Member States. We can clearly see that the two groups of observations are distinct and we could fit two linear regression models using the observations in the two groups. We use the slope of these linear models as an estimate of the import price of the flows in the respective groups.

There is also an observation, the open circle on the bottom-left, that does not follow the regression line fitted to the observations of the same category. Rather, it appears in the distribution of the solid dots. This "abnormal" black circle is a single flow to a Member

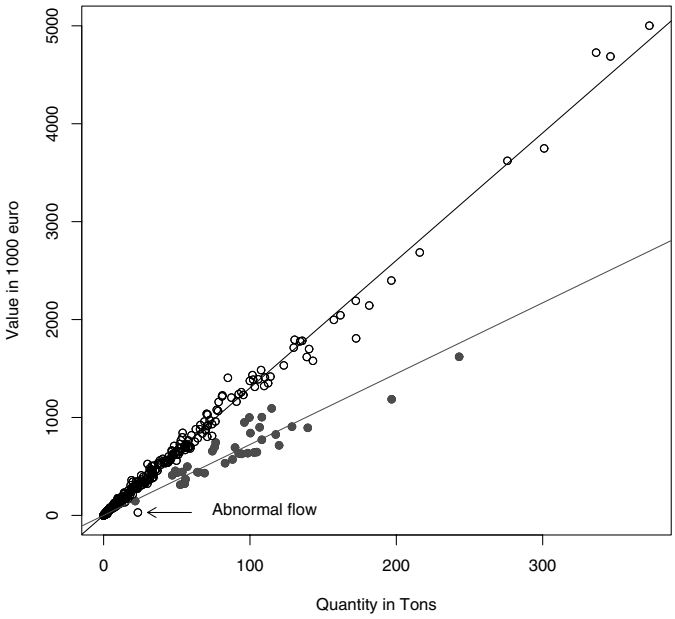

Figure 3. Quantities (in tons) and values (in thousands of euros) of 677 monthly imports of a fishery product from a third country into the EU, over a period of three years. Flows to MS7 (solid dots) and flows to the other Member States (open circles) form distinct groups following different regression lines. On the bottom-left an abnormal single flow to MS11.

State that we identify as MS11. The *unit value* of this flow, obtained by dividing the value by the corresponding quantity, is so small (~ 1.27€/Kg) compared to the market price of this specific fishery product (12.5€/Kg in 2005[2]) that we may suspect an error in recording the data. Although from a data quality point of view it might be worth investigating the validity of this data record, from the economical point of view we are unlikely to be interested in a trade flow of such volume (~ 20 Tons).

Much more importantly, the distribution of the solid dots indicates that the imports of MS7 are systematically underpriced in comparison with the imports of the other Member States. This indication is of appreciable economic relevance since MS7 imported about 20% (~ 3300 Tons) of the total EU imports of this product in our reference period.

Our data sets consist of thousands of samples similar to this example. Therefore we need to detect the outliers and to estimate the mixtures of linear components automatically and *efficiently*. We require high computational and statistical efficiency of the algorithms; they should detect a manageable number of outliers in reasonable time and with reasonable statistical power. But, for the anti-fraud subject matter experts, the concept of efficiency is also related to the problem of extracting cases of possible operational interest from the statistically relevant patterns that we detect with our algorithms. We will not address this issue here.

[2]Source: "European Fish Price Report", a GLOBEFISH publication (http://www.globefish.org).

4.2. Fitting Mixtures of Regression Lines

Our data have a clear regression structure (Figure 3) and we propose now to analyse them with the regression approach of Section 3.1. This approach uses the squared regression residuals for progression in the search and the minimum deletion residual among the observations not in the subset to monitor the search and infer departures from linearity. We show that the iterative application of this approach, detailed in Section 3.2, suggests modelling the data with a mixture of at least four linear components of rather clear interpretability by subject-matter experts.

The initial subset is chosen by robust fitting, using the least median of squares (LMS) criterion. The first iteration leads relatively straightforwardly to the identification of a first linear regression component with 344 homogeneous observations. In accordance with our algorithm, we remove these observations from the data and repeat the procedure on the remaining 333 observations. We describe the second iteration in greater detail.

Figure 4 shows the forward plot of minimum deletion residuals for these 333 observations. Clearly they are not homogeneous, but the question for this iteration is where is the end of the major group? There is a sharp dip in the plot at around $m = 120$ caused by the inclusion in $S(m)$ of a unit with high leverage. Otherwise, we first obtain a signal indicative of model failure at $m = 225$, the first point at which the calculated minimum deletion residual lies above the 99.9% envelope. However, as the plotted envelopes show, we can expect larger values of the residual as m approaches n even when there are no outliers; the value of n for this group may be somewhat larger than 225. Indeed this does seem to be the case.

The upper left-hand panel of Figure 5 shows the envelopes for $n = 225$, together with the deletion residuals up to this sample size. With these new, more curved, envelopes it is clear that the group is homogeneous up to this size. The same is true for the upper right-hand panel for $n = 235$. However, in the lower left-hand panel with $n = 245$, the observed values have already crossed the 99% envelope. For $n = 248$ the 99.9% envelope is crossed, so there is evidence of non-homogeneity. When n is one less, namely 247, there is no exceedance and we take the second component as containing 247 observations.

There are two points about this process. The minor one is that the envelopes are recalculated for each panel but the calculated values of the minimum deletion residuals are from the search plotted in Figure 4; as n increases the plots reveal more of this sequence of values. The major point is that the envelopes we have found have the stated probability, but for each m. Thus the probability of exceeding the 99.9% envelope for any m is 0.1%. However, the probability that the 99.9% envelope is exceeded at least once during a search is much greater than 0.1%. The calculations for regression are in [3]. Here the envelopes are much more like 99% overall, which is a more reasonable level at which to detect a change in structure. However, in our application, the exact significance level of this part of our analysis is not crucial.

The procedure of identification and deletion continues for another three iterations, leading to additional homogeneous populations of 247, 38 and 22 observations. Figure 6 shows the four components of the mixture estimated with this procedure and the remaining 26 observations (the '+' symbols).

The slopes of the four mixture components, from 1 to 4, are: 14.044, 12.542, 13.134 and 5.83. We recall that these values are estimates of the import price of the flows as-

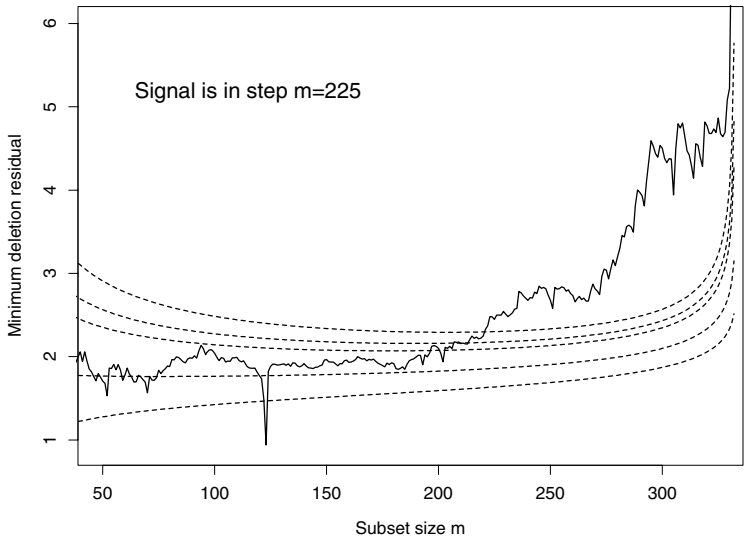

Figure 4. Fisheries data: the 333 observations in iteration 2. Forward plot of minimum deletion residuals with 1, 50, 99, 99.9 and 99.99% envelopes. There is a signal at step 225.

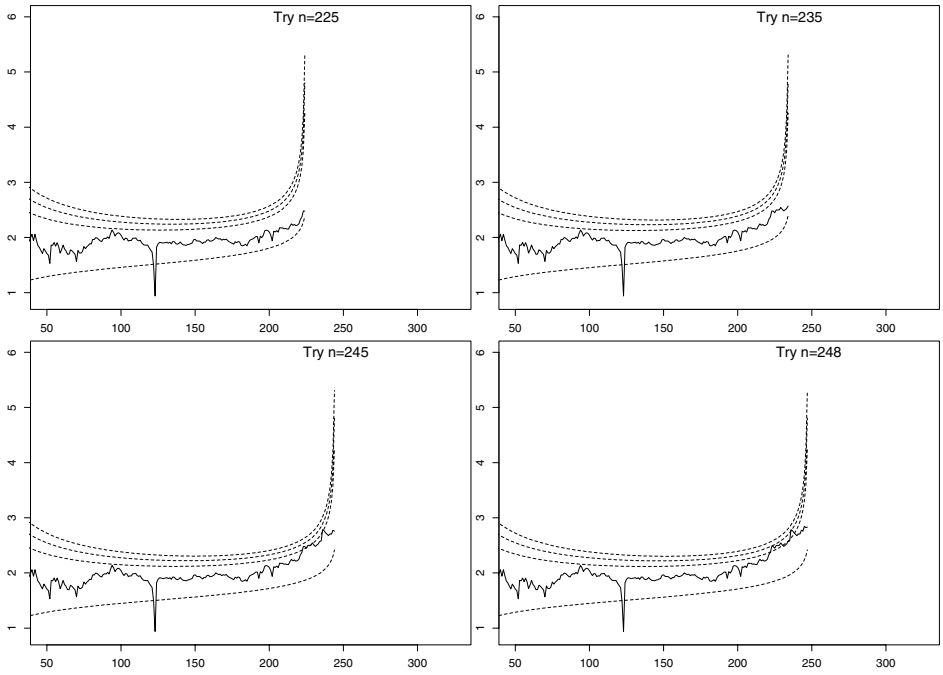

Figure 5. Fisheries data: the 333 observations in iteration 2. Forward plot of minimum deletion residuals with 1, 99, 99.9 and 99.99% envelopes for various sample sizes. Iteration 2 identifies a group of 247 observations.

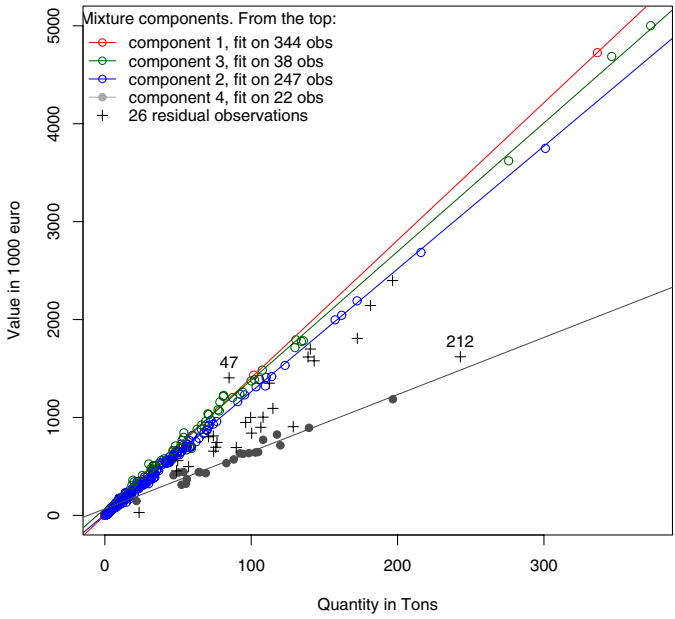

Figure 6. Fisheries data: a mixture of four linear regression lines estimated by iterating the FS on the 677 EU import flows in the dataset. The "residual" flows are identified with a '+' symbol. Those marked with their record number, 212, 47, are detected as outliers by a final FS on the residual flows.

signed by the FS to the four groups. Interestingly, we have verified in our dataset that the group fitted by component 4 (estimated import price 5.83€) is exclusively formed by import flows to MS7 that took place in 22 consecutive months. In addition, there are no flows to MS7 in the other three groups.

Since the prices, that is the slopes of the regression lines, are the major output of our analysis, we give in Figure 7 the forward plots of the estimated slopes during the forward searches for the different iterations. That for the 344 in group 1 is amazingly stable. That for group 2 is also stable, although it does show some slight fluctuations, as well as a small jump around $m = 120$ that we noticed in Figure 3. The much smaller group 3 has a slope between groups 2 and 3. As our other analyses have shown, the slope for group 4 is markedly different. All of these slopes are sensibly constant during the search. However, that for the residual group decreases rapidly, suggesting that these 26 observations are far from homogeneous.

Information on the degree of homogeneity of the observations in the groups can be also obtained from a plot of the estimates of the squared correlation coefficient R^2 in the five iterations (for the general definition of R^2 see, for instance, [15]). This is shown in Figure 8, which also reflects the high strength of the linear regression fit in the four groups.

At this point we accordingly ran the FS also on the 26 "residual" observations. Among those entering the subset in the last steps, two are detected as outliers and cause a visible increase in the plot of deletion residuals. In fact Figure 6 shows that one of these observations, record number 47 in the original dataset, appears almost in line with the first mixture component while the other, 212, is virtually in line with the fourth component.

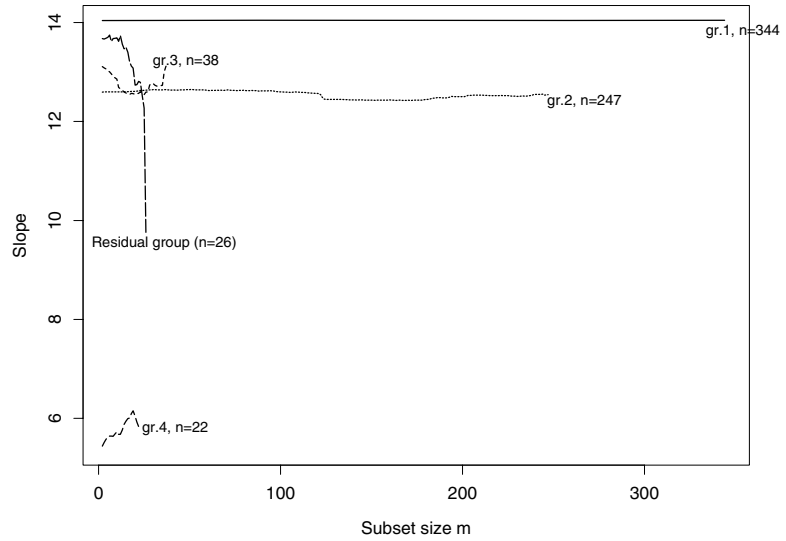

Figure 7. Fisheries data: forward plot of estimated regression coefficients in the five iterations. The first and second groups are very stable with coefficients approximately twice that for the suspicious Group 4.

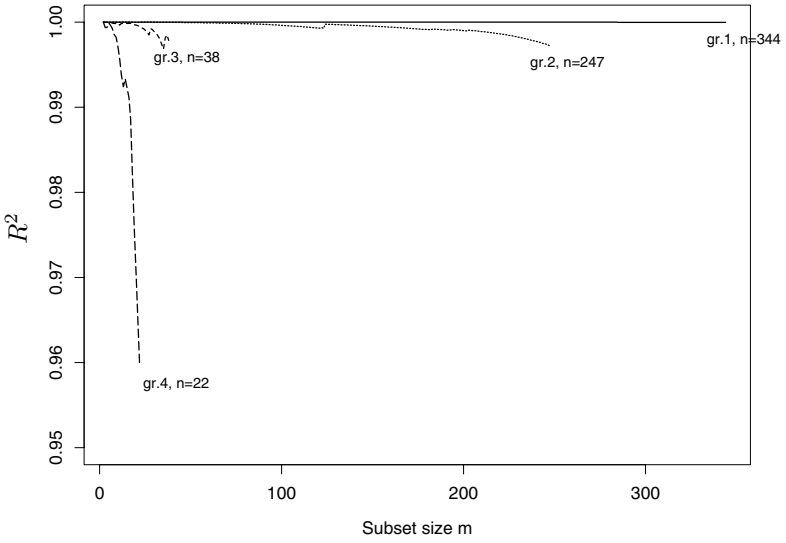

Figure 8. Fisheries data: forward plot of the estimates of the squared correlation coefficient R^2 in the five iterations. The strength of the linear regression fit in the four groups is high.

We now consider in more detail the composition of the residual observations, which are plotted in Figure 9. We represent the flows to MS7 with open circles and the flows to the other Member States with solid dots. The dashed line fits the flows to MS7 but excludes 212 since, as we have just remarked, it is very close to the fourth component. The slope of this line is 7.66. Again this group of 13 flows to MS7 took place in consecutive months. Among the other 11 residual flows, 7 refer to a single Member State, MS2.

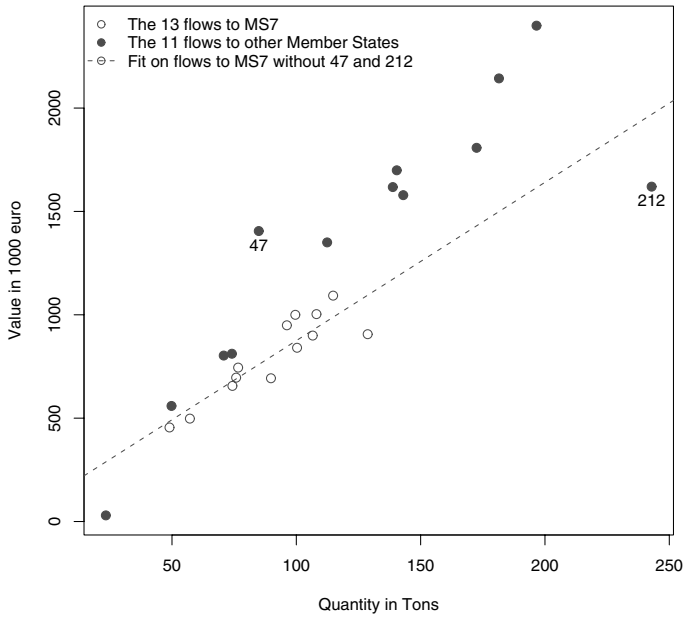

Figure 9. Fisheries data: zoom of the residual flows marked with a '+' in Figure 6. The open circles are the flows to MS7. The solid dots are the flows to other Member States. Again, the two outlying flows are labelled with their record numbers. The regression line is fitted using the flows to MS7 excluding 212, which is outlying.

4.3. Implications for Anti-fraud

The flows to MS7 have been clustered into two homogeneous groups: the first which we called component 4 and the second a fitted subset of the residual observations. Historically, the flows in the first of these two groups took place after those in the second. The estimated import prices for the two periods are considerably different: 5.83€ and 7.66€, and also considerably lower than the prices estimated for the other groups and Member States, 14.05€, 12.54€ and 13.13€. In short, in the period analysed, MS7 lowered the import price of this fishery product, up to half of the import price reported by the other Member States. In earlier analyses this type of pattern was not considered. Its operational evaluation, for example in relation to possible evasion of import duties, is the responsibility of the anti-fraud services.

5. Discussion and Directions for Further Work

The procedure of Section 3.2 indicated four clear sub-populations. A limitation of the procedure is the lack of a criterion to decide automatically about the nature of the residual flows: some of them may form separate homogeneous groups, others are very close to existing groups and could be re-assigned (e.g. flows 47, 212) and yet others may be outliers in the entire dataset (e.g. flow 355). In fact, we have used a rather pragmatic approach to the analysis of the residual observations. A confirmatory analysis invoking simultaneous searches including all established regression lines (not given here for lack of space) can help to infer the degree to which each unit belongs to each group.

The focus in this example has been on clustering linear regression models with motivation for the FS coming from the clustering of multivariate data. We would like to stress that the FS is of much wider applicability; examples, not all in [2], include applications to multiple and curvilinear regression, to nonlinear and generalized linear models and to the estimation of response transformations in regression and data transformation in multivariate analysis. Given any quantity of interest, such as a parameter estimate or a test of departures from a model, its properties can be studied using the FS. The distributional properties of the quantity can be found, often by simulation. Any significant departure from this distribution may indicate outliers, ignored structure or a systematically inadequate model, depending on the quantity being studied. In our anti-fraud application we require techniques for large numbers of observations. For the very large data sets encountered in "data-mining" we use a FS in which $s > 1$ units enter at each forward step; thus we move directly from the subset $S(m)$ to the subset $S(m + s)$.

Acknowledgements

The work of Cerioli, Riani and Atkinson was partially supported by the grants "Metodi statistici multivariati per la valutazione integrata della qualità dei servizi di pubblica utilità: efficacia-efficienza, rischio del fornitore, soddisfazione degli utenti" and "Metodologie statistiche per l'analisi di impatto e la valutazione della regolamentazione" of Ministero dell'Università e della Ricerca – PRIN 2006.

The work of Perrotta and Torti was conducted in the research action "Statistics and Information Technology for Anti-Fraud and Security" of the Joint Research Centre of the European Commission, under the institutional work-programme 2007-2013.

References

[1] A.C. Atkinson, Fast very robust methods for the detection of multiple outliers, *Journal of the American Statistical Association* **89** (1994), 1329–1339.

[2] A.C. Atkinson and M. Riani, *Robust Diagnostic Regression Analysis*, Springer, New York, 2000.

[3] A.C. Atkinson and M. Riani, Distribution theory and simulations for tests of outliers in regression, *Journal of Computational and Graphical Statistics* **15** (2006), 460–476.

[4] A.C. Atkinson and M. Riani, Exploratory tools for clustering multivariate data, *Computational Statistics and Data Analysis* (2007).

[5] A.C. Atkinson, M. Riani and A. Cerioli, *Exploring Multivariate Data with the Forward Search*, Springer, New York, 2004.

[6] A.C. Atkinson, M. Riani and A. Cerioli, Random start forward searches with envelopes for detecting clusters in multivariate data, in: S. Zani, A. Cerioli, M. Riani and M. Vichi (Eds.), *Data Analysis, Classification and the Forward Search*, Springer-Verlag, Berlin, 2006.

[7] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York.

[8] A. Cerioli, M. Riani and A.C. Atkinson, Robust classification with categorical variables, in: A. Rizzi and M. Vichi (Eds.), *COMPSTAT 2006: Proceedings in Computational Statistics, Physica-Verlag, Heidelberg*, 2006.

[9] B. Flury and H. Riedwyl, *Multivariate Statistics: A Practical Approach*, Chapman and Hall, London, 1988.

[10] C. Fraley and A.E. Raftery, Enhanced Model-Based Clustering, Density Estimation, and Discriminant Analysis Software: MCLUST, *Journal of Classification* **20** (2003), 263–286.

[11] A.S. Hadi, Identifying multiple outliers in multivariate data, *Journal of the Royal Statistical Society, Series B* **54** (1992), 761–771.

[12] R.A. Maronna, R.D. Martin and V.J. Yohai, *Robust Statistics: Theory and Methods*, Wiley, Chichester, 2006.
[13] M. Riani, A.C. Atkinson and A. Cerioli, Results in Finding an Unknown Number of Multivariate Outliers in Large Data Sets, Research Report 140, London School of Economics, Department of Statistics.
[14] P.J. Rousseeuw and A.M. Leroy, *Robust Regression and Outlier Detection*, Wiley, New York, 2006.
[15] G.A.F. Seber, *Linear Regression Analysis*, Wiley, New York, 1977.
[16] H. Späth, *Cluster Dissection and Analysis*, Ellis Horwood, Chichester, 1985.