Problems and Challenges in the Analysis of Complex Data: Static and Dynamic Approaches

Marco Riani, Anthony Atkinson and Andrea Cerioli

Abstract This paper summarizes results in the use of the Forward Search in the analysis of corrupted datasets, and those with mixtures of populations. We discuss new challenges that arise in the analysis of large, complex datasets. Methods developed for regression and clustering are described.

1 Introduction

Data are an overwhelming feature of modern life. As the amount of data increases so do the challenges facing the statistician in trying to extract information from ever larger data sets. We argue that larger data sets are also more complex and require flexible multiple analyses in order to reveal their structure. Only then can all information be efficiently extracted.

The analysis of large data sets may be complicated by the high dimensionality of responses, large numbers of observations and complexity of the choices to be made among explanatory variables. Although appreciable, these challenges to the statistician are not different in kind from those faced in the analysis of smaller data sets. We, however, focus on problems that become severe in large complex data sets, such as our inability to find a single model for all the data.

The simplest situation is that of a single model with possibly many outliers. In the presence of a core population and some isolated or clustered outliers, traditional robust methods (Maronna et al. 2006) can be successfully used to find proper models. However, when there are several populations and different sources of

M. Riani (✉) · A. Cerioli
Department of Economics, University of Parma, Italy
e-mail: mriani@unipr.it; andrea.cerioli@unipr.it

A. Atkinson
Department of Statistics, London School of Economics, London WC2A 2AE, UK
e-mail: a.c.atkinson@lse.ac.uk

A. Di Ciaccio et al. (eds.), *Advanced Statistical Methods for the Analysis of Large Data-Sets*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-21037-2_14, © Springer-Verlag Berlin Heidelberg 2012

heterogeneity, traditional robust methods fail to recover the real structure of the data and more sophisticated procedures, such as those derived from the Forward Search (FS) (Atkinson and Riani 2000; Atkinson et al. 2004) are required. Some examples are presented in the next section. Section 3 introduces an example of complex data in a situation where automatic procedures need to be developed. We conclude with a longer example of robust model building.

2 Some Difficulties in Data Analysis

2.1 The Presence of Outliers

The presence of atypical observations may strongly and wrongly influence the output of statistical analyses. When the number of observations is large it is likely that there will be several atypical observations which mask one another. They will not be revealed by a single static analysis, although the dynamic analysis of many subsets of data through the FS will reveal such structure. However, the outliers should not be seen only as bad observations that estimation procedures must avoid; they may themselves contain valuable information. The discovery of the hole in the ozone layer is one example. In drug development, the existence of a subset of individuals with an adverse reaction to the drug might be one target of the analysis.

2.2 Calibration of Test Procedures

Particulary as the sample size grows, it is necessary to calibrate tests of the outlyingness of individual observations. The repeated application of statistical tests makes it necessary to calibrate for simultaneity. Use of procedures that are correctly calibrated to provide tests of the desired size will keep false alarms under control (Riani et al. 2009).

2.3 Subpopulations

Large datasets often contain hidden groups, which are not revealed by application of single population methods, even in their robustified forms. For multivariate data there are well established techniques of cluster analysis, which may work well for normal populations. However, automatic methods such as MCLUST (Fraley and Raftery 1999) for establishing cluster membership often indicate too many clusters. Standard clustering procedures are also badly affected by the presence of outliers. Procedures based on the forward search have been shown to work well in identifying clusters and establishing cluster membership, even in the presence of outliers, but are far from automatic, requiring appreciable input from the statistician.

3 An Example of Large Complex Corrupted Data

As an illustration of the problems involved with the analysis of complex data, consider the example given in Fig. 1 referred to the quantity (x) and the value (y) of 4719 import declarations of a specific technological product. This is an example of one of the thousands of datasets provided by the "Office Européen de Lutte Anti-Fraude" (OLAF) or by its partners in the Member States. The purpose is to find atypical transactions, which might correspond to cases of potential fraud (e.g. the evasion of import duties) or to potential money laundering activities.

The observations appear roughly distributed along three main lines departing from the origin of the coordinate axes. However, there seem also to be horizontal strips of concentrated data. It is certainly not clear how many groups are present in the data. Traditional methods which assume one single regression population will fail in revealing the real structure as will their robust counterparts. The general structure is of a mixture of linear models heavily contaminated by observations that do not follow the general pattern (Riani et al. 2008). Outliers may be isolated, originating from recording errors during the data collection process, or they may be clustered, when they represent some systematic behaviour. In the context of anti-fraud the outliers themselves are important. However, the size of any outlier tests needs to be calibrated: prosecutions which fail encourage fraudsters while law enforcement agencies will become discouraged.

Use of the "multiple random starts forward search" (Atkinson and Riani 2007) enables us to dissect these data into components and outliers. However, the clustering of regression lines is again a procedure that involves considerable statistical intervention. The context of anti-fraud indicates several important directions for statistical development.

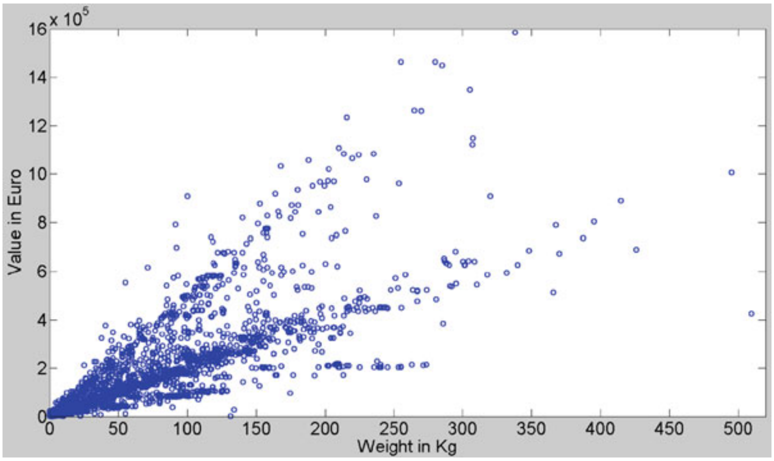

Fig. 1 An example of international trade data

4 Forward Directions for the Forward Search

4.1 Automatic Classification Procedures

While for a small number of datasets it is possible to envisage human intervention for each dataset, including the use of exploratory data analysis tools, in the presence of a huge amount of data only automatic procedures are feasible. These developments are required both for the clustering of multivariate data mentioned in Sect. 2 and for the mixtures of regression lines of Sect. 3.

4.2 Timeliness and On-Line Systems

The context of anti-fraud data analysis motivates the need for timeliness, which may only be achievable through on-line analysis. If a fraud is being committed it needs to be detected and prevented as quickly as possible. An important challenge in on-line analysis is to disseminate the results in a form that is again understandable by the final users. The importance of timeliness and on-line systems accentuates the need for research into the theoretical and practical aspects of dynamic updating methods.

4.3 Automatic Model Selection Procedures

For simple regression models with several explanatory variables the FS provides a robust form of the Cp statistic for selection of regression models. However, for high-dimensional data, the phrase "model selection" refers in addition to the choices of distribution of the responses and the functional form between the explanatory variables and the response.

5 Choosing Regression Models with Mallow's C_p

The analysis of the issues raised in the previous section requires book-length treatment. In this section we concentrate on the issue of model selection to illustrate how a robust flexible trimming approach (specifically that provided by the forward search), makes it possible to get inside the data in a manner impossible using standard statistical methods, be they robust or non-robust.

5.1 Background and Aggregate C_p

Mallows' C_p is widely used for the selection of a model from among many non-nested regression models. However, the statistic is a function of two residual sums of squares; it is an aggregate statistic, a function of all the observations. Thus C_p suffers from the well-known lack of robustness of least squares and provides no evidence of whether or how individual observations or unidentified structure are affecting the choice of model. In the remainder of this paper we describe a robust version of C_p that relies on the forward search to choose regression models in the presence of outliers. Theoretical details are given by Riani and Atkinson (2010). Here we provide a brief survey of the main results, before concentrating on a complex example.

There are n univariate observations y. For the linear multiple regression model $y = X\beta + \epsilon$, X is an $n \times p$ full-rank matrix of known constants, with ith row x_i^T. The normal theory assumptions are that the errors ϵ_i are i.i.d. $N(0, \sigma^2)$. The residual sum of squares from fitting this model to the data is $R_p(n)$.

In the calculation of C_p, σ^2 is estimated from a large regression model with $n \times p^+$ matrix X^+, $p^+ > p$, of which X is submatrix. The unbiased estimator of σ^2 comes from regression on all p^+ columns of X^+ and can be written $s^2 = R_{p^+}(n)/(n - p^+)$. Then

$$C_p = R_p(n)/s^2 - n + 2p = (n - p^+)R_p(n)/R_{p^+}(n) - n + 2p. \qquad (1)$$

Provided the full model with p^+ parameters and the reduced model with p parameters yield unbiased estimates of σ^2, it follows that $\mathrm{E}(C_p)$ is approximately p.

Models with small values of C_p are preferred. Statements are often made that those models with values of C_p near p are acceptable. However, we find it helpful to use the distribution of the statistic which Mallows (1973) shows is a scaled and shifted F.

5.2 The Forward Search and Forward C_p

The forward search for a single regression model fits subsets of observations of size m to the data, with $m_0 \leq m \leq n$. Least squares on the subset of m observations yields a residual sum of squares $R_p(m)$. The C_p criterion (1) for all observations is a function of the residual sums of squares $R_p(n)$ and $R_{p^+}(n)$. For a subset of m observations we can define the forward value of C_p as

$$C_p(m) = (m - p^+)R_p(m)/R_{p^+}(m) - m + 2p. \qquad (2)$$

For each m we calculate $C_p(m)$ for all models of interest.

Some care is needed in interpreting this definition. For each of the models with p parameters, the search may be different, and outliers, if any, may not enter in the same order for all models.

The distributional results of Mallows apply when C_p is calculated from the full sample. But, in the forward search with $m < n$ we order the observations during the search and take the central m residuals to calculate the sums of squares $R_{p^+}(m)$ and $R_p(m)$. These sums of squares are accordingly based on truncated samples and will have smaller expectations than those based on a full sample of m observations. However, Riani and Atkinson (2010) show that the full sample distribution holds to a good approximation with n replaced by m. That is

$$C_p(m) \sim (p^+ - p)F + 2p - p^+, \qquad \text{where} \qquad F \sim F_{p^+ - p, m - p^+}, \qquad (3)$$

which is Mallows' result with a change in the degrees of freedom of the F distribution.

6 Credit Card Data

6.1 Background and Aggregate Model Selection

As an example we analyse data on factors influencing the use of credit and other cards. The data are appreciably larger and more complex than those customarily used to illustrate aggregate C_p. There are $1,000$ observations on the most active customers of a bank operating in the north of Italy. There is one response and nine explanatory variables which are listed in Appendix A.

Figure 2 gives the traditional C_p plot for these data, showing only those models with the smallest values of C_p for values of p from 3 to 9. The minimum value of C_p is for $p = 6$, for a model containing a constant and variables 1 2 4 5 and 6. There is a pleasing structure to this plot in that there is a well defined series of submodels that have the smallest C_p values. For $p = 5$ we have similar values for 1 2 4 and 5 and for 1 2 4 and 6. For $p = 3$ the best model seems to be 1 2 and 4, with 1 and 2 best for $p = 3$, although the value of C_p for this model lies above the 97.5% bound.

The models suggested by the plot are summarised in Table 1 in which the hierarchy of suggested models is clearly indicated. The table indicates that statistical procedures for checking the model should start with $p \leq 6$.

6.2 The Generalized Candlestick Plot

When we apply the forward search to model selection we obtain a forward plot of $C_p(m)$ for each model. Thus the points in Fig. 2 are replaced by the curves

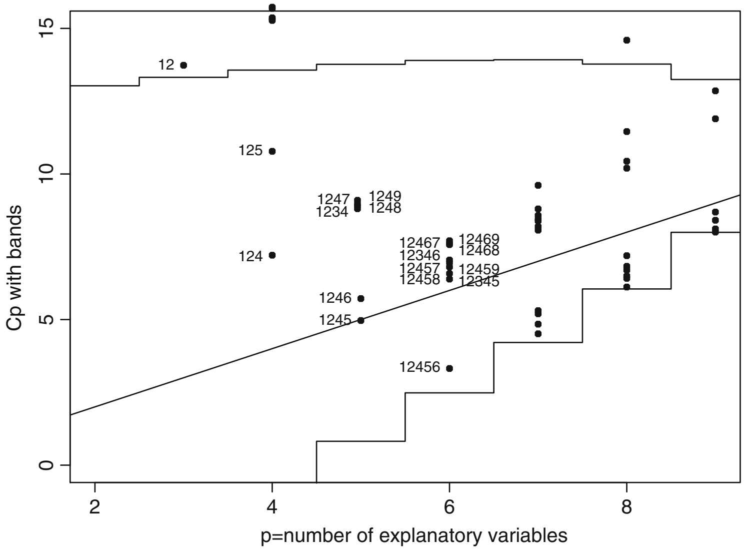

Fig. 2 Credit card data: C_p plot. There is a simple hierarchy of good models. Bands are the 2.5% and 97.5% points of the scaled and shifted F distribution of C_p

Table 1 Credit card data: some models selected by C_p and by $C_p(m)$ in the candlestick plot of Fig. 3

p	Variables	p	Variables
Non-robust C_p		$C_p(m)$	
		7	1 2 3 4 5 8
6	1 2 4 5 6	6	1 2 3 4 5
5	1 2 4 5	6	1 2 3 5 8
5	1 2 4 6	5	1 2 3 5
4	1 2 4	4	1 2 5
3	1 2		

of forward plots for all values of m that are of interest. For example, Atkinson and Riani (2008) give separate plots of forward C_p for values of p from 4 to 7. The resulting quantity of graphical output can be overwhelming. We accordingly illustrate the plot introduced by Riani and Atkinson (2010) that cogently summarises this information.

The plot summarises, for each model, the information in the trajectory of the forward plots of $C_p(m)$. The starting point is the “candlestick” plot used to summarise such quantities as the high, low and closing values of stocks. Google provides many references. However, we need a generalization of this plot. Since we expect any outliers to enter the search towards the end, the last few values of $C_p(m)$ are of particular interest, as is the comparison of these values with earlier average behaviour.

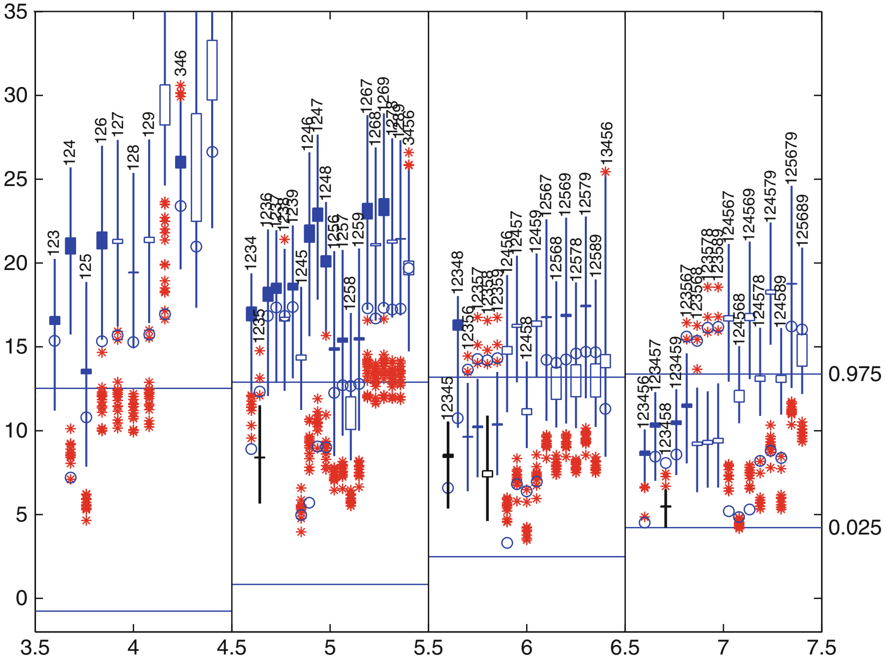

Fig. 3 Credit card data: generalized candlestick plot of $C_p(m)$ for the best models in the range $m = 900-1{,}000$. The last 20 observations to enter the individual searches are marked if they lie outside the candlestick. Models 1 2 3 5, 1 2 3 4 5, 1 2 3 5 8 and 1 2 3 4 5 8 are highlighted by a thick vertical line

Figure 3 gives the generalized candlestick plot for the values of $C_p(m)$ for the credit card data. The figure includes all models that were among the five best for $m \geq 900$, with symbols for the last 20 values if they are extreme. In order to check the indication of $p = 6$ as providing the largest model, we plot values in the range $p = 4$ to $p = 7$.

The plot depends on the range of values of m which define a "central part" of the plot. With 1,000 observations we take as the central part of the search values of m in the range 900–980. The figure includes all models that were among the five best for $m \geq 900$, with symbols for the last 20 values if they lie outside the "candle." The vertical lines in the plot summarise the values of $C_p(m)$ for each model in the central part of the search. The definition of the candlesticks is:

Lowest Value; minimum in the central part of the search;
Highest Value; maximum in the central part of the search;
Central Box; mean and median of the values in the central part of the search; filled if mean < median;
Stars; the values in steps "central part" + 1 to $n - 1$ if these lie outside the box;
Unfilled Circle; the final value.

Thus each point in the standard non-robust C_p plot such as Fig. 2 is replaced by a single vertical line and a series of extra symbols.

We start by looking at models for $p = 6$. The value of $C_p(m)$ for model 1 2 3 4 5 seems unaffected by exclusion of the last 20 observations. However, that for 1 2 4 5 6, which was the indicated model at the end of the search, increases to lie mostly above the bound when the observations are excluded. On the contrary, under the same conditions the values for 1 2 4 5 8 decrease, for it to become one of the two best models. If we now turn to $p = 7$, we see that the union of these models, that is 1 2 3 4 5 8, has a stable small value of $C_p(m)$.

The conclusions for $p = 5$ are straightforward: 1 2 3 5 is the only model which lies within the bounds for the central part of the search. This is a special case of the two models for $p = 6$ suggested above. Figure 3 indicates clearly that there is no satisfactory model with $p = 4$, although 1 2 5 is the best of a bad bunch. These models are also listed in Table 1.

The general shape of the plot in Fig. 3 is similar to that of the non-robust C_p plot in Fig. 2. However, for small values of p, many models have relatively small values of $C_p(m)$ only over the last values of m whereas, for larger p, there are many models with small values of $C_p(m)$ over most of the range. There is also a decrease in variability in the values of $C_p(m)$ as p increases. When p is too small, the values of $C_p(m)$ respond with extreme sensitivity to the addition of extra observations.

6.3 Outlier Detection

The ordering of observations by the forward search enables us to pinpoint the influential effect of individual observations. Table 1 shows appreciable change in the models selected as the last twenty observations are deleted. We accordingly now see whether there is evidence that some of these observations are outlying.

To detect outliers we use forward plots of minimum deletion residuals, with envelopes (Riani and Atkinson 2007). The left-hand panel of Fig. 4 is a forward plot of all such residuals for all 1,000 observations when the model fitted is 1 2 3 5. It is clear, from the exceedances of the upper threshold in the range m from 980 to 995, that there are some outliers, although the exact number is not obvious. With a large sample, the presence of several outliers has led to masking, so that departures are less extreme when $m = n$ than they are earlier in the search. Similar phenomena occur for multivariate data when forward plots of the minimum Mahalanobis distance are used for the detection of outliers. Riani et al. (2009) propose a rule that allows for masking and simultaneous inferences to provide an outlier detection rule with a size close to 1%. Torti and Perrotta (2010) amend the rule for regression. In the credit card data we detect between eight and ten outliers, the exact number depending on the model being fitted. Fortunately, the set of ten outliers contains that of nine or eight for all models of interest.

The forward plot of minimum deletion residuals for 990 observations is shown in the right-hand panel of Fig. 4. It shows that, for this model, there are no more

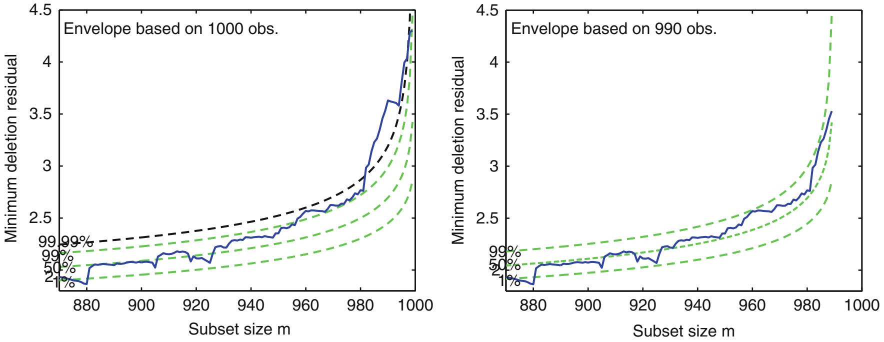

Fig. 4 Credit card data, outlier detection, model with variables 1 2 3 and 5. Upper panel, forward plot of minimum deletion residual $r_i(m)$, with 1%, 50%, 99% and 99.99% envelopes for $n = 1,000$. Lower panel, forward plot after deletion of ten observations; envelopes for $n = 990$

than ten outliers. Compared with the left-hand panel, the most extreme values occur at the end of the search, indicating that the observations are correctly ordered by the search and that there are no remaining outliers. The presence of these outliers explains the structure of the "starred" observations in Fig. 3. The outliers are causing the majority of the models to have small values of $C_p(m)$ towards the end of the search.

6.4 Model Building and Checking

The independent identification of outliers in the credit card data justifies the selected models listed in Table 1. It is interesting to investigate some of these models a little further.

Table 2 gives t values, when $n = 990$, for the terms of the best models in Table 1 for $p = 5$, 6 and 7. The models were selected by our interpretation of the generalized candlestick plot of Fig. 3. Model 1 2 3 4 5 8 is highlighted in the figure as the best model for $p = 7$. If we remove the least significant term, that for x_8, we obtain the stable model 1 2 3 4 5 with little change of the t values for the remaining terms. Now x_4 is the least significant term. Its omission, however, causes an increase in the t statistic for x_2 from 5.84 to 9.47. In this model all terms are significant at the 1% level.

Figure 5 give forward plots of $C_p(m)$ for a few selected models. This extends the information available from the generalized candlestick plot. For model 1 2 3 4 5 the values of $C_p(m)$ remain within the 97.5% bound throughout and are stable at the end of the search. But the figure shows how the values of $C_p(m)$ for the simpler model 1 2 3 5 are affected by the last few observations to enter. The plot also shows how model 1 2 4 5 6 was chosen by its small value of C_p at the very end of the search. However, values of $C_p(m)$ earlier in the search lie well above the 97.5% bound.

Table 2 Credit card data: t statistics of the three best models of Table 1 after removing outliers ($n = 990$)

Term	Model		
	1 2 3 5	1 2 3 4 5	1 2 3 4 5 8
Intercept	49.11	49.18	49.34
x_1	6.37	6.05	6.22
x_2	9.47	5.84	5.78
x_3	2.64	2.70	2.73
x_4	–	2.28	2.71
x_5	2.78	2.52	3.00
x_8	–	–	−2.29

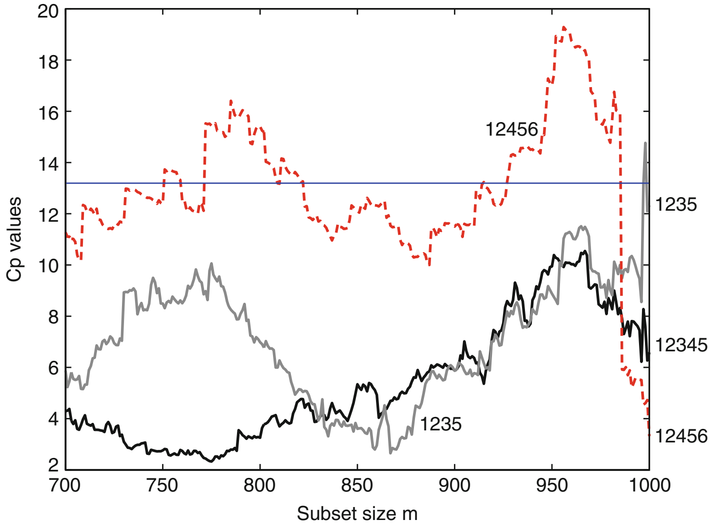

Fig. 5 Credit card data: forward plots of $C_p(m)$ for selected models from Table 1

A final comment on model choice is to compare 1 2 3 5 and 1 2 3 4 5 over the values of m from 700. Although the two models give very similar values of $C_p(m)$ for $m = 850$–980, the larger model is superior in the first part of the plot. Since the value of $C_p(m)$ is also unaffected by the outlying observations, we would recommend this as the chosen model.

7 Computation

The Matlab software used in this paper is part of the FSDA (Forward Search Data Analysis) toolbox which can be downloaded, free of charge, from the webpage `www.riani.it` in the section "Matlab code". Full documentation is included.

Appendix A: The Credit Card Data

Variables that are given as amount are in euros and are either annual totals or averages, depending on the nature of the variable.

y Amount of use of credit, debit and pre-paid card services of the bank

x_1 Direct debts to the bank

x_2 Assigned debts from third parties

x_3 Amount of shares (in thousands of Euros)

x_4 Amount invested in investment funds (in thousands of Euros)

x_5 Amount of money invested in insurance products from the bank (in thousands of Euros)

x_6 Amount invested in bonds (in thousands of Euros)

x_7 Number of telepasses (Italian electronic toll collection system) of the current account holder

x_8 Number of persons from the bank dealing with the management of the portfolio of the customer (min=0, max=4). This variable has many zeroes

x_9 Index of use of point of sale services

In x_7 the telepass is a debit card issued for each car. Since other forms of payment are possible, this variable also contains many zeroes.

References

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

Atkinson, A. C. and M. Riani (2007). Exploratory tools for clustering multivariate data. *Computational Statistics and Data Analysis 52*, 272–285. doi:10.1016/j.csda.2006.12.034.

Atkinson, A. C. and M. Riani (2008). A robust and diagnostic information criterion for selecting regression models. *Journal of the Japanese Statistical Society 38*, 3–14.

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

Fraley, C. and A. E. Raftery (1999). Mclust: software for model-based cluster analysis. *Journal of Classification 16*, 297–306.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics 15*, 661–675.

Maronna, R. A., D. R. Martin, and V. J. Yohai (2006). *Robust Statistics: Theory and Methods*. New York: Wiley.

Riani, M. and A. C. Atkinson (2007). Fast calibrations of the forward search for testing multiple outliers in regression. *Advances in Data Analysis and Classification 1*, 123–141. doi:10.1007/s11634-007-0007-y.

Riani, M. and A. C. Atkinson (2010). Robust model selection with flexible trimming. *Computational Statistics and Data Analysis 54*, 3300–3312. doi:10.1016/j.csda.2010.03.007.

Riani, M., A. C. Atkinson, and A. Cerioli (2009). Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society, Series B 71*, 201–221.

Riani, M., A. Cerioli, A. Atkinson, D. Perrotta, and F. Torti (2008). Fitting mixtures of regression lines with the forward search. In F. Fogelman-Soulié, D. Perrotta, J. Piskorski, and R. Steinberger (Eds.), *Mining Massive Data Sets for Security*, pp. 271–286. Amsterdam: IOS Press.

Torti, F. and D. Perrotta (2010). Size and power of tests for regression outliers in the forward search. In S. Ingrassia, R. Rocci, and M. Vichi (Eds.), *New Perspectives in Statistical Modeling and Data Analysis*. Heidelberg: Springer-Verlag, pp. 377–384.

Tufte, E. (2001). *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.