Journal of the Korean Statistical Society 39 (2010) 117–134

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

journal homepage: www.elsevier.com/locate/jkss

The forward search: Theory and data analysis

Anthony C. Atkinson [a,*], Marco Riani [b], Andrea Cerioli [b]

[a] Department of Statistics, London School of Economics, London WC2A 2AE, UK
[b] Università di Parma, Via Kennedy 6, I-43100, Parma, Italy

ARTICLE INFO

Article history:
Received 26 January 2010
Accepted 23 February 2010
Available online 27 March 2010

Keywords:
Box–Cox transformation
Building models
Clustering
Constructed variable
C_p
Flexible trimming
Fraud detection
Kalman filter
Outliers
Ozone data
Robustness
Simultaneous inference

ABSTRACT

The Forward Search is a powerful general method, incorporating flexible data-driven trimming, for the detection of outliers and unsuspected structure in data and so for building robust models. Starting from small subsets of data, observations that are close to the fitted model are added to the observations used in parameter estimation. As this subset grows we monitor parameter estimates, test statistics and measures of fit such as residuals. The paper surveys theoretical development in work on the Forward Search over the last decade. The main illustration is a regression example with 330 observations and 9 potential explanatory variables. Mention is also made of procedures for multivariate data, including clustering, time series analysis and fraud detection.

Contents

1. Introduction 118
2. The forward search and outliers 119
3. Regression 119
   3.1. Least squares and likelihood 119
   3.2. Testing for outliers 120
   3.3. Envelopes from order statistics 120
   3.4. Transformations of the response and the fan plot 121
   3.5. Building regression models 122
4. The ozone data 123
5. One multivariate sample 130
6. Clustering 131
7. The analysis of time series 131
8. Integrated analysis 132
   References 132

* Corresponding author.
E-mail addresses: a.c.atkinson@lse.ac.uk (A.C. Atkinson), mriani@unipr.it (M. Riani), andrea.cerioli@unipr.it (A. Cerioli).

1226-3192/$ – see front matter © 2010 The Korean Statistical Society. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.jkss.2010.02.007

1. Introduction

Data are often corrupted, that is they contain outliers. Models are often also inadequate, that is they may contain systematic failures for specific sets of data. The identification of outliers and the immunization of data analysis against both outliers and failures of modelling are important aspects of modern statistics. The Forward Search discussed in this paper is a graphics rich approach that leads to the formal detection of outliers and to the detection of model inadequacy combined with suggestions for model enhancement. The key idea of our approach to the Forward Search is to monitor quantities of interest, such as parameter estimates and test statistics, as the model is fitted to data subsets of increasing size.

The detection of outliers has a long history. Over the centuries scientists have used robust common sense to discard data that look suspicious. For example, the data on the boiling point of water presented in Forbes (1857) contain one outlier from the regression model. According to Weisberg (2005, p. 38), Forbes noted on his copy of the paper that this pair of values was "evidently a mistake". With larger and more complicated data sets more formal methods of outlier detection have had to be developed.

The development of methods for outlier detection and accommodation is more advanced than those for model enhancement. Box (1953) is credited with being the first to introduce the term 'robust' for procedures that are little affected by the presence of outliers. However, there are several approaches to the presence and detection of outliers:

1. Outlier detection may be subsidiary to model fitting by, for example, least squares. Survey data are often pre-processed to remove outliers (such as 17 year old mothers with five children).
2. The outliers themselves may be the primary research interest, for example in anti-fraud and counter-terrorism.
3. In many applications, such as marketing and medicine, both aspects are important.

Given these different aims it is not surprising that the systematic study of outliers has led to remarkably different points of view:

- Testing: an emphasis on formal tests of hypotheses for precise outlier identification;
- Diagnostics: less formal diagnostic tools for exploratory data analysis and the enhancement of the intuitive inspection of data;
- Robustness: formal methods for outlier accommodation through the downweighting of observations.

The theme of this paper is that the Forward Search provides a reconciliation of these three approaches.

There is an appreciable history of the ideas of outlier detection and robustness, which is summarised in the opening chapters of several books. Hawkins (1980) is concerned with tests for outlier detection. The emphasis in the books of Cook and Weisberg (1982) and Atkinson (1985) is on diagnostic procedures, especially in regression. The Princeton Robustness Study, Andrews, Bickel, Hampel, Tukey, and Huber (1972), reports the results of a large simulation study on the properties of estimators of location in univariate samples. This was followed by the more mathematical approach to more general problems of Hampel, Ronchetti, Rousseeuw, and Stahel (1986) and of Huber (1981), Huber and Ronchetti (2009). The focus of Rousseeuw and Leroy (1987) is high-breakdown regression. A succinct historical survey of robust regression opens Rousseeuw (1984). A recent exposition of the theory and methods of robust statistics is Maronna, Martin, and Yohai (2006), while a thoughtful survey is provided by Morgenthaler (2007). Hadi, Imon, and Werner (2009) survey outlier detection.

The history of the Forward Search is much shorter. The idea of fitting a model to subsets of an increasing size was introduced by Hadi (1992) and Atkinson (1994) for multivariate data and by Hadi and Simonoff (1993) for regression. The ensuing development of the forward search, combining flexible downweighting with a focus on data analysis, is covered in two books, Atkinson and Riani (2000) on regression and Atkinson, Riani, and Cerioli (2004), which is concerned with multivariate procedures. The emphasis in our survey is on work since the appearance of these books which has emphasized the importance of distributional results and tests of hypotheses as an adjunct to graphical displays. Forbes's data are used as an introductory example for regression by Weisberg (2005) and to introduce the Forward Search by Atkinson and Riani (2000).

The structure of our paper is as follows. The theoretical background is outlined in Section 2. Section 3 presents distributional theory for outlier tests in regression during the Forward Search and describes methods for testing response transformation and for model building, using a forward version of C_p. In the next, longest, section we exemplify the use of our procedure through the analysis of a complex regression dataset on ozone concentration in the Los Angeles area. Sections 5–7 briefly review recent work on the Forward Search in multivariate analysis, clustering and time series. Section 8 concludes, providing suggestions for additional research; the discussion is focused on fraud detection.

These are only some of the applications and developments that we could have discussed. We stress that this approach to data analysis is very general and can be applied in many scientific disciplines and to many data structures. For example, Crosilla and Visentini (2007) used the Forward Search in the classification and segmentation of 3-D objects surveyed with high density laser scanning techniques. Mavridis and Moustaki (2010) extended the technique to latent factor analysis for binary data. Solaro and Pagani (2010) used the Forward Search in the context of multidimensional scaling and, finally, Cheng and Biswas (2008) develop methods for the analysis of mixed continuous and categorical data.

2. The forward search and outliers

To test for outliers requires division of the data into two groups. The parameters are estimated from m observations, believed to come from the uncontaminated distribution model. The estimates are then used in the test for the outlyingness of one or more of the remaining $n-m$ observations. Sequential incrementation of the set of observations believed to be good by observations close to the fitted model provides an ordering of the data. We can then monitor any properties of interest, such as parameter estimates and test statistics to see whether our inferences change as m increases.

For most of the paper we are concerned with independent observations, so that the loglikelihood is a sum of the contributions of individual observations. Let the loglikelihood of all observations be $L_n(\theta)$. The forward search fits subsets of observations of size m to the data, with $m_0 \leq m \leq n$. Let $S_*^{(m)}$ be the subset of size m found by the forward search, for which the m.l.e. of the $p \times 1$ parameter θ is $\hat{\theta}_*(m)$. Then the loglikelihood for these m observations is

$$L_m\{\hat{\theta}_*(m)\} = \sum_{i \in S_*^{(m)}} l_i\{\hat{\theta}_*(m)\}, \tag{1}$$

where $l_i\{\hat{\theta}_*(m)\}$ is the likelihood contribution of observation i. Likelihood contributions can be calculated for all observations including those not in $S_*^{(m)}$.

The search moves forward with the augmented subset $S_*^{(m+1)}$ consisting of the observations with the $m+1$ largest values of $l_i\{\hat{\theta}_*(m)\}$. Usually one observation enters the subset at each step, but sometimes two or more enter, when one or more then leave. Such occurrences are indicative of changes in structure or of clusters of outliers. The estimate of the parameter from this new subset is $\hat{\theta}_*(m+1)$.

To start we take $m_0 = p$ or slightly larger. The details of finding the initial subset depend on the particular application. There are, however, two broad strategies. If the data are believed to be homogeneous, with perhaps a few outliers, we start from a point carefully chosen by trimming. However, if the data may be clustered, we run searches from random starting points, allowing exploration of a wide number of data partitions into 'good' and 'bad'. In both cases our trimming is flexible, the proportion being determined by the data. In this there is a contrast with the trimmed likelihood methods of Neykov and co-workers, for example Müller and Neykov (2003), in which the trimming proportion is fixed in advance.

3. Regression

3.1. Least squares and likelihood

The literature on the detection of multiple outliers in regression is large. See, for example, Beckman and Cook (1983) or Barnett and Lewis (1994). A more recent review and comparison of methods is Wisnowski, Montgomery, and Simpson (2001). The central problem is that of "masking": if there are several outliers, least squares estimation of the parameters of the model from all n observations will not lead to identification of the m uncontaminated observations. The single deletion methods that form the core of diagnostic methods (for example Atkinson, 1985; Cook & Weisberg, 1982) will also fail. Hawkins (1983) argues for exclusion of all possibly outlying observations, which are then tested sequentially for reinclusion. Instead we develop the theory of the forward search in which observations are sequentially included.

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

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

For regression the likelihood contribution (1), omitting constants not depending on i, becomes

$$l_i\{\hat{\theta}_*(m)\} = -\{y_i - x_i^T\hat{\beta}_*(m)\}^2/2s_*^2(m) = -e_{i*}^2(m)/2s_*^2(m). \tag{2}$$

In (2) $\hat{\beta}_*(m)$ and $s_*^2(m)$ are the parameter estimates from the subset $S_*^{(m)}$.

The search thus moves forward with the augmented subset $S_*^{(m+1)}$ consisting of the observations with the $m+1$ smallest absolute values of $e_{i*}(m)$. An inferentially important consequence is that the estimates of the parameters are based only on those observations giving the central m residuals.

To start we take $m_0 = p$ and search over subsets of p observations to find the subset, out of 5000 or some such large number, that yields the least median of squares (LMS) estimate of β (Hampel, 1975; Rousseeuw, 1984). Alternatively we can use the least trimmed squares estimate (LTS) (Rousseeuw & Leroy, 1987). The forward search for a single population is not sensitive to the choice of starting point and there seems little difference in the searches from these two starting points, except at the very beginning of the search (Atkinson, 2009).

3.2. Testing for outliers

Before each observation is introduced into the subset used in fitting we test whether it is an outlier by calculating the deletion residual. These $n - m$ residuals are

$$r_{i*}(m) = \frac{y_i - x_i^T \hat{\beta}_*(m)}{\sqrt{s_*^2(m)\{1 + h_{i*}(m)\}}} = \frac{e_{i*}(m)}{\sqrt{s_*^2(m)\{1 + h_{i*}(m)\}}}, \tag{3}$$

where $h_{i*}(m) = x_i^T \{X_*(m)^T X_*(m)\}^{-1} x_i$; the leverage of each observation depends on $S_*^{(m)}$. Let the observation nearest to those constituting $S_*^{(m)}$ be $i_{\min}$ where

$$i_{\min} = \arg \min_{i \notin S_*^{(m)}} |r_{i*}(m)|$$

denotes the observation with the minimum absolute deletion residual among those not in $S_*^{(m)}$. To test whether observation $i_{\min}$ is an outlier we use the absolute value of the minimum deletion residual

$$r_{i_{\min *}}(m) = \frac{e_{i_{\min *}}(m)}{\sqrt{s_*^2(m)\{1 + h_{i_{\min *}}(m)\}}} \tag{4}$$

as a test statistic. If the absolute value of (4) is too large, the observation $i_{\min}$ is considered to be an outlier, as well as all other observations not in $S_*^{(m)}$.

The null distribution of each deletion residual from a sample of size n is t_{n-p-1}. However, during the search, we are taking subsamples of observations from the centre of the distribution. One way of finding the distribution of the residuals during the search is to simulate envelopes from perhaps 10,000 searches. A more flexible alternative is to find an analytical approximation to the envelopes based on order statistics.

3.3. Envelopes from order statistics

Since the test statistic (4) is the $(m + 1)$st ordered value of the deletion residuals, we can use distributional results for order statistics to obtain very good approximate envelopes for our plots.

Let $Y_{[m+1]}$ be the $(m + 1)$st order statistic from a sample of size n from a univariate distribution with c.d.f. $G(y)$. Then the c.d.f. of $Y_{[m+1]}$ is given exactly by

$$P\{Y_{[m+1]} \le y\} = \sum_{j=m+1}^{n} \binom{n}{j} \{G(y)\}^j \{1 - G(y)\}^{n-j}. \tag{5}$$

See, for example, Casella and Berger (2002, p. 228). Further, it is well known that we can apply properties of the beta distribution to the RHS of (5) to obtain

$$P\{Y_{[m+1]} \le y\} = I_{G(y)}(m + 1, n - m), \tag{6}$$

where

$$I_p(A, B) = \int_0^p \frac{1}{\beta(A, B)} u^{A-1}(1 - u)^{B-1} \mathrm{d}u$$

is the incomplete beta integral. From the relationship between the F and the beta distribution it is possible to rewrite Eq. (6) as

$$P\{Y_{[m+1]} \le y\} = P\left\{F_{2(n-m),2(m+1)} > \frac{1 - G(y)}{G(y)} \times \frac{m + 1}{n - m}\right\}, \tag{7}$$

where $F_{2(n-m),2(m+1)}$ is the F distribution with $2(n-m)$ and $2(m+1)$ degrees of freedom (Guenther, 1977). Thus, the required quantile of order γ of the distribution of $Y_{[m+1]}$ say $y_{m+1,n;\gamma}$ can be obtained as

$$y_{m+1,n;\gamma} = G^{-1}\left(\frac{m + 1}{m + 1 + (n - m)x_{2(n-m),2(m+1);1-\gamma}}\right), \tag{8}$$

where $x_{2(n-m),2(m+1);1-\gamma}$ is the quantile of order $1 - \gamma$ of the F distribution with $2(n - m)$ and $2(m + 1)$ degrees of freedom.

In our case we are considering the absolute values of the deletion residuals. If the c.d.f. of the t distribution on ν degrees of freedom is written as $T_\nu(y)$, the absolute value has the c.d.f.

$$G(y) = 2T_\nu(y) - 1, \quad 0 \le y < \infty. \tag{9}$$

To find percentage points of $Y_{(k)}$ we only need numerical inversion of the t and F distributions. For a subset of size m we put $k = m + 1$. With $G(y)$ given by (9) and $\nu = m - p$ we obtain $V_{m,\alpha}$ as the $100\alpha\%$ point of the distribution of the k-th order statistic of the absolute value of a t random variable, that is of the t distribution folded at the origin.

If we had an unbiased estimator of σ^2 the envelopes would be given by $V_{m,\alpha}$ for $m = m_0, \ldots, n-1$. However, the estimator $s^2(m^*)$ is based on the central m observations from a normal sample. (Strictly the m observations with smallest squared residuals based on the parameter estimates from $S_*^{(m-1)}$, which may not always be quite the same thing). The variance of the truncated normal distribution containing the central m/n portion of the full distribution, called $\sigma_T^2(m)$ by Riani and Atkinson (2007, Section 7), follows from standard results (Johnson, Kotz, & Balakrishnan, 1994, pp. 156–162). Since the outlier tests we are monitoring are divided by an estimate of σ^2 that is too small, we need to scale up the values of the order statistics to obtain the envelopes

$$V^*_{m,\alpha} = V_{m,\alpha}/\sigma_T(m).$$

An entertaining alternative derivation is given by Riani, Atkinson, and Cerioli (2009, Appendix 2) who use results on elliptical truncation due to Tallis (1963). On the way their result links the c.d.f. of the χ^2_3 in an unexpected way to the density and c.d.f. of the standard normal distribution.

The resulting envelopes give an excellent approximation to the pointwise distribution of the t-statistics and so provide a test of size $1-\alpha$ for the outlyingness of individual observations. The results of Atkinson and Riani (2006) show that there is a surprisingly high probability that the sequence of deletion residuals produced by the forward search will fall outside, for example, the 95% limits of the distribution at least once. However, we are concerned with simultaneous tests of outlyingness and want procedures that, in the null case of no outliers, indicate outliers in $100(1-\alpha)\%$ of the datasets. Riani et al. (2009) give a decision rule for the detection of multivariate outliers that reacts to several possible patterns of exceedance of envelopes of different levels. Torti and Perrotta (in press) show that this rule can be directly used for regression, with good size and power and we employ it here.

The procedure of Riani et al. (2009) is based on a two-stage process. In the first stage a search is run on the data, monitoring the bounds for all n observations until a "signal" is obtained. This signal indicates that observation $m^\dagger$, and therefore succeeding observations, may be outliers, because the value of the statistic lies beyond the chosen threshold. In the second part of the process, if a signal has been detected, the envelopes are recomputed for increasing sample sizes $n^* = m^\dagger - 1, m^\dagger, m^\dagger + 1, \ldots$. These new envelopes are then superimposed on the trajectory of absolute $r_{i_{\min}*}(m)$ until the observation to enter is recognised as causing the presence of an outlier. An example is in Section 4.

In addition to outlier detection we are interested in inference about aspects of the regression model, such as the evolution of the values of the parameter estimates, of R^2 and of test statistics. In addition we may need to transform the response in the regression model and to select among the many regressors. We now consider the application of the forward search to these problems.

3.4. Transformations of the response and the fan plot

To explore the dependence of the information about transformation of the response on particular observations we use a forward version of the approximate score test introduced by Atkinson (1973) for the value of the transformation parameter λ in the Box and Cox (1964) family of normalized power transformations

$$z(\lambda) = \begin{cases} (y^\lambda - 1)/(\lambda \dot{y}^{\lambda-1}) & \lambda \neq 0 \\ \dot{y} \log y & \lambda = 0, \end{cases} \tag{10}$$

where the geometric mean of the observations is written as $\dot{y} = \exp(\Sigma \log y_i/n)$.

The statistic is derived by Taylor series expansion of (10) as

$$\begin{aligned} z(\lambda) &\doteq z(\lambda_0) + (\lambda - \lambda_0) \left.\frac{\partial z(\lambda)}{\partial \lambda}\right|_{\lambda=\lambda_0} \\ &= z(\lambda_0) + (\lambda - \lambda_0) w(\lambda_0), \end{aligned} \tag{11}$$

which only requires calculations at the hypothesized value λ_0. In (11) $w(\lambda_0)$ is the constructed variable for the transformation. Differentiation of $z(\lambda)$ for the normalized power transformation yields

$$w(\lambda) = \begin{cases} y^\lambda \{\log(y/\dot{y}) - 1/\lambda\}/(\lambda \dot{y}^{\lambda-1}) + c(\lambda) & \lambda \neq 0 \\ \dot{y} \log y (0.5 \log y - \log \dot{y}) + c(0) & \lambda = 0, \end{cases} \tag{12}$$

where $c(\lambda)$ and $c(0)$ are constants depending on the data only through $\dot{y}$.

The approximate score statistic for testing the transformation $T_p(\lambda_0)$ is the t statistic for regression on $w(\lambda_0)$ in

$$z(\lambda) = x^T \beta + \gamma\ w(\lambda_0) + \epsilon, \tag{13}$$

where $\gamma = -(\lambda - \lambda_0)$. The statistic can either be calculated directly from the regression in (13), or from the formulae for added variables in Atkinson and Riani (2000, Chapter 4) in which multiple regression on x is adjusted for the inclusion of an

additional variable. The t test for $\gamma = 0$ in (13) is then the test of the hypothesis $\lambda = \lambda_0$. If the model includes a constant term, the $c(.)$ in (12) do not affect the value of the statistic and can be treated as zero.

In the Forward Search we use the added variable formulation, running the search for regression of $z(\lambda_0)$ on $X\beta$. Because the constructed variable $w(\lambda_0)$ is not included in the search, the t-statistic has approximately a t distribution. It cannot be exactly a t distribution since the constructed variable (12) is a function of the response, so that the response and the constructed variable are not independent. Atkinson and Riani (2002) explore the distribution of the statistic and show that it becomes more nearly t as the strength of the regression, for example the value of R^2, increases.

For moderate sized data, such as the $n = 48$ of one of the examples of Box and Cox (1964), we run forward searches for five values $\lambda : -1, -0.5, 0, 0.5$ and 1. For each we plot the value of the t statistic against m, combining the curves for the five values of λ in a single plot, which we call the fan plot. An example is in Section 4. For larger data sets a finer grid of values of λ is sometimes more informative.

3.5. Building regression models

We now turn attention to the choice of the predictors that constitute X in the regression model.

For individual variables we can use t tests. However, during a search on uncontaminated data, the values of $\hat{\beta}(m)$ remain sensibly constant, whilst the values of $s^2(m)$ increase with m. The values of the t statistics for the individual parameters then decrease, often highly, during the search. It is clear from such plots as those in Figures 1.8, 1.14 and 3.4 of Atkinson and Riani (2000) that forward plots of these t statistics are uninformative. Instead we extend the idea of the constructed variable test described in Section 3.4.

The central idea is to rewrite the regression model as

$$y = X\beta + \epsilon = X_{-j}\theta + x_j\gamma + \epsilon, \tag{14}$$

where γ is a scalar. We in turn take each of the columns of X as the vector $w = x_j$ (except the column corresponding to the constant term in the model). Thus if the columns of X are a constant and the $p - 1$ regression variables x_2 to x_p, we exclude each variable in turn and reinclude it as w. We perform a forward search using only the variables in each X_{-j} and then use the approach of added variables to calculate the t test for the inclusion of x_j in a manner orthogonal to the search.

Added variable t-tests were introduced by Atkinson and Riani (2002) who argued that the orthogonality of the residual added variable to the residuals of the remaining explanatory variables used in constructing the t test was sufficient to cancel the effect of ordering during the search and so to guarantee an exact t distribution for the test statistic. However, Riani and Atkinson (in press) note that, in addition, account needs to be taken of the truncated nature of the errors of observation at the beginning of the search. Results such as those of Box and Watson (1962) suggest that such truncations of error have a negligible effect on the distribution of test statistics which may be taken as, in this case, t. Indeed, an extensive simulation by Riani and Atkinson (in press) shows no departure from the null distribution even at the very early stages of the search, that is for values of m which are almost always unimportant for inferences. The forward distribution of the added variable tests can then be taken as t to a high degree of accuracy.

It is well known that t tests of single parameters can be misleading for finding the best subset of correlated predictors. Atkinson and Riani (2008) accordingly introduced a forward version of C_p (Mallows, 1973).

For all n observations let the residual sum of squares from fitting a p parameter model to the data be $R_p(n)$. The calculation of C_p also requires an estimate of σ^2 which comes from a large regression model with $n \times p^+$ matrix X^+, $p^+ > p$, of which X is submatrix. The unbiased estimator of σ^2 can then be written

$$s^2 = R_{p^+}(n)/(n - p^+). \tag{15}$$

Then

$$C_p = R_p(n)/s^2 - n + 2p = (n - p^+)R_p(n)/R_{p^+}(n) - n + 2p. \tag{16}$$

Models with small values of C_p are preferred. Although it is often said that models with values of C_p near p are acceptable, we find it helpful to consider the distribution of the statistic which is given, for example, by Mallows (1973) and by Gilmour (1996) who show that

$$C_p \sim (p^+ - p)F + 2p - p^+, \quad \text{where } F \sim F_{p^+-p,n-p^+}. \tag{17}$$

In (17) F is the test statistic for the inclusion of the extra $p^+ - p$ explanatory variables that are in X^+ but not in X.

The C_p criterion (16) 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 Atkinson and Riani (2008) define the forward value of C_p as

$$C_p(m) = (m - p^+)R_p(m)/R_{p^+}(m) - m + 2p. \tag{18}$$

For each m they calculate $C_p(m)$ for all models of interest. However, some care is needed in interpreting this definition. For each of the models with p parameters, the search may be different, so that the subset $S(m)$ will depend on which model is being fitted. This same subset is used to calculate $R_{p^+}(m)$, so that the estimate s^2 in (15) will also depend on the particular

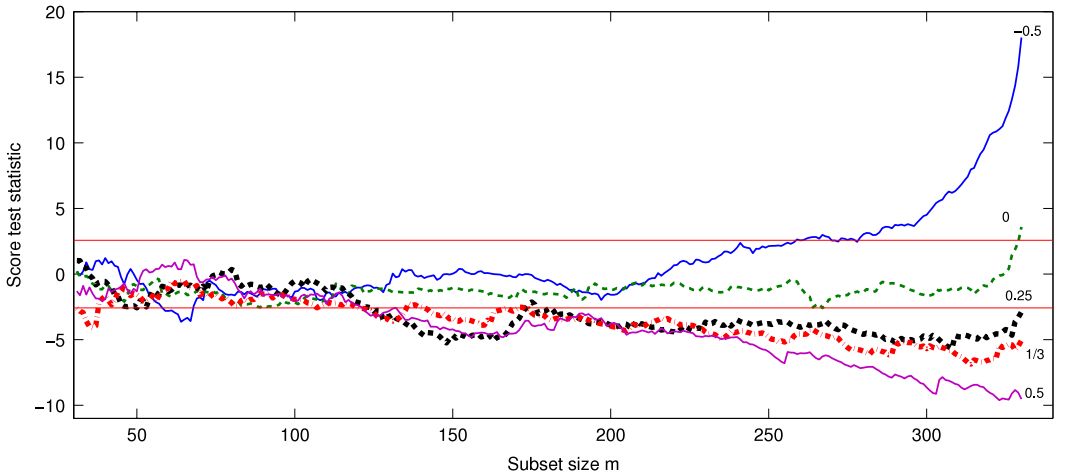

Fig. 1. Ozone data, all variables: fan plot—forward plot of the constructed variable t test for transformations for five values of λ. The logarithmic transformation is preferred except for the last few steps of the search.

model being evaluated as well as on m. Since the searches are for different models, outliers will not necessarily enter in the same order for all models.

It follows from the definition of F in (17) that $C_p(n)$ is the generalization of the t test for inclusion of a single variable to the simultaneous inclusion of $p^+ - p$ variables. In the forward version $C_p(m)$ is calculated from searches including just the p variables of interest. This procedure is therefore the generalization of the added variable t test of Atkinson and Riani (2002) and the same distributional arguments apply. Indeed, mention has already been made of the distributional and simulation results of Riani and Atkinson (in press), which show no divergence of the null distribution of $C_p(m)$ from that in (17) when the degrees of freedom $p^+ - p$, $n - p^+$ for F are replaced by $p^+ - p$, $m - p^+$.

We now exemplify these forward procedures by an analysis of a complex data set of non-trivial size with several explanatory variables which are likely to be highly correlated. The data may contain multiple masked outliers. In addition, since the response variable is non-negative, with a range 1–38, it is likely that it will need to be transformed.

4. The ozone data

The ozone data (Breiman & Friedman, 1985) consist of 366 readings of maximum daily ozone concentration at a hot spot in the Los Angeles (USA) basin and nine meteorological predictor variables such as temperature and humidity. Eliminating one variable with many missing values and a few other cases leaves a data set with 330 complete cases and eight variables.

The data have been much analyzed. Hastie, Tibshirani, and Friedman (2009), Figure 6.9 and p. 336, recommend the cube root of ozone as response. In a series of forward analyses Atkinson and Riani (2000, 2007a, 2008) find evidence for the logarithmic transformation of the response when only the first 80 observations are analysed. They also find strong evidence of a linear trend; hardly surprising since the data start on January 1st.

To extend these forward analyses to all 330 observations we replace the linear trend with one that is triangular, increasing linearly to time 165 and then decreasing, again linearly. This serves as a rough approximation to the possibly underlying sinusoidal variation in this and all other variables. To approximate this behaviour for the explanatory variables x_j we form quadratics $(x_{i,j} - x_{165,j})^2$. These eight transformed variables plus the trend again give a model with nine explanatory variables ($p^+ = 10$). Although more subtle detailed modelling is possible, these variables serve to illustrate the main points connected with the use of the forward search.

We start by exploring an appropriate transformation of the response. Fig. 1 is the fan plot when the full model, that is with all explanatory variables, is fitted to all 330 observations. Because of the large number of observations we take a finer grid of λ values than that suggested in Section 3.4, ranging from -0.5 to 0.5. The plot shows that the log transformation ($\lambda = 0$) is supported by virtually all the data, lying within the bounds of ± 2.58 until the last few observations are included in the subset $S_*(m)$. The plot also shows why slightly larger values of λ might be chosen, although the value of 1/3 suggested by Hastie et al. (2009) is rejected after around half the data have been fitted. The value of 1/4 is just on the 1% level of significance at the end of the search, although rejected earlier.

The indication from the fan plot is that the last few observations are influencing the selected transformation. To determine whether there are any outliers we show in Fig. 2 the forward plot of minimum deletion residuals amongst observations not in the subset, namely the outlier detection statistic (4). For most of the search the trace of observed values fluctuates around the 50% point of the envelopes. In the last few steps the level lies between the 99% and 99.9% simultaneous envelopes, but the rule of Riani et al. (2009), that avoids problems of multiple testing, indicates that there are no outliers for this model and transformation. However, this is not the end of the analysis; the t statistics in Table 1 indicate that apart from the constant

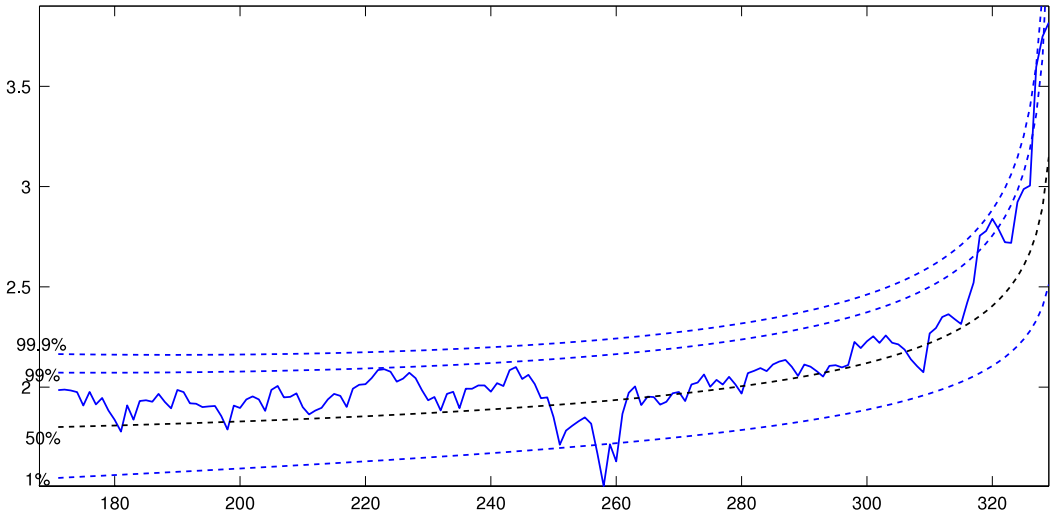

Fig. 2. Logged ozone data, full model: outlier test—forward plot of the minimum deletion residuals of observations not in the subset (4). There is no evidence of any outliers.

Table 1
Logged ozone data: t tests of coefficients in the best models for $p = 4$, 5 and 10.

Term	Value of p		
	10	4	5
Constant	32.34	35.42	35.49
Time	7.65	9.30	9.74
x_1	−2.24		
x_2	−3.75		−3.34
x_3	−4.98	−7.96	−6.57
x_4	1.22		
x_5	−0.39		
x_6	−0.61		
x_7	−2.49	−17.83	−8.80
x_8	−1.42		
R^2 (%)	72.88	71.01	71.98

and the time trend, only variables 1, 2, 3 and 7 have appreciable t statistics. The other four variables are not significant. Because of the correlation between the variables, the t statistics are not necessarily a good guide to significance after some variables are deleted; variable selection is required.

Riani and Atkinson (in press) introduce a "generalized candlestick" plot that contains a summary, for each model, of the information in the trajectory of the forward plots of $C_p(m)$. Since any outliers are expected 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. Here we display results over the last 10% of the search. This value is arbitrary; we have experimented with other percentages for the logged ozone data but found that the conclusions drawn from the plot do not change. For data that contain many outliers, a value larger than 10% might well be appropriate.

Fig. 3 shows the generalized candlestick plot for models in the range $p = 3$ to 7. There is a panel for each value of p, within which the 95% limits of the distribution of $C_p(m)$ are given. All models shown have trajectories that, over the last 10% of the search (that is in the last 33 steps) had one of the five smallest values of $C_p(m)$ for their particular value of p. The vertical lines for each model summarise the values of $C_p(m)$ for each model in the "central" part of the search, that is for $m = n - 33$ through $n - 6$. Individual symbols are reserved for the last six values. 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 $n - 5, \ldots, n - 1$ if these lie outside the box;

Unfilled circle; the final value.

Thus each point in the standard non-robust C_p plot against p is replaced by a single vertical line and several extra symbols.

The general shape of the plot in Fig. 3 is similar to that of the standard C_p plot. For small values of p all models have large values of $C_p(m)$ over the last 33 values of m. Conversely, for large p there are many models with small values of $C_p(m)$ over

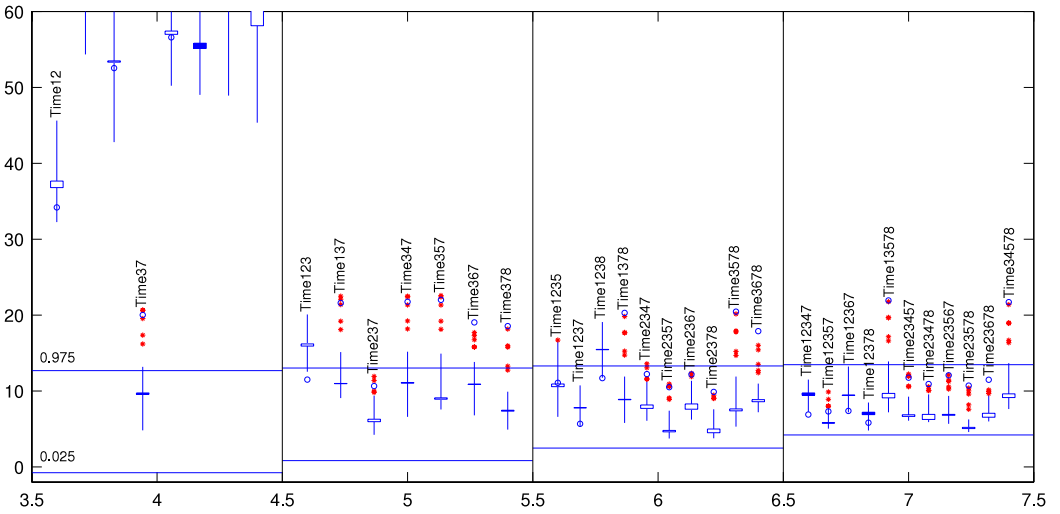

Fig. 3. Logged ozone data: Candlestick plot of values of $C_p(m)$: ∘ the value of $C_p(n)$.

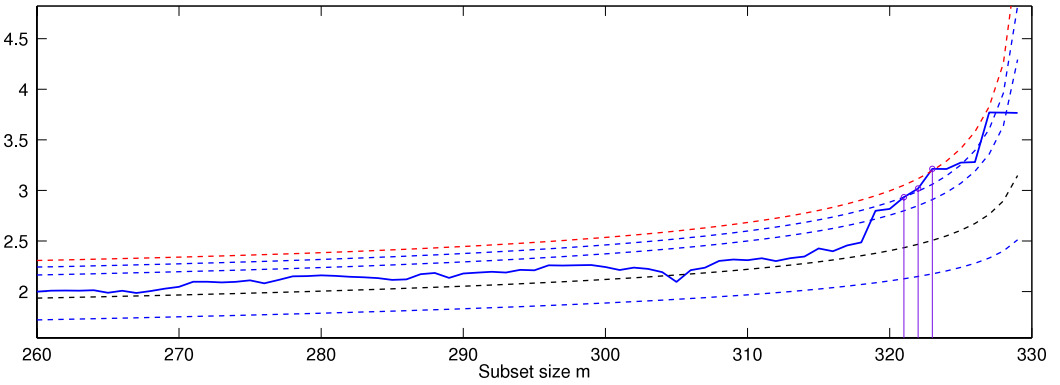

Fig. 4. Logged ozone data: model with trend, x_3 and x_7. Forward plot of the minimum deletion residuals. 1%, 50%, 99%, 99.9% and 99.99% envelopes. The first significant exceedance of the envelopes is signalled at $m = 322$.

most of the range. What is striking is the 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.

The indication of Fig. 3 is that values of p of 4 or 5 should yield a satisfactory model. For $p = 4$ the best model has the trend, x_3 and x_7, although the plot shows the values of $C_p(m)$ increasing towards the end of the search. By far the most stable model for $p = 5$ adds x_2 to these variables. The t statistics for $m = n$ for these two models are given in Table 1. Comparison of the three models in the table shows stable values of the t statistics for the constant and the time trend, and for the value of R^2. The most surprising change is in the t statistic for x_7, which is -17.83 for the model with x_3 and x_7, but drops to -8.80 when x_2 is included. Such behaviour is an indication of high correlation among the explanatory variables.

Accordingly, we monitor the outlier test during the forward search for this model. The resulting Fig. 4 shows that the bounds derived from Riani et al. (2009) start to be violated at $m = 322$. To see whether individual observations have an appreciable effect on the choice of a smaller model we give, in Fig. 5, the forward plot of deletion t statistics. All three variables become steadily more significant as m increases, although that for the time trend does not start to become significant until m is around 200.

It is just possible to consider that the statistic for x_3 becomes less significant at the end of the search. The forward plot of $C_p(m)$ in Fig. 6 does indeed show a significant increase at the end of the search, corresponding to the candlestick plot for this model in Fig. 3; the final 5 values all lie above the upper limit of the distribution of C_p.

We now proceed to establish which are the central and which are the outlying observations. We confirm the outliers by resuperimposing envelopes for a series of tentative values of n, until we find the maximum value for which no outliers are declared. In our example the procedure has provided a signal at $m = 322$ because $r_{\min}(322, 330) >$ the 99.9% envelope, $r_{\min}(321, 330) >$ the 99% envelope and $r_{\min}(323, 330) >$ the 99.9% envelope. Due to masking, the plot returns within the envelopes at the end of the search. This procedure ensures that we overcome simultaneity and so have a test with an overall size close to 1%, even if the existence of outliers manifests itself before, but not at, the end of the search.

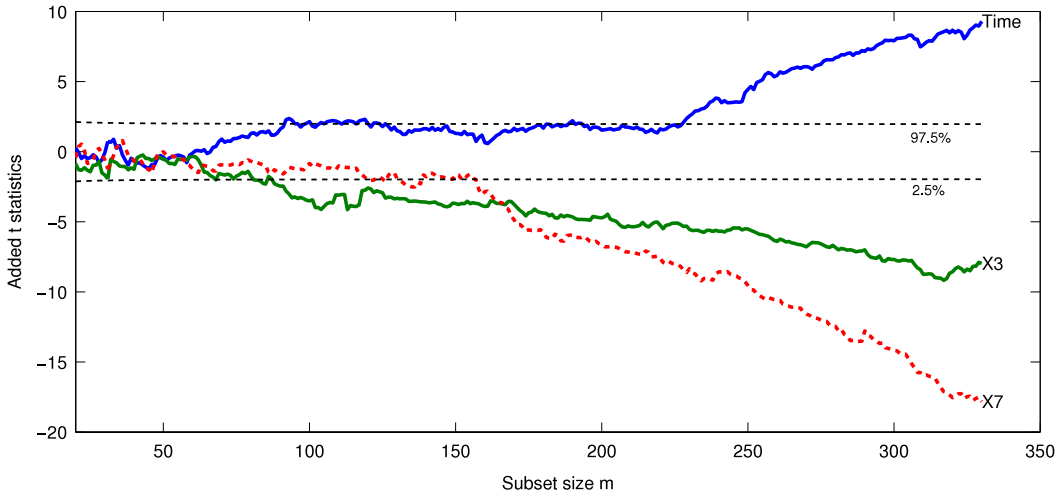

Fig. 5. Logged ozone data: model with trend, x_3 and x_7. Added variable t tests for the explanatory variables.

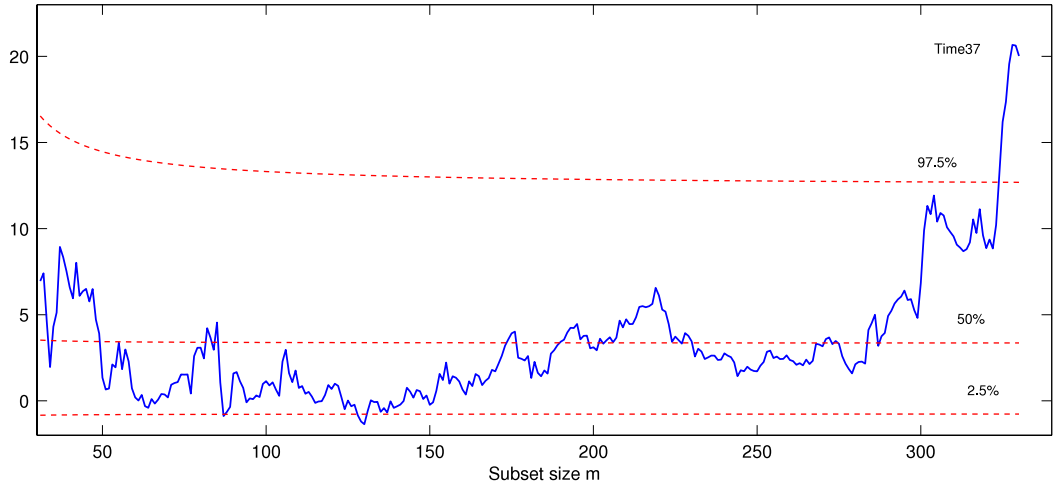

Fig. 6. Logged ozone data: model with trend, x_3 and x_7. Forward plot of $C_p(m)$. The values are too large at the end of the search.

Fig. 7 shows forward plots of $r_{\min}(m, n)$ for several values of n. The envelopes depend on the assumed sample size n whereas the continuous line for the observed values of the statistic is an increasing part of that in Fig. 4 as n increases. In the first panel ($n = 321$) the observed values lie close to the centre of the envelopes throughout the search. As n increases in successive panels to 324, 327 and 328 the end of the curve of observed values comes close to the upper envelope, which is 99.9%. In Fig. 8, for $n = 329$, $r_{\min}(323, 329) >$ the 99.9% envelope and we stop, declaring two outliers. Although the outlier was detected when $m^{\dagger} = 323$, this observation is not outlying when $n = 328$. The two observations that have not entered the search when $m = 328$ are the outliers.

The indication from Fig. 5, the forward plot of added t statistics for the model including x_3 and x_7, is that the properties of this model are not sensitive to the last few observations, even if they are outlying, although the value of $C_p(m)$ in Fig. 4 does become significantly large. The model formed by the inclusion of x_2 is however slightly more sensitive.

The relative stability of the model including x_3 and x_7 contrasts with that when x_2 is also included. Fig. 9 shows the forward plot of added variable t statistics for this model, which can be compared with Fig. 5. The major difference is that the new variable x_2 is significant only at the very end of the search. We explore the effect of deleting the last four observations. These are 286, 327, 205 and 150. When they are deleted the t value for x_2 is -3.07.

It is often instructive to move from such an analysis back to the data. Fig. 10 shows scatterplots of y against x_2, x_3 and x_7. The last four units to enter the subset are highlighted. They are clearly all extreme. The more important question is what inferential effect they have.

We have already seen that deleting these observations decreases the evidence for inclusion of x_2. A further effect is visible in Fig. 11, the fan plot for this model with the last four observations highlighted. They are having an appreciable effect on the evidence against the log transformation. The associated score test statistic when $m = n - 4 = 326$ is equal

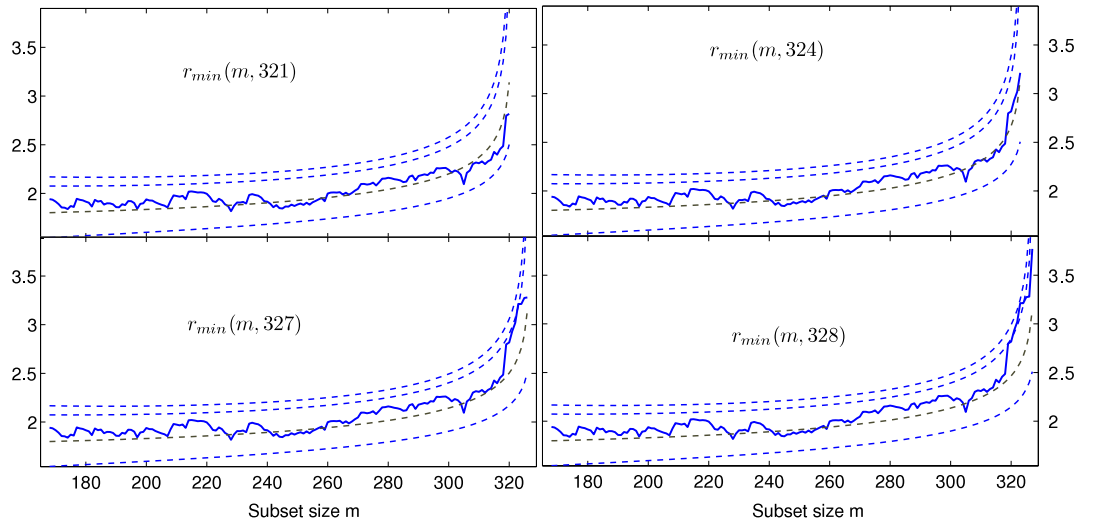

Fig. 7. Logged ozone data: model with trend, x_3 and x_7. Resuperimposition of envelopes. There is no evidence of outliers for tentative $n = 321$, 324, 327 and 328. 1%, 50%, 99% and 99.9% envelopes.

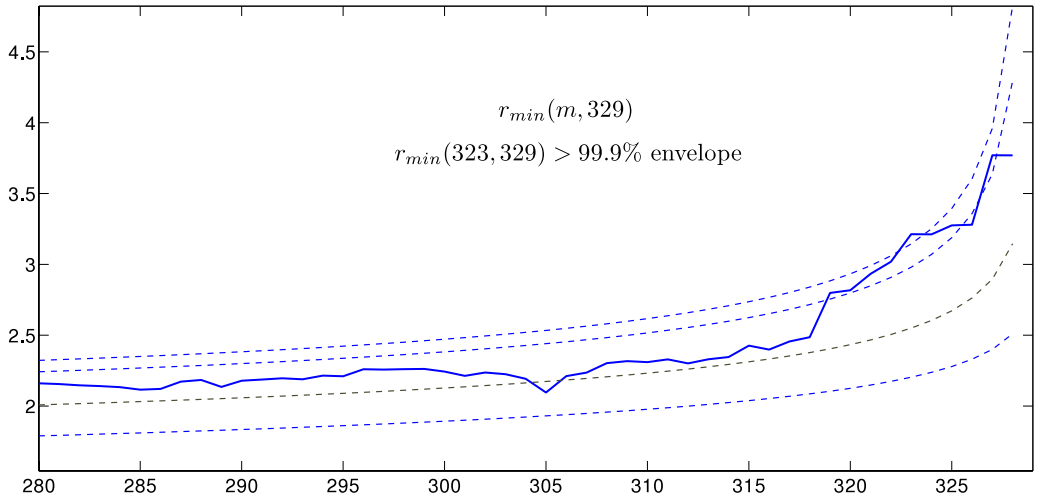

Fig. 8. Logged ozone data: model with trend, x_3 and x_7. Resuperimposition of envelopes for tentative $n = 329$. There is evidence that this sized sample does not follow the distributional assumptions underlying the regression model. 1%, 50%, 99% and 99.9% envelopes.

to 1.99, while when $n = 330$ the value is 4.21. Finally, Fig. 12 shows a forward plot of all scaled residuals using a colour map whose intensity increases as the trajectories depart from zero. The observations which were selected in the fan plot are highlighted. All have negative residuals. For most of the search the observations included in $S_*(m)$ have a symmetrical distribution of residuals between -2 and 2, so that the log transformation is appropriate. The majority of negative residuals start to decrease slightly in magnitude towards the end of the search. An exception is observation 327, whose residual is very close to the bulk of the data when we do a robust fit but tends to increase in absolute value, entering only at the final step of the search. Such monitoring of individual trajectories can lead to the detection of hidden subgroups of units whose trajectories will have a similar pattern.

Our conclusion from this analysis, intended in part to illustrate some of the procedures associated with regression in the forward search, is that a model with logged response, a V-shaped time trend and variables 3 and 7 represents the data well, especially if a few outliers are deleted. On the other hand, the model in which the variables also include x_2 is highly sensitive to deletion of the last four observations.

Our last plot, Fig. 13, is of the full-sample C_p values. No models are shown for $p = 4$. As the candlestick plot (Fig. 3) shows, the increase in the value of C_p in the last five steps of the search makes the value too large for inclusion. For $p = 5$ there are two models with virtually the same values of C_p, those with variables 1, 2, 3 and 2, 3, 7. But Fig. 3 shows that the model including x_1, x_2 and x_3 has too large values of C_p until the last step of the search, the opposite behaviour from the other model. Our forward analysis has enabled us to select a better model and to elucidate the effects of particular observations on this choice.

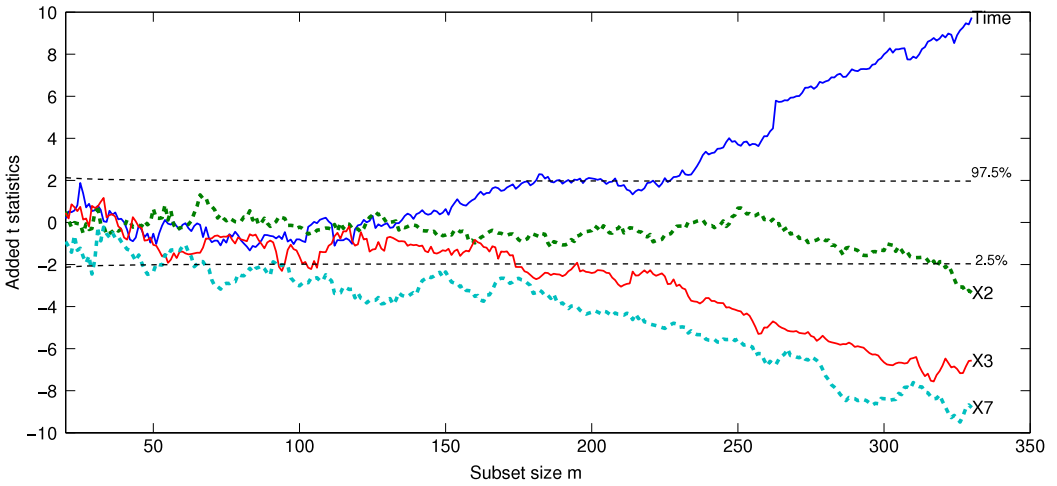

Fig. 9. Logged ozone data: model with trend, x_2, x_3 and x_7. Added variable t tests for the explanatory variables. The inclusion of x_2 is significant only at the end of the search.

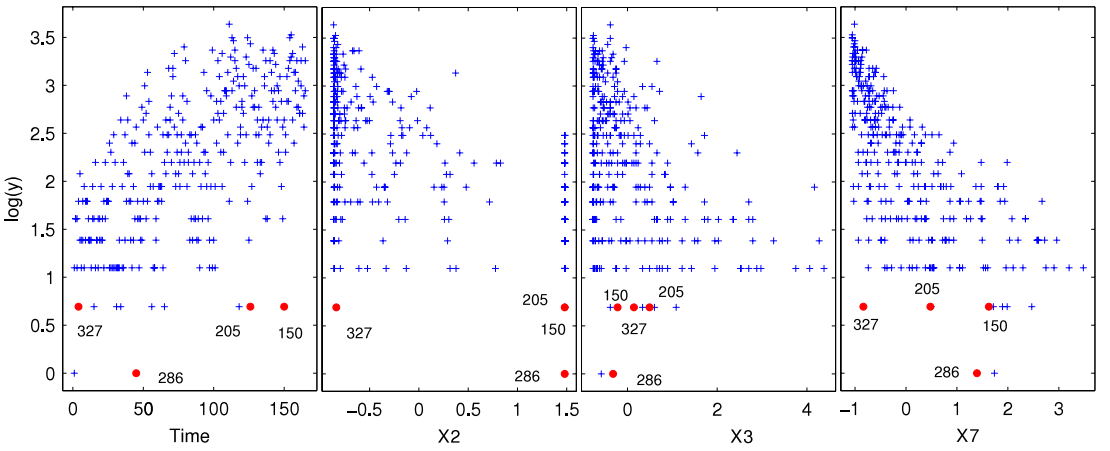

Fig. 10. Logged ozone data: model with trend, x_2, x_3 and x_7. Scatterplots of $\log(y)$: highlight—the last four observations to enter.

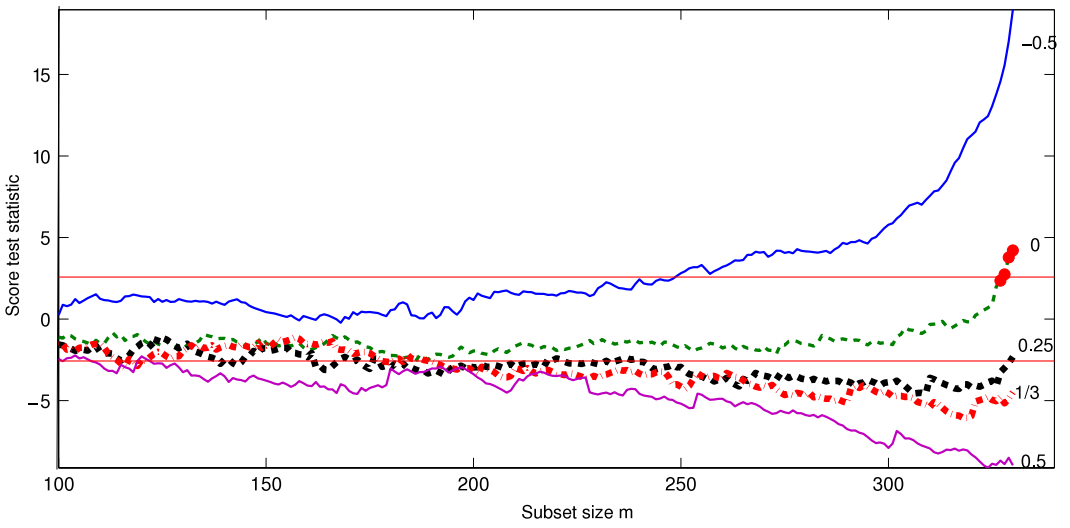

Fig. 11. Ozone data: model with trend, x_2, x_3 and x_7. Fan plot: highlight—the last four observations to enter when $\lambda = 0$.

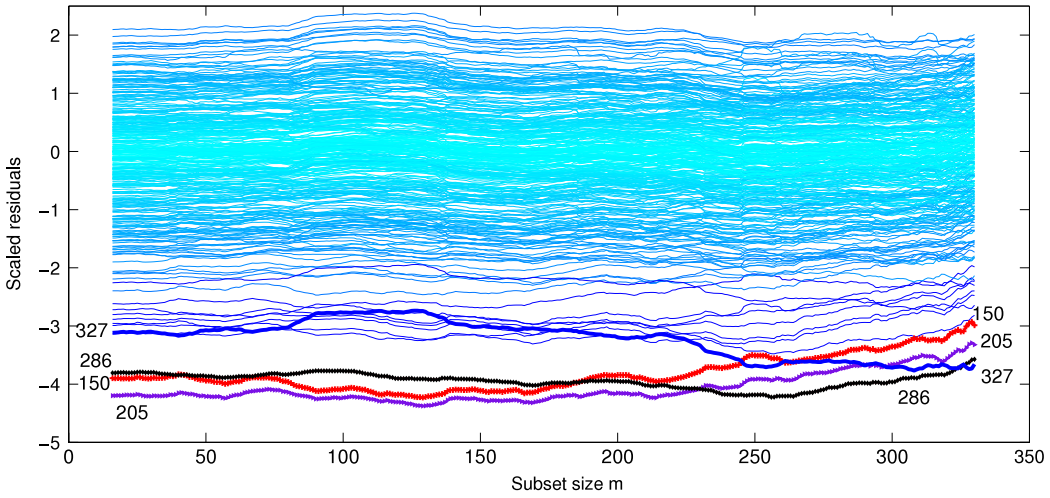

Fig. 12. Logged ozone data: model with trend, x_2, x_3 and x_7. Forward plot of all scaled residuals: highlight—the last four observations to enter.

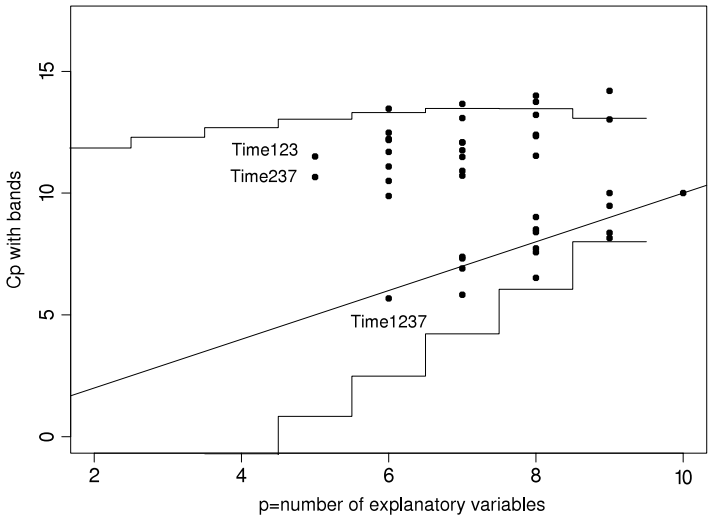

Fig. 13. Logged ozone data. Standard C_p plot. Due to the presence of outliers our preferred model is not considered.

In such a forward analysis of data it is extremely helpful to be able to switch from plot to plot, brushing and linking units between different displays. Examples, for other data sets, are given by Perrotta, Riani, and Torti (2009). The advantages of linking are particularly clear in the online version of the paper where colours are used in the figures, something that is lost on the printed page (as it is also indeed for our paper).

Finally, we compare our results with those of the robust C_p of Ronchetti and Staudte (1994) which uses M-estimation. In a comparative analysis Riani and Atkinson (in press) find that the forward search, combined with using values of $C_p(m)$, gives smaller models than those found by the use of the robust C_p which likewise downweights observations. However, in the ozone example of this paper, the robust methods gives model 3,7 for $p = 4$ and 2, 3, 7 for $p = 5$. This may be related to the clear nature of the outlying observations in Fig. 12. These are not only the last to enter the search but will be those that are downweighted during M-estimation.

A general point that emerges from our analysis is that there was no indication of outliers when we fitted the full model (Fig. 2). However, our procedure did detect outliers in Fig. 4 with the model in which the explanatory variables were x_3 and x_7. An explanation is that the addition of extra explanatory variables can help to accommodate slight outliers, although the parameter estimates for these variables may fluctuate depending on the presence or absence of particular observations. Examples of this effect for models with too many variables are shown in Figures 1 and 3 of Atkinson and Riani (2007a).

An obvious extension of our procedure is to the building and analysis of generalized linear models. Indeed, Atkinson and Riani (2000, Chapter 6) show how the Forward Search can be adapted in this case. However, work remains to be done in establishing distributional results of the kind we have employed in this section for the calibration and interpretation of forward plots.

5. One multivariate sample

The identification of outliers in unstructured multivariate data carries some additional difficulties with respect to the structured regression example that we have examined so far. To summarize, the most relevant issues are:

- Outlyingness should be judged with respect to several, or even many, dimensions simultaneously;
- There is no natural ordering of multivariate data on which "extremeness" of an observation can be ascertained;
- Simple graphical diagnostic tools like the boxplot are difficult to construct in more than one or two dimensions (see Zani, Riani, & Corbellini, 1998).

As we emphasized at the end of the Introduction, a crucial advantage of the forward search over alternative robust approaches (Maronna et al., 2006) is that the main principles are readily extended to any data analysis context, with only minor changes in the implementation reflecting the technicalities of the chosen context. In the case of a sample $y = (y_1, \ldots, y_n)^T$ of v-variate observations from the multinormal distribution $N(\mu, \Sigma)$, the likelihood contribution (1), again omitting constants not depending on i, becomes

$$l_i\{\hat{\mu}_*(m), \hat{\Sigma}_*(m)\} = -\{y_i - \hat{\mu}_*(m)\}^T \hat{\Sigma}_*(m)^{-1}\{y_i - \hat{\mu}_*(m)\}/2, \tag{19}$$

where $\hat{\mu}_*(m)$ and $\hat{\Sigma}_*(m)$ are the estimates of μ and Σ computed from the fitting subset $S_*^{(m)}$ for which the loglikelihood is $L_m(.)$. Comparison of (19) with (2) shows that scaled residuals are now replaced by the squared Mahalanobis distances

$$d_{i*}^2(m) = \{y_i - \hat{\mu}_*(m)\}^T \hat{\Sigma}_*(m)^{-1}\{y_i - \hat{\mu}_*(m)\} \quad i = 1, \ldots, n, \tag{20}$$

which are used for progressing in the search and for detecting multivariate outliers.

The distribution of $d_{i*}^2(n)$ was found by Wilks (1963) to be

$$d_{i*}^2(n) \sim \frac{(n-1)^2}{n}\text{Beta}\left(\frac{v}{2}, \frac{n-v-1}{2}\right)$$

when all the observations come from the prescribed multivariate normal distribution. However, any diagnostic procedure based on $d_{i*}^2(n)$ is prone to masking if there are multiple outliers, as a consequence of the low breakdown of the estimates $\hat{\mu}_*(n)$ and $\hat{\Sigma}_*(n)$.

Reasoning as in Section 3.2, we define

$$i_{\min} = \arg\min_{i \notin S_*^{(m)}} d_{i*}^2(m)$$

to be the observation with the minimum squared Mahalanobis distance among those not in $S_*^{(m)}$. We treat $d_{i_{\min}*}^2(m)$ as a squared deletion distance on $m-1$ degrees of freedom, whose distribution is (Atkinson et al., 2004, pages 43–44)

$$\frac{(m^2-1)v}{m(m-v)}F_{v,m-v}, \tag{21}$$

while $S_*^{(m)}$ remains outlier free. We can then use the order statistics arguments of Section 3.3 to obtain the percentage points of $d_{i_{\min}*}^2(m)$.

Unfortunately, the squared distance (20) is based on $\hat{\Sigma}_*(m)$, which is a biased estimate of Σ, being calculated from the m observations in the subset that have been chosen as having the m smallest distances. In the notation of Section 3.3, the appropriate scaling factor for the envelopes of $d_{i_{\min}*}^2(m)$ from the scaled F distribution (21) can be shown to be

$$\sigma_T(m)^{-1} = \frac{m/n}{P(X_{v+2}^2 < \chi_{v,m/n}^2)}, \tag{22}$$

where $\chi_{v,m/n}^2$ is the m/n quantile of χ_v^2. Plots in Riani, Atkinson, and Cerioli (2007), based on Monte-Carlo simulation, show the superiority of this F approximation over the asymptotic χ_v^2 distribution. The theoretical results of Cerioli (in press) provide further motivation on the adequacy of the F approximation with correction (22), at least in the second half of the search.

Riani et al. (2009) develop a formal statistical test of multivariate outlyingness based on the envelopes

$$V_{m,\alpha}^* = V_{m,\alpha}/\sigma_T(m) \tag{23}$$

with $V_{m,\alpha}$ now the $100\alpha\%$ cut-off point of the $(m+1)$-th order statistic of the scaled F distribution (21). Their procedure uses the two-stage process outlined in Section 3.2.

The bounds defined in the first stage of the procedure are very severe, so that they can be exceeded only when there is strong evidence of contamination. This, in combination with the accurate distributional approximation provided by (23), guarantees good control of the size of the test of no outliers. We argue that such control is an important property of an

outlier detection method. In fact, this was the view behind the seminal work of Wilks (1963). More recent appreciation of this requirement can be found in the techniques developed by Hadi (1994), Becker and Gather (1999), García-Escudero and Gordaliza (2005), Hardin and Rocke (2005) and Cerioli, Riani, and Atkinson (2009).

On the other hand, in the superimposition stage, α is chosen to be the size of an individual testing procedure. This ensures a considerable increase in power, comparable to that obtained by outlier detection methods that do not adjust for multiplicity (Huber, Rousseeuw & Van Aelst, 2008). The only price one has to pay is a small increase in the number of falsely declared outliers, but only in a contaminated framework. A similar approach to multivariate outlier detection using high-breakdown estimators is proposed by Cerioli (in press), while Cerioli and Farcomeni (submitted for publication) describe alternative error rates that may be of interest in this context, and their relationship to masking and swamping effects.

We conclude this section by pointing out that the analysis of multivariate data which is available through the forward search is not confined to multiple outlier detection. For instance, the statistical machinery of Section 3.4 readily extends to multivariate transformations. We refer the reader to Atkinson et al. (2004, Chapters 4–5) for a detailed description of this topic, as well as for the use of the forward search in Principal Component Analysis.

6. Clustering

The distinction between cluster analysis and multivariate outlier identification is somewhat elusive, since a relatively large group of outliers could quite sensibly be considered as a separate subset of observations to be treated as a cluster in their own right. This questionable distinction is even more intriguing when, as is often the case, one has to deal with several contaminating sources at the same time. The forward search has greater potential than other robust techniques to become a comprehensive technique through which cluster analysis and outlier detection could be performed under the same umbrella. The main reason is that the level of trimming associated with the forward search varies from $m_0/n \approx p/n$ to 1, without being restricted to lie in the range $(n/2, 1)$.

Our main tool for cluster analysis is again $d^2_{i_{\min}*}(m)$, the minimum squared Mahalanobis distance for the units not belonging to $S^{(m)}_*$ introduced in Section 5. However, 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. This strategy seemingly requires that we know the clusters, at least approximately, before running the searches. But instead we use many searches with random starting points to provide information on cluster existence and definition. Atkinson, Riani, and Cerioli (2006), Atkinson and Riani (2007b) and Atkinson, Riani, and Cerioli (2008) discuss random starts and argue that one set of simulation envelopes is appropriate whether the membership of $S^{(m_0)}_*$ is robustly or randomly chosen.

One bonus of running many randomly started searches for the purpose of cluster analysis is that they provide reliable information about the true number of groups. Atkinson and Riani (2007b) compare the results provided by the forward search with those obtained through the popular BIC criterion, as implemented in the `mclust` library (Fraley & Raftery, 2003). The conclusion is that BIC for cluster choice tends to overfit the data and to indicate an excessive number of clusters. Furthermore, BIC may be highly sensitive to small changes in the data, while the approach of random start forward searches provides a robust method of establishing cluster numbers and membership. The extension of the statistical test of Riani et al. (2009) to the clustering framework is the subject of ongoing research.

7. The analysis of time series

Clearly all the problems which we have seen so far in terms of model identification and the presence of outliers or of influential observations become more complicated when the observations have a correlation structure. More specifically, the presence of irregularities or structural changes in the observed time series may seriously damage identification and estimation of the suggested ARIMA or structural model (e.g. Chen & Liu, 1993; Harvey & Koopman, 1992). The correlation structure of the time series for each subset size during the forward search can be retained using the Kalman filter and treating as missing the observations not belonging to the subset (Riani, 2004). Further, least squares residuals can be replaced by one-step-ahead standardized prediction residuals. The initial subset can be chosen among several blocks of contiguous observations of fixed dimension in order to retain the same dependence structure as in the original data set (Cerioli & Riani, 2002; Haegerty & Lumley, 2000). Score tests using the added variables of de Jong and Penzer (1998) can be used to check for shocks, such as outliers, level shifts and switches. Grossi and Laurini (2009) extended the procedure to the financial time series which are likely to contain ARCH effects. In addition, the score test for transformations mentioned in Section 3.4 can be applied to time series, adding an extra recursion in the transition equation of the state–space formulation of the Kalman filter (Riani, 2009). This machinery can be used not only for outlier detection but also for robust seasonal adjustment under transformations (Proietti & Riani, 2009). However, many things remain to be developed in the application of the Forward Search in the time series context, such as:

- The extension of the automatic construction of the envelopes based on order statistics to the one step ahead prediction residuals;
- The implementation of a robust model selection procedure;
- The construction of an algorithm which can automatically distinguish among the different types of outliers and level shifts.

8. Integrated analysis

The model of statistical behaviour implicit so far has been of unpressured statistical analysis of single sets of data. We conclude with a brief discussion of problems that can arise in the analysis of anti-fraud data (Deng, Joseph, Sudjianto, & Wu, 2009; Riani, Cerioli, Atkinson, Perrotta, & Torti, 2008). Briefly these are that there are many interlinked large data sets which may contain a variety of informative deviations from simple models. Further, any fraudulent deviations need to be quickly detected, so that the related illegal activities can be promptly prevented.

The statistical models may be either simple regression or the time series models of Section 7. As an example from the monitoring of international trade, an unexpected strong decrease in trade between two countries of a product can be coupled with a sudden increase in trade between two other countries or in trade in a similar product in order to avoid taxes or import duties. More generally, the presence of irregularities in a multitude of datasets can give rise to signals or alerts which may indicate possible instances of a wide range of fraud control problems. These include stockpiling before EU enlargements, fraud in payment of export refunds, deflection of trade to evade import duties or quotas in force for imports into the EU, trade based money laundering etc. Obviously not all abnormal prices are associated with fraudulent transactions. For example, it is possible to find errors in recording, although their frequency and distribution are not known.

Clearly the need to analyze thousands of data sets involves a series of additional problems. These include:

- Calibration for simultaneity in order to keep false alarms under control;
- The requirement to rank outliers and other irregularities by their financial importance, not just by disagreement with a statistical model. If the data have low variability many unimportant outliers will seem important;
- The presence of different populations within a single dataset;
- The need to enhance the Forward Search through the development of interactive dynamic plots with user-friendly graphical interfaces.

The problem of distinguishing between the presence of multiple outliers or distinct models for subsets of the data has strong implications in the context of international trade data, such as that monitored by the Joint Research Centre of the European Community at Ispra (Italy). Given that the relationship between value and quantity of a traded good is linear, trade data should not require transformation. Any indicated transformation may be due to the presence of multiple populations: that is a combination of fair trade declarations together with price under-declarations to reduce import duties and over-declarations to carry out money laundering activities. Perrotta et al. (2009) give an example of the imports of a fishery product into the EU where the inclusion of the last observations to enter the Forward Search causes strong rejection of the hypothesis of no transformation. This set of units forms a well defined cluster below the main cloud in the scatterplot of value against quantity and is associated with a different population whose import prices are much lower than those of the bulk of the data.

We are currently exploring the detection of multiple groups in regression data by the extension of the method of random starts described for the clustering of multivariate data in Section 6. This procedure, which does not force all units to be classified, has connections with the procedure of García-Escudero, Gordaliza, San Martin, Van Aelst, and Zamar (2009) which is based on the extension of least trimmed squares to linear clustering.

The final goal is to have an automatic robust monitoring system which can classify data structures and also focus attention on individual cases among tens of thousands of similar situations.

References

Andrews, D. F., Bickel, P. J., Hampel, F. R., Tukey, W. J., & Huber, P. J. (1972). *Robust estimates of location: survey and advances*. Princeton, NJ: Princeton University Press.
Atkinson, A. C. (1973). Testing transformations to normality. *Journal of the Royal Statistical Society, Series B*, *35*, 473–479.
Atkinson, A. C. (1985). *Plots, transformations, and regression*. Oxford: Oxford University Press.
Atkinson, A. C. (1994). Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association*, *89*, 1329–1339.
Atkinson, A. C. (2009). Econometric applications of the forward search in regression: robustness, diagnostics and graphics. *Econometric Reviews*, *28*, 21–39.
Atkinson, A. C., & Riani, M. (2000). *Robust diagnostic regression analysis*. New York: Springer-Verlag.
Atkinson, A. C., & Riani, M. (2002). Tests in the fan plot for robust, diagnostic transformations in regression. *Chemometrics and Intelligent Laboratory Systems*, *60*, 87–100.
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. (2007a). Building regression models with the forward search. *Journal of Computing and Information Technology—CIT*, *15*, 287–294. doi:10.2489/cit.1001135.
Atkinson, A. C., & Riani, M. (2007b). 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., & Riani, M. (2008). A robust and diagnostic information criterion for selecting regression models. *Journal of the Japanese Statistical Society*, *38*, 3–14.
Atkinson, A. C., Riani, M., & Cerioli, A. (2004). *Exploring multivariate data with the forward search*. New York: Springer-Verlag.
Atkinson, A. C., Riani, M., & Cerioli, A. (2006). Random start forward searches with envelopes for detecting clusters in multivariate data. In S. Zani, A. Cerioli, M. Riani, & M. Vichi (Eds.), *Data analysis, classification and the forward search* (pp. 163–171). Berlin: Springer-Verlag.
Atkinson, A. C., Riani, M., & Cerioli, A. (2008). Monitoring random start forward searches for multivariate data. In P. Brito (Ed.)., *COMPSTAT 2008* (pp. 447–458). Heidelberg: Physica-Verlag.
Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). New York: Wiley.
Becker, C., & Gather, U. (1999). The masking breakdown point of multivariate outlier identification rules. *Journal of the American Statistical Association*, *94*, 947–955.

Beckman, R. J., & Cook, R. D. (1983). Outlier.....s (with discussion). *Technometrics*, *25*, 119–163.
Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika*, *40*, 318–335.
Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B*, *26*, 211–246.
Box, G. E. P., & Watson, G. S. (1962). Robustness to non-normality of regression tests. *Biometrika*, *49*, 93–106.
Breiman, L., & Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and transformation (with discussion). *Journal of the American Statistical Association*, *80*, 580–619.
Casella, G., & Berger, R. L. (2002). *Statstical inference* (2nd ed.). Pacific Grove: Duxbury.
Cerioli, A. (2010). Multivariate outlier detection with high-breakdown estimators. *Journal of the American Statistical Association* (in press).
Cerioli, A., & Farcomeni, A. (2010). Error rates for multivariate outlier detection. Unpublished manuscript (submitted for publication).
Cerioli, A., & Riani, M. (2002). Robust methods for the analysis of spatially autocorrelated data. *Statistical Methods and Applications—Journal of the Italian Statistical Society*, *11*, 335–358.
Cerioli, A., Riani, M., & Atkinson, A. C. (2009). Controlling the size of multivariate outlier tests with the MCD estimator of scatter. *Statistics and Computing*, *19*, 341–353.
Chen, C., & Liu, L.-M. (1993). Joint estimation of model parameters and outlier effects in time series. *Journal of the American Statistical Association*, *88*, 284–297.
Cheng, T.-C., & Biswas, A. (2008). Maximum trimmed likelihood estimator for multivariate mixed continuous and categorical data. *Computational Statistics and Data Analysis*, *52*, 2042–2065.
Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. London: Chapman and Hall.
Crosilla, F., & Visentini, D. F. S. (2007). An automatic classification and robust segmentation procedure of spatial objects. *Statistical Methods and Applications*, *15*, 329–341.
de Jong, P., & Penzer, J. (1998). Diagnosing shocks in time series. *Journal of the American Statistical Association*, *93*, 796–806.
Deng, D., Joseph, V. R., Sudjianto, A., & Wu, C. F. J. (2009). Active learning through sequential design, with applications to detection of money laundering. *Journal of the American Statistical Association*, *104*, 969–981.
Forbes, J. D. (1857). Further experiments and remarks on the measurement of heights by the boiling point of water. *Transactions of the Royal Society of Edinburgh*, *21*, 235–243.
Fraley, C., & Raftery, A. E. (2003). Enhanced model-based clustering, density estimation and discriminant analysis: MCLUST. *Journal of Classification*, *20*, 263–286.
García-Escudero, L. A., & Gordaliza, A. (2005). Generalized radius processes for elliptically contoured distributions. *Journal of the American Statistical Association*, *100*, 1036–1045.
García-Escudero, L. A., Gordaliza, A., San Martin, R., Van Aelst, S., & Zamar, R. (2009). Robust linear clustering. *Journal of the Royal Statistical Society, Series B*, *71*, 301–308.
Gilmour, S. G. (1996). The interpretation of Mallows's C_p-statistic. *The Statistician*, *45*, 49–56.
Grossi, L., & Laurini, F. (2009). A robust forward weighted Lagrange multiplier test for conditional heteroscedasticity. *Computational Statistics and Data Analysis*, *53*, 2251–2263.
Guenther, W. C. (1977). An easy method for obtaining percentage points of order statistics. *Technometrics*, *19*, 319–321.
Hadi, A. S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society, Series B*, *54*, 761–771.
Hadi, A. S. (1994). A modification of a method for the detection of outliers in multivariate samples. *Journal of the Royal Statistical Society, Series B*, *56*, 393–396.
Hadi, A. S., Imon, A. H. M. R., & Werner, M. (2009). Detection of outliers. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*, 57–70.
Hadi, A. S., & Simonoff, J. S. (1993). Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association*, *88*, 1264–1272.
Haegerty, P., & Lumley, T. (2000). Window subsampling of estimating functions with application to regression models. *Journal of the American Statistical Association*, *95*, 197–211.
Hampel, F. R. (1975). Beyond location parameters: Robust concepts and methods. *Bulletin of the International Statistical Institute*, *46*, 375–382.
Hampel, F., Ronchetti, E. M., Rousseeuw, P., & Stahel, W. A. (1986). *Robust statistics*. New York: Wiley.
Hardin, J., & Rocke, D. M. (2005). The distribution of robust distances. *Journal of Computational and Graphical Statistics*, *14*, 910–927.
Harvey, A. C., & Koopman, S. J. (1992). Diagnostic checking of unobserved components time series models. *Journal of Business and Economic Statistics*, *10*, 377–389.
Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning. Data mining, inference and prediction* (2nd ed.). New York: Springer.
Hawkins, D. M. (1980). *Identification of outliers*. London: Chapman and Hall.
Hawkins, D. M. (1983). Discussion of paper by Beckman and Cook. *Technometrics*, *25*, 155–156.
Huber, P. J. (1981). *Robust statistics*. New York: Wiley.
Huber, P. J., & Ronchetti, E. M. (2009). *Robust statistics* (2nd ed.). New York: Wiley.
Hubert, M., Rousseeuw, P. J., & Van Aelst, S. (2008). High-breakdown robust multivariate methods. *Statistical Science*, *23*, 92–119.
Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions—1* (2nd ed.). New York: Wiley.
Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, *15*, 661–675.
Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust statistics: Theory and methods*. Chichester: Wiley.
Mavridis, D., & Moustaki, I. (2010). The forward search algorithm for detecting aberrant response patterns in factor analysis for binary data. *Journal of Computational and Graphical Statistics*, *18*, 1016–1034.
Morgenthaler, S. (2007). A survey of robust statistics. *Statistical Methods and Applications*, *15*, 271–293; *Statistical Methods and Applications, 16*, 171–172 (erratum).
Müller, C., & Neykov, N. (2003). Breakdown points of the trimmed likelihood and related estimators in GLMs. *Journal of Statistical Planning and Inference*, *116*, 503–519.
Perrotta, D., Riani, M., & Torti, F. (2009). New robust dynamic plots for regression mixture detection. *Advances in Data Analysis and Classification*, *3*, 263–279. doi:10.1007/s11634-009-0050-y.
Proietti, T., & Riani, M. (2009). Seasonal adjustment and transformations. *Journal of Time Series Analysis*, *30*, 47–69.
Riani, M. (2004). Extensions of the forward search to time series. *Studies in Nonlinear Dynamics and Econometrics*, *8*, 1–23.
Riani, M. (2009). Robust transformations in univariate and multivariate time series. *Econometric Reviews*, *28*, 262–278.
Riani, M., & Atkinson, A. C. (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., & Atkinson, A. C. (2010). Robust model selection with flexible trimming. Computational Statistics and Data Analysis (in press).
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., Atkinson, A. C., & Cerioli, A. (2009). Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society, Series B*, *71*, 447–466.
Riani, M., Cerioli, A., Atkinson, A., Perrotta, D., & Torti, F. (2008). Fitting mixtures of regression lines with the forward search. In F. Fogelman-Soulié, D. Perrotta, J. Piskorski, & R. Steinberger (Eds.), *Mining massive data sets for security* (pp. 271–286). Amsterdam: IOS Press.
Ronchetti, E., & Staudte, R. G. (1994). A robust version of Mallows's C_p. *Journal of the American Statistical Association*, *89*, 550–559.
Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, *79*, 871–880.
Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York: Wiley.
Solaro, N., & Pagani, M. (2010). The forward search for classical multidimensional scaling when the starting data matrix is known. In C. Lauro, F. Palumbo, & M. Greenacre (Eds.), *Data analysis and classification: From the exploratory to the confirmatory approach* (pp. 101–109). Berlin: Springer-Verlag.

Tallis, G. M. (1963). Elliptical and radial truncation in normal samples. *Annals of Mathematical Statistics*, *34*, 940–944.
Torti, F., & Perrotta, D. (2010). Size and power of tests for regression outliers in the forward search. In: Ingrassia, S., Rocci, R., Vichi, M. (Eds.), New perspectives in statistical modeling and data analysis. Springer-Verlag, Heidelberg (in press).
Weisberg, S. (2005). *Applied linear regression* (3rd ed.). New York: Wiley.
Wilks, S. S. (1963). Multivariate statistical outliers. *Sankhya A*, *25*, 407–426.
Wisnowski, J. W., Montgomery, D. C., & Simpson, J. R. (2001). A comparative analysis of multiple outlier detection procedures in the linear regression model. *Computational Statistics and Data Analysis*, *36*, 351–382.
Zani, S., Riani, M., & Corbellini, A. (1998). Robust bivariate boxplots and multiple outlier detection. *Computational Statistics and Data Analysis*, *28*, 257–270.