Robust methods for complex data (⋆)

Marco Riani[1], Andrea Cerioli[1], Domenico Perrotta[2], Francesca Torti[2]

[1] Dipartimento di Economia, Università di Parma, Italy
e-mail: mriani@unipr.it

[2] European Commission, Joint Research Centre, Institute for the Protection and Security of the Citizens, Support to External Security Unit, Ispra, Italy

Riassunto: Questo lavoro mostra le potenzialità di un approccio robusto fondato sulla Forward Search per l'adattamento di una mistura di modelli di regressione. Il metodo è motivato dai problemi classificatori che sorgono nel tentativo di identificare frodi fiscali che coinvolgono aziende operanti all'interno dell'Unione Europea. La struttura dei dati è resa complessa dai meccanismi economici che governano le transazioni, dalla necessità di operare in modo automatico su migliaia di mercati differenti e dalla presenza di numerosi outliers. Dal punto di vista statistico, l'approccio proposto consente di coniugare tecniche esplorative e stumenti inferenziali per la scelta del numero di componenti della mistura e per l'identificazione delle situazioni anomale.

Keywords: forward search, outliers, mixtures of regressions, fraud detection

1. Introduction

The protection of the budget of the European Community against fraud is a standing obligation for the Commission and the Member States. The successful protection of the financial interests of the Community is of paramount importance. Fraud against revenues and expenditures of the Community has an impact on the successful conduction of practically all Community activities. As a result of an enduring collaboration with the European Anti-fraud Office, the European Commission's Joint Research Centre routinely collects data sets including millions of trade flows grouped in a large number of small to moderate size samples. These data samples are then analyzed with the purpose of detecting anomalies of various kinds (e.g. recording errors), specific market price dynamics (e.g. discounts in trading big quantities of product) and cases of unfair competition or fraud. The statistically relevant cases are presented for evaluation and feed-back to subject matter experts of the Anti-fraud Office and of its partner services in the Member States.

The statistical analysis of such data shows several levels of complexity. One basic issue is that any diagnostic method must be applied sequentially to millions of records and must be reliable on all of them. Reliability should be measured both in terms of sensitivity, i.e. the number of false positives produced by the method, and of specificity, i.e. the ability of detecting truly anomalous transactions. Satisfactory performance on these grounds translates into high rewards, when frauds are correctly detected, and reduced efforts by the anti-fraud staff, if only few non-anomalous transactions are examined. Both features are

(⋆) The work of Riani and Cerioli was partially supported by grants 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.

crucial for the anti-fraud subject matter experts. They ask for the adoption of techniques that can combine high power for outlier detection with consistency when the data are not contaminated. It is also necessary to summarize the outlyingness of each observation through a single measure, typically a *p*-value from a test statistic, to guarantee routine application of the methods. Hundreds of diagnostic tests must then be computed within thousands of samples and a very severe problem of multiplicity occurs.

The structure of the data adds further complexity to the analysis. The typical pattern of trade flows shows aggregation of points around an unknown number of regression lines whose parameters must be estimated. One major problem is that the regression lines are not well separated. Their degree of overlapping depends on the economic features of the markets where the flows originate. In addition, this archetypal mixture of linear models is heavily contaminated by observations that do not follow the general pattern. Outliers may be isolated, e.g. when produced by recording errors, or clustered, when they represent a systematic behaviour. Perhaps the most useful information for fraud detection purposes comes from outliers that cluster themselves around an "anomalous" regression line. Also inliers that do not follow any specific regression structure and are intermediate between different lines may be of interest, especially if they represent economically important transactions. Figure 1 provides an example of such a complex data structure. The goals of the statistical analysis can be summarized as follows:

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

All the three steps must be performed robustly, to ensure that multiple outliers do not mask each other. Our proposal is to perform them through the Forward Search (FS), a powerful general method for detecting unidentified subsets and masked outliers and for determining their effect on models fitted to the data (Atkinson and Riani, 2000; Atkinson *et al.*, 2004). Unlike most robust methods, in the FS the amount of trimming is not fixed in advance but is chosen conditionally on the data. Many subsets of the data 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 §4 we show how well the FS is able to describe the complex structure of the data pictured in Figure 1. Our assumptions are comparable to those underpinning latent class and model-based clustering methods (Bishop, 2006), but our results are not affected by outliers and our output is richer. Furthermore, we apply distributional results for precise identification of the outliers and of the clusters. These distributional results are particularly appropriate for the simultaneous testing scenario implied by sequential screening of millions of records.

2. Essentials of the Forward Search

The basic idea of the Forward Search is to start from a small, robustly chosen, subset of the data 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. 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 outliers or groups there will be a point where the stable progression of fits is interrupted. Our tools for outlier detection and clustering in regression are then developed from forward plots of residuals.

In the general regression framework we have one univariate response Y and v explanatory variables $X_1, \ldots, X_v$ satisfying

$$E(y_i) = \beta_0 + \beta_1 x_{i1} + \cdots \beta_v + x_{iv} \tag{1}$$

under the usual assumptions. Suppose that a sample $S(n)$ of n observations on Y and on the explanatory variables is available. Standard least-squares theory focuses on $\hat{\beta}$, the estimate of the $(v+1)$-dimensional parameter vector $\beta = (\beta_0, \beta_1, \cdots, \beta_v)^T$, where ^{T} denotes transpose, computed by fitting the regression hyperplane to all the observations in $S(n)$. On the contrary, in the FS we are interested in the *sequence of estimators* $\hat{\beta}(m_0), \hat{\beta}(m_0+1), \ldots, \hat{\beta}(n)$, obtained by fitting the regression hyperplane to subsamples $S(m) \subseteq S(n)$ of m observations, with $m = m_0, \ldots, n$. Each subsample $S(m+1)$ in this sequence is obtained by looking at the n squared regression residuals

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

computed from the estimate of β at step m. $S(m+1)$ is defined as the subset of observations corresponding to the $m+1$ smallest squared residuals $e_i^2(m)$. The search starts from an outlier-free subset of m_0 observations. Usually $m_0 = v+1$, with $S(m_0)$ chosen through the least median of squares criterion of Rousseeuw and Leroy (1987).

To detect outliers we examine the minimum absolute deletion residual amongst observations not in the subset

$$r_{\min}(m) = \min \frac{|e_i(m)|}{s(m)\sqrt{[1 + x_i^T \{X^T(m) X(m)\}^{-1} x_i]}} \quad \text{for } i \notin S(m), \tag{3}$$

where $s(m)$ is the square root of the estimate of the residual variance $\sigma^2 = E\{y_i - E(y_i)\}^2$ computed from the observations in $S(m)$, $x_i = (x_{i1}, \ldots, x_{iv})^T$ is the ith row of the design matrix X and $X(m)$ is the block of X with rows indexed by the units in $S(m)$. If the observation for which (3) is computed does not follow the regression model, the modulus of its deletion residual will be large if compared to the maximum amongst observations belonging to the subset. At that step all other observations not in the subset will, by definition, have absolute deletion residuals greater than $r_{\min}(m)$ and will therefore also be outliers. We call the graph of $r_{\min}(m)$ for $m = m_0, \ldots, n$ a *forward plot*.

In order to provide sensitive inferences it is necessary to augment the forward plot of $r_{\min}(m)$ with envelopes of its distribution. Detailed examples of such envelopes and of their use in the FS with moderate sized regression data sets are presented by Atkinson and Riani (2006), while Atkinson and Riani (2007) consider the multivariate framework. For small data sets we can use envelopes from bootstrap simulations to determine the threshold of our statistic during the search. For larger data sets we can instead use polynomial approximations. Theoretical arguments not involving simulation, which are increasingly attractive as n grows, are provided by Riani *et al.* (2007) together with a formal test that allows for simultaneity in outlier detection. Therefore, the procedure of Riani *et al.* (2007) is a particularly suitable inferential framework for detecting multiple outliers in complex data sets such as those described in this work.

3. Mixtures of Regression Hyperplanes

We now suppose that the observations come from g regressions models (1) 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 (Späth, 1985). A more modern probabilistic approach is to fit the joint density of the n observations as a mixture of regressions models (Bishop, 2006, §14.5; Van Aelst *et al.*, 2006). However, both methods may suffer from the presence of outliers and/or strongly overlapping clusters as shown, e.g., by Neykov *et al.* (2007). 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 fitting the g components of the regression mixture. Our forward algorithm is summarized as follows.

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 initialized robustly through the least median of squares criterion 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 in the $n^*(j)$ observations. The test is performed using (3) and the technique developed by Riani *et al.* (2007) to keep simultaneity into account.
4. If the sequence of tests performed in step 3 does not lead to the identification of any outlier, the sample of $n^*(j)$ observations is declared to be homogeneous and the algorithm stops by fitting the regression model (1) 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 (1). Fit the regression model (1) 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 Trade Data

The EU has a common commercial policy which aims at protecting European businesses from obstacles to trade[(1)]. The trade needs to be monitored and cases of unfair competition

[(1)] http://europa.eu/pol/comm

and fraud should be detected as early as possible. This can be achieved by appropriate analysis of trade data. Figure 1 is an example of trade dataset: the quantity (x axis) and the value (y axis) of the importations of a specific technological product into a Member State (MS) of the European Union (EU) are plotted. The commercial activity is between importers in the MS and exporters in non-EU Member Countries.

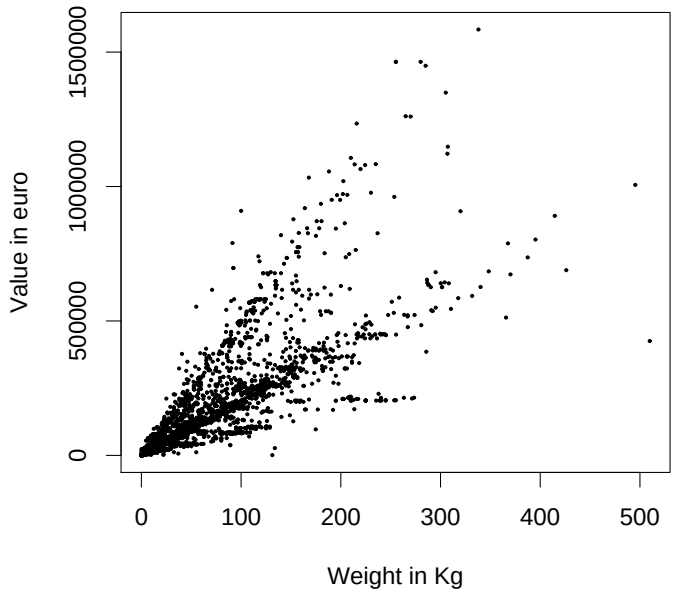

Figure 1: *An example of external trade dataset: quantities (in kilograms) and values (in thousands of euros) of 4719 import transactions into a Member State of the European Union for a technological product, in a period of one year.*

The observations appear roughly distributed along three main different lines departing from the origin of the coordinate axes. One group of observations extends mostly over the upper part of plot, a second group over the central part and a third over the lower part. We will refer to these three informal groups as upper, central and lower groups. The dataset is formed by 4719 observations. Like for the majority of similar trade datasets that we have analysed so far, observations are rather concentrated towards the origin. To appreciate this fact we can compute the Gini coefficients for the quantity and the values (they are respectively 0.63 and 0.67) or more intuitively we can observe that more than 50% of the transactions (precisely 2669) are below 30 Kg weight and cover only 10% of the total trade volume considered in this dataset.

We used the FS to cluster automatically the observations, in order to estimate the import price of the transactions in the respective groups. It is up to subject matter experts to judge the price estimates, eventually examine the transactions in each cluster and draw conclusions. Here, we will only check the homogeneity of the transactions throughout the clusters based on the country of origin of the goods, although the rich structure of this dataset allows different and more sophisticated evaluation criteria. We have conducted similar analyses elsewhere (e.g. Riani *et al.*, 2008) on somewhat less complex data, where flows consisting of monthly aggregations by Member State were available.

The iterative application of the procedure described in Section 3 has produced ten mixture components which are shown in Figure 2. Components 1, 2, 3, 4 and 6 ('$+$' symbols) partially overlap and capture well most of the central group of observations.

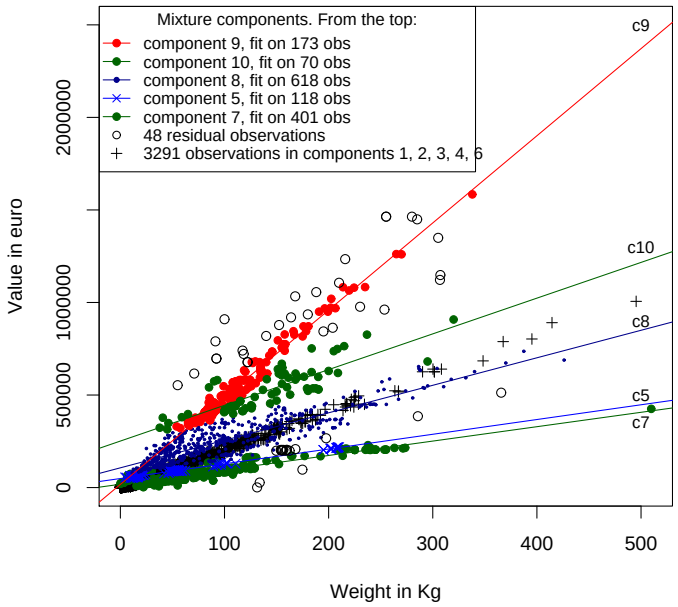

Figure 2: *Components found by the FS. The regression lines of components 1, 2, 3, 4 and 6 almost overlap and are therefore omitted.*

Component 8 ('·' symbols) extends over the central group, but it is much more dispersed. Components 5 and 7 are rather aligned and, together, cover most of the lower group. Finally, components 9 and 10 cover rather well the observations of the upper group. The "residual" observations ('○' symbols) that are not assigned by the FS to any cluster, are very dispersed and could be reassigned to the existing tentative clusters with a confirmation procedure.

component	average price	#obs$(Q > 30)$	#obs	ρ^2	slope	intercept
1	2,036.23	99	1432	1	2032.7	66.6
2	2,122.50	146	570	0.9995	2070	2877.2
3	1,943.40	152	564	0.9995	2010.4	-3217.7
4	2,279.88	269	451	0.9985	2121.1	9372.7
5	1,517.94	73	118	0.995	797	48857.7
6	1,829.87	141	274	0.999	2001.1	-11821.9
7	952.98	317	401	0.853	777.4	17916.1
8	2,987.02	559	618	0.7825	1486.1	106730
9	4,876.89	173	173	0.9745	4708.9	17137.7
10	4,060.60	70	70	0.7125	1930.3	251326.5

Table 1: *Basic statistics on the components.*

Table 1 gives a few basic statistics for the components. The average price (value/weight) is computed excluding the transactions below 30 Kg, because in this context they are operationally irrelevant and, at some extent, subject to approximations due to the Customs data collection procedures. These prices confirm that, for operational purposes, components 1, 2, 3, 4 and 6 could be merged into a single group. As expected, the first six components are highly homogeneous and capture observations

densely concentrated. The corresponding estimate of the squared correlation coefficient ρ^2 is very close to 1. The progression of this estimate during the searches is shown in

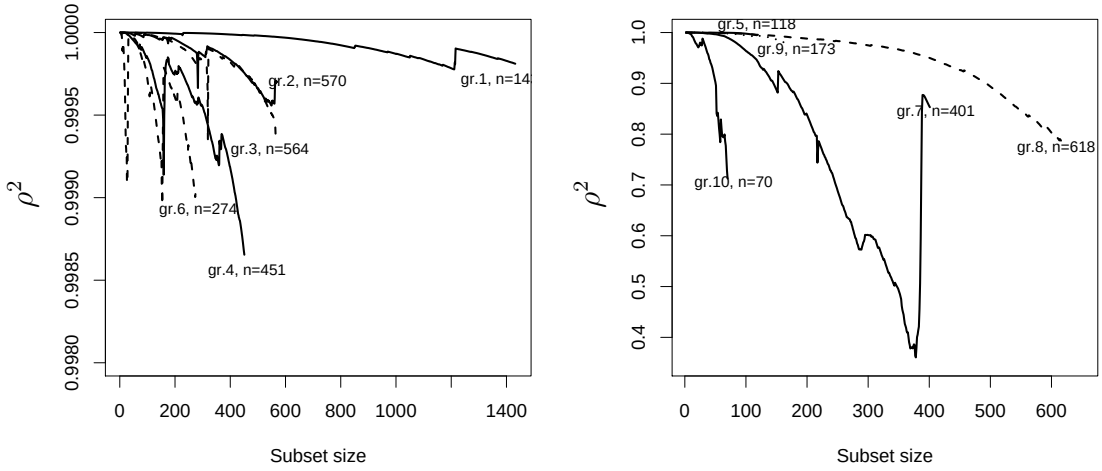

Figure 3: *Forward plots of the estimates of ρ^2 for components 1, 2, 3, 4, and 6 (on the left) and for components 5, 7, 8, 9, 10 (on the right).*

Figure 3. Figure 4 also gives the progression of the slopes for the components of bigger operational relevance, namely those which correspond to the smallest and the highest prices (components 5, 7 and 9).

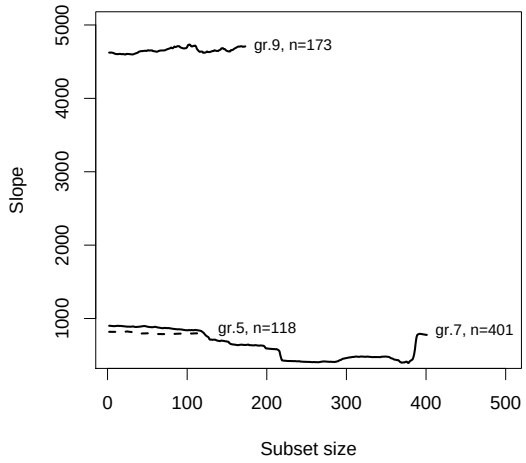

Figure 4: *Forward plots of the estimated slopes for components 5,7 and 9.*

We may get insights from the mixture components, for anti-fraud or for a simple academic exploration of the European trade dynamic for the product in question, by focusing on the country of origin of the goods. The cheapest goods, those in components 5 and 7, originate mainly from a single third country (Malaysia, MY). The most expensive goods (component 9) have origin mainly in Korea (KR). KR is also the dominant exporting country in component 10 (on the other hand the average prices for components 9 and 10 are comparable). The biggest exporting country is China (CN), which mainly appears in the central components (1, 2, 3, 4 and 6). In some cases the FS components also suggest unexpected patterns. For example, by looking at the date of the transactions

in components 5 and 7, we note that gradually the price of imports from MY raised from a lower to a higher value. Such patterns are difficult to detect using alternative techniques. In other applications the identification of clusters of anomalous transactions proved to be an important tool for directing the attention of anti-fraud services to possible instances of unfair competition and fraud, such as evasion of import duties (Riani *et al.*, 2008).

5. Discussion

With complex data sets such as those considered in this paper many additional things can be tried. For instance, we could have worked with log-transformed data in §4, or we could have forced the fitted regression lines to pass through the origin of Figure 1. Due to lack of space in this article we have been unable to report all these variations. Due to similar reasons we did not have space to show the comparison of the suggested procedure with other methods. However, the message of the article is clear: even in presence of highly overlapping groups with multiple masked outliers, the application of the forward search provides the user with a plethora of information about the data under analysis and produces a reasonable tentative classification. Finally, it is worth noting that the suggested approach does not require optimization routines, provides the user with a simple way of determining the number of clusters and does not force all units to be firmly clustered.

References

Atkinson A. C., Riani M. (2000), *Robust Diagnostic Regression Analysis*, Springer, New York.

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

Atkinson A. C., Riani M. (2007), Exploratory tools for clustering multivariate data, *Computational Statistics and Data Analysis*, 52, 272–285.

Atkinson A. C., Riani M., Cerioli A. (2004), *Exploring Multivariate Data with the Forward Search*, Springer, New York.

Bishop C. M. (2006), *Pattern Recognition and Machine Learning*, Springer, New York.

Neykov N., Filzmoser P., Dimova R., Neytchev P. (2007), Robust fitting of mixtures using the trimmed likelihood estimator, *Computational Statistics and Data Analysis*, 52, 299-308.

Riani M., Atkinson A. C., Cerioli A. (2007), Results in Finding an Unknown Number of Multivariate Outliers in Large Data Sets, Research Report 140, London School of Economics, Department of Statistics.

Riani M., Cerioli A., Atkinson A. C., Perrotta D., Torti F. (2008), Fitting Robust Mixtures of Regression Lines to European Trade Data, To appear in *Mining Massive Datasets for Security Applications*, IOS Press, 2008.

Rousseeuw P. J., Leroy A. M. (1987), *Robust Regression and Outlier Detection*, Wiley, New York.

Späth H. (1985), *Cluster Dissection and Analysis*, Ellis Horwood, Chichester.

Van Aelst S., Wang X., Zamar R. H., Zhu R. (2006), Linear grouping using orthogonal regression, *Computational Statistics and Data Analysis*, 50, 1287–1312.