The Forward Search

Anthony C. Atkinson *
The London School of Economics, London WC2A 2AE, UK

March 20, 2002

Abstract

The paper is based on an introductory talk to a session on the forward search at Compstat 2002 in Berlin. Four examples are considered: multiple regression, response transformation, generalized linear models with gamma errors and multivariate data, exemplified by measurements on Swiss banknotes. The method reveals the effect of multiple outliers on inferences about suitable models for the data. References are given to more extensive treatments in joint work with Marco Riani.

Keywords: Added-variable t test, Box-Cox transformation, fan plot, forward search, generalized linear model, masked outliers, multivariate data, robustness

1 Introduction

This paper summarises joint research with Marco Riani on the forward search, a powerful general method for detecting multiple masked outliers and for determining their effect on models fitted to the data. Atkinson and Riani (2000) describe its use in linear and nonlinear regression, response transformation and in generalized linear models. These examples are here extended to include multivariate analysis. Riani and Atkinson (2001) describe an application to multivariate transformations and discriminant analysis.

The regression example is in §2. Section 3 is concerned with the effect of outliers on the Box-Cox transformation of the response in regression. The models considered are extended in §4 to include generalized linear models. Finally, in §5, an example is given of six measurements on Swiss bank notes. The observations are seen to fall into two groups, with some further structure.

*a.c.atkinson@lse.ac.uk

2 Regression and the Surgical Unit Data

Let the standard linear regression model be written as

$$y = Q\theta + \epsilon = X\beta + w\gamma + \epsilon, \tag{1}$$

where the errors ϵ are independent and identically distributed with constant variance σ^2. The vector of p parameters θ is estimated by least squares applied to a subset of m observations to give the estimate $\hat{\beta}(m)$. We start the search with m small, usually p or perhaps $p+1$, and randomly select 1,000 subsamples. The initial subset $S^*(p)$ provides the least median of squares estimator $\hat{\beta}(p^*)$, that is it minimises the median squared residual (Rousseeuw 1984). We then order the residuals and augment the subset.

When m observations are used in fitting, the optimum subset $S^*(m)$ yields n residuals $e(m^*)$. We order the squared residuals $e^2(m^*)$ and take the observations corresponding to the $m+1$ smallest as the new subset $S^*(m+1)$. Usually this process augments the subset by one observation, but sometimes two or more observations enter as one or more leave. Due to the form of the search, outliers, if any, tend to enter as m approaches n.

In our example we look at a forward plot of the residuals $e(m^*)$, scaled by the final estimate of σ. We also look at forward plots of t tests from (1) where we, in turn, take each of the columns of Q as the vector w (except the column corresponding to the constant term in the model). The well-established approach of added variables (for example §2.2 of Atkinson and Riani 2000) leads to an expression for the least squares estimate $\hat{\gamma}$ as a function of residuals from the regression of y and w on X. The added variable t test for w is

$$t_\gamma = \hat{\gamma}/\{s_w^2/(w^T A w)\}^{1/2}, \tag{2}$$

where $A = I - H = I - X(X^T X)^{-1} X^T$ and s_w^2 is the residual mean square estimate of σ^2.

As an example we use the **Surgical Unit Data**, (Neter, Kutner, Nachtsheim, and Wasserman 1996, pp.334 & 438) with 108 observations on the time of survival of patients who had liver surgery. There are four explanatory variables. To illustrate the detection of masked outliers we changed the values of 12 observations. The details are in Atkinson and Riani (2002a).

The effect of the outliers is clear. The upper panel of Figure 1 shows the forward plot of the standardised residuals. For most of the search the 12 outliers are clearly separated from the rest of the data. However, towards the end of the search, these outliers begin to be included in the subset of m observations used in fitting and become less apparent and so increasingly hard to detect.

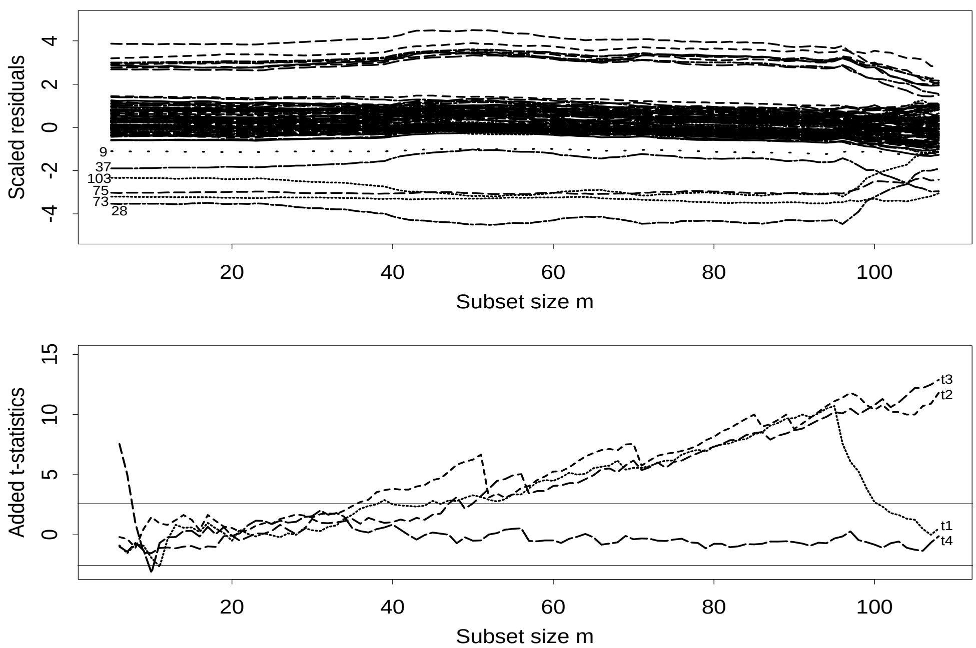

Figure 1: Modified Surgical Unit data: upper panel, forward plot of scaled residuals; lower panel, added variable t test for each explanatory variable

The t statistics for the variables are plotted in the lower panel of Figure 1. The effect of the modification has been to make variable one non-significant: in the unmodified data it is the most important variable. The plot very clearly shows the effect of the outliers on the t tests for regression. Variable selection using t tests on all 108 observations would lead to the incorrect dropping of variable one. It is clear that a subset of observations are indicating a different model from the majority of the data. Which these observations are follows from the order in which the observations enter the search.

The plot of residuals in the upper panel of Figure 1 come from a single search fitting the model $E(Y) = A\theta$. The lower panel of the figure shows the results of four searches, one for each of the submodels $E(Y) = X\beta$. Exclusion from the model of the variable w which is to be tested ensures that the null distribution of t_γ is indeed Student's t (Atkinson and Riani 2002a).

3 Transformations of the Univariate Response in Regression and the Doubly Modified Poison Data

We consider tests of the value of the transformation parameter λ in the Box and Cox (1964) family of normalized power transformations

$$z(\lambda) = \begin{cases} (y^\lambda - 1)/\lambda \dot{y}^{\lambda-1} & \lambda \neq 0 \\ \dot{y} \log y & \lambda = 0, \end{cases} \tag{3}$$

where the geometric mean of the observations is written as $\dot{y} = \exp(\Sigma \log y_i/n)$. The model is multiple regression with the response in (1) replaced by $z(\lambda)$. An approximate score test can be found by Taylor series expansion of $z(\lambda)$ about the null value λ_0 which adds an extra, constructed, variable to the regression model. To form an approximate t test for the significance of this variable, and so of the need for transformation, the variable w in (1) is replaced by the constructed variable which, provided X includes a constant, can be written

$$w(\lambda) = \begin{cases} y^\lambda \{\log(y/\dot{y}) - 1/\lambda\}/(\lambda \dot{y}^{\lambda-1}) & \lambda \neq 0 \\ \dot{y} \log y (0.5 \log y - \log \dot{y}) & \lambda = 0. \end{cases} \tag{4}$$

We combine calculation of the test statistic with the forward search. Since observations which are outlying on one scale may not be outlying for a different transformation, we conduct several searches for different values of λ. In most applications, including the example here, we use five searches for the values $\lambda = -1$, -0.5, 0, 0.5 and 1. If there are outliers for a particular λ they will enter the search last and influence the value of the test statistic.

As an example we use the **Poison Data** from Box and Cox (1964) in which two observations have been modified. For all the data the score statistic has a value of 0.64 when $\lambda = 0$ and the log transformation seems appropriate. The adjacent values of -0.5 and 0.5 are firmly rejected. However forward plots of the five t statistics, combined in a “fan” plot, reveal the effect of the two changed observations.

The effect of the two outliers is clearly seen in Figure 2. The effect is greatest for $\lambda = -1$, where addition of the two outliers at the end of the search causes the statistic to jump from an acceptable 1.08 to 10.11. The effect is similar, although smaller, for $\lambda = -0.5$. At the end of the search the log transformation is acceptable, but the forward plot shows that this conclusion is determined solely by the two outliers.

Although forward plots of the test statistic are easily interpreted, the statistic cannot have exactly a t distribution since the constructed variable

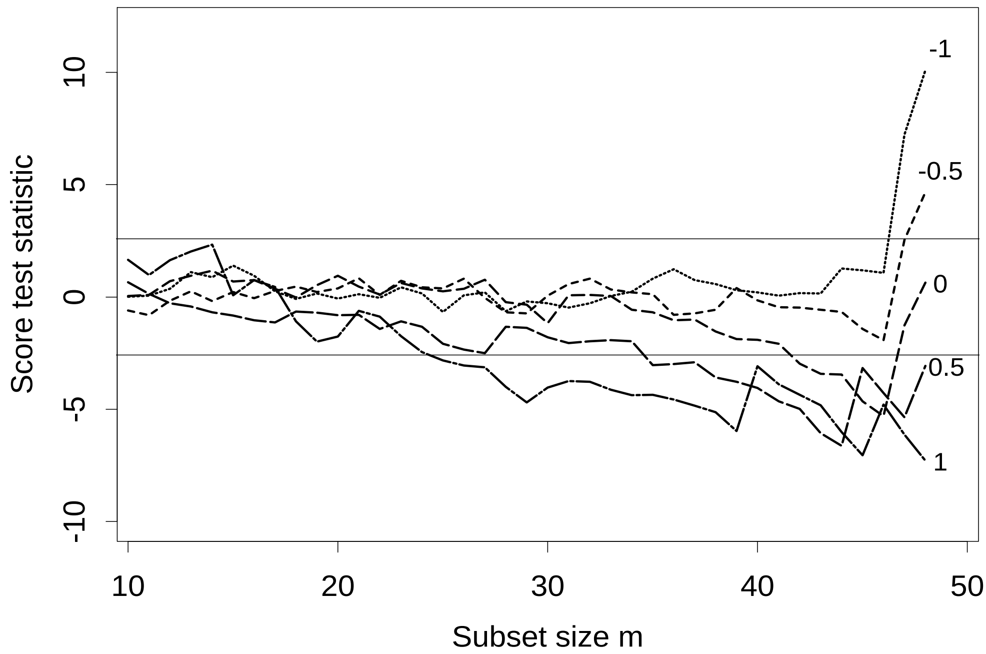

Figure 2: Doubly modified poison data: fan plot – forward plot of $T_p(\lambda)$ for five values of λ. The curve for $\lambda = -1$ is uppermost; the effect of the two outliers is clear

(4) is a function of the response. The simulations reported in Atkinson and Riani (2002b) indicate that the distribution is close to the t distribution of §2 if there is "strong" regression on X - values of R^2 around 80% or higher are necessary. A theoretical justification for these results is given by Atkinson and Riani (2002a). A fuller analysis of the doubly modified Poison Data is in Atkinson and Riani (2000, §4.6).

4 Generalized Linear Models and Dielectric Breakdown Strength

The structure provided by the theory of generalized linear models allows us to apply the forward search to, particularly, gamma, Poisson and binomial data in a manner analogous to that used for regression. Chapter 6 of Atkinson and Riani (2000) contains theory and examples.

In generalized linear models we have a response y, a vector of linear predictors with elements $\eta = x^T\beta$ and a link function $g(\mu) = \eta$ connecting the two. As well as the customary problems about individual outliers and

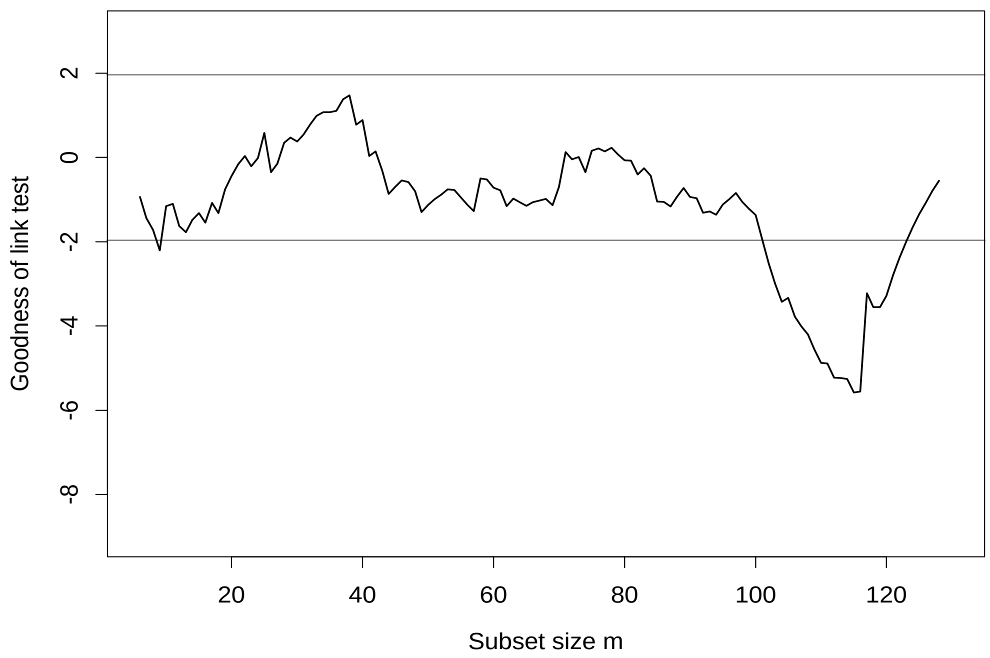

Figure 3: Dielectric data, Box–Cox link with $\lambda = 2$: forward plot of goodness of link test

the correct form of the linear model, there is also a need to specify the correct form of link function, a need which is particulary acute with gamma data. One approach is to use a goodness of link test, which comes from a Taylor series expansion of the link function (for example, Atkinson and Riani 2000, §6.6.4). This results in a similar t test to those of the two previous sections, but now with constructed variable $\hat{\eta}^2$, where $\hat{\eta}$ is the fitted linear predictor from the model to be tested.

Atkinson and Riani (2000), §6.9, analyse data from Nelson (1981) on performance degradation from accelerated tests. The response, **dielectric breakdown strength**, is measured at the points of an 8×4 factorial with four observations per cell. The response is nonnegative with a skew distribution. The gamma distribution is appropriate.

The analysis starts by fitting a predictor with linear terms in temperature and log(time) and exploring the family of power (Box-Cox) links. Analysis of all 128 observations suggests a link with $\lambda = 2$. However the forward plot of the goodness of link statistic in Figure 3 reveals that, although the value of the statistic may be acceptable for all the data, it is not so earlier on, having a maximum absolute value of 5.58 when $m = 115$. The forward plot of deviance residuals, Figure 4, shows that around this extreme value there

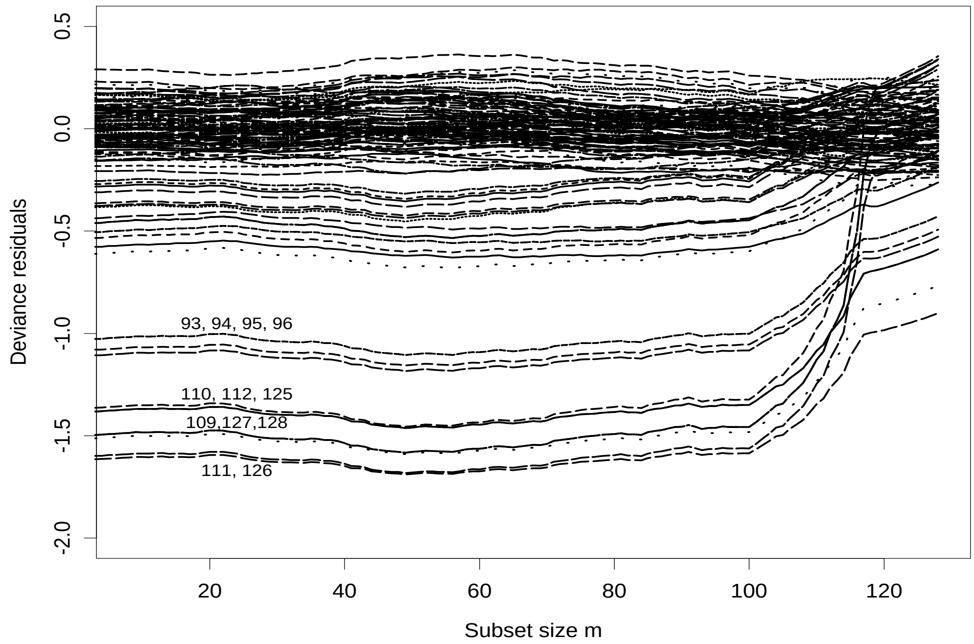

Figure 4: Dielectric data, Box–Cox link with $\lambda = 2$: forward plot of deviance residuals

is a rapid change in the values of some residuals. The figure also reveals that there are groups of residuals that form clusters. Much of the subsequent analysis is concerned with finding a structure for the linear predictor that accommodates these groups of observations from individual cells of the factorial arrangement.

Here the forward search revealed the groups of observations that do not agree with that fitted to the rest of the data. The search also shows the effect of these observations on the proposed link.

5 Multivariate Observations and the Swiss Banknote Data

With multivariate observations we replace the squared residuals $e_i^2(m^*)$ with the squared Mahalanobis distances

$$d_i^2(m^*) = \{y_i - \hat{\mu}(m^*)\}^T \hat{\Sigma}^{-1}(m^*)\{y_i - \hat{\mu}(m^*)\}, \tag{5}$$

where $\hat{\mu}(m^*)$ and $\hat{\Sigma}(m^*)$ are estimates of the mean and covariance matrix of the observations based on the subset $S^*(m)$. These distances are used

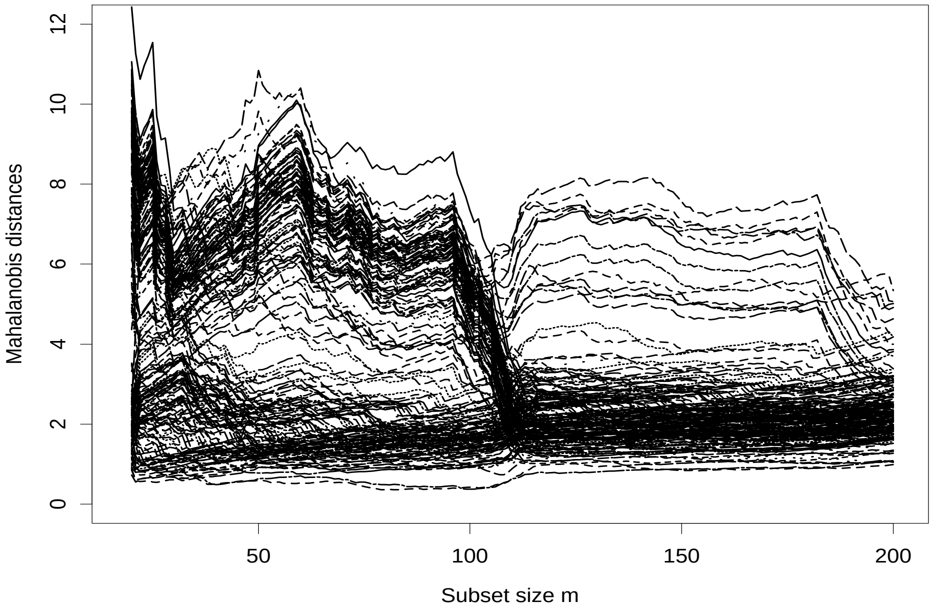

Figure 5: Swiss Banknote Data, forward plot of Mahalanobis distances for both groups

for ordering the observations and for determining how we move forward in the search. We use the robust bivariate boxplots of Zani, Riani, and Corbellini (1998) to determine an initial subset which is not outlying in any two-dimensional plot of the data. The content of the contours is adjusted to give an initial subset of the required size.

The data are readings on six dimensions of 200 Swiss bank notes, 100 of which may be genuine and 100 forged. All notes have been withdrawn from circulation, so some of the notes in either group may have been misclassified. Also, the forged notes may not form a homogeneous group. For example, there may be more than one forger at work. The data, and a reproduction of the bank note, are given by Flury and Riedwyl (1988, p.4-8).

Figure 5 shows the scaled Mahalanobis distances from a forward search starting with 20 observations on notes believed genuine. In the first part of the search, up to $m = 93$, the observations seem to fall into two groups. One has small distances and is composed of observations within or shortly to join the subset. Above these there are some outliers and then, higher still, a concentrated band of outliers, all of which are behaving similarly. The two groups are apparent.

The structure of the group of forgeries is also readily revealed by the

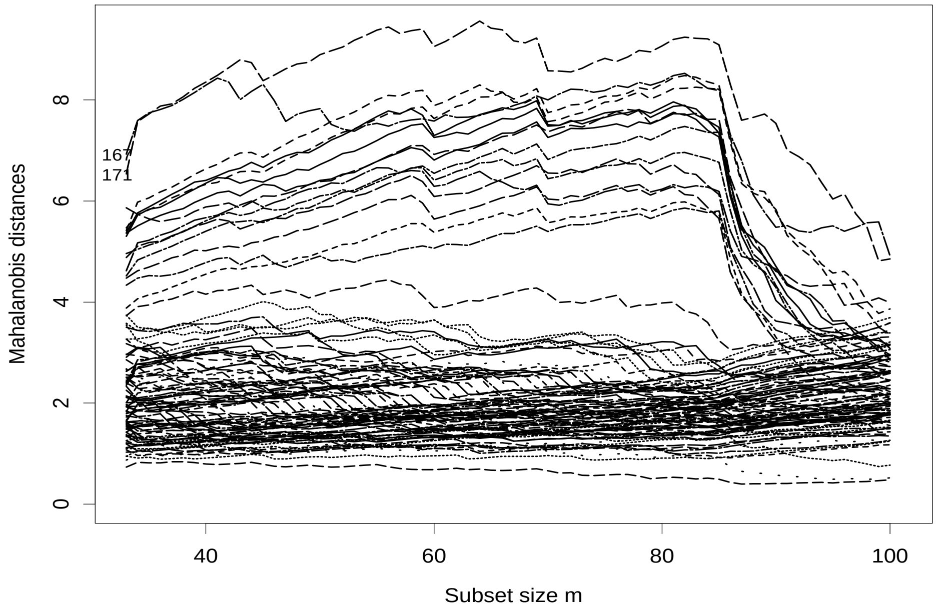

Figure 6: Swiss Banknote Data, forward plot of Mahalanobis distances for the forgeries, showing evidence of a third group

forward search. Figure 6 is a forward plot of the scaled Mahalanobis distances just for the forgeries. In the centre of the plot, around $m = 70$ this shows a clear structure of a central group, one outlier from that group and a second group of 15 outliers. As successive units from this cluster enter after $m = 85$, they become less remote and the distances decrease.

In this example the forward search clearly indicates not only the presence of two groups of notes, but also that the group of forgeries is not homogeneous, itself consisting of two subgroups.

References

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

Atkinson, A. C. and M. Riani (2002a). Forward search added variable t tests and the effect of masked outliers on model selection and transformation. (Submitted).

Atkinson, A. C. and M. Riani (2002b). Tests in the fan plot for robust, diagnostic transformations in regression. *Chemometrics and Intelligent*

Laboratory Systems 60, 87–100.

Box, G. E. P. and D. R. Cox (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B 26*, 211–246.

Flury, B. and H. Riedwyl (1988). *Multivariate Statistics: A Practical Approach*. London: Chapman and Hall.

Nelson, W. (1981). The analysis of performance-degradation data. *IEEE Transactions on Reliability R-30*, 149–155.

Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman (1996). *Applied Linear Statistical Models, 4th edition*. New York: McGraw-Hill.

Riani, M. and A. C. Atkinson (2001). A unified approach to outliers, influence, and transformations in discriminant analysis. *Journal of Computational and Graphical Statistics 10*, 513–544.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association 79*, 871–880.

Zani, S., M. Riani, and A. Corbellini (1998). Robust bivariate boxplots and multiple outlier detection. *Computational Statistics and Data Analysis 28*, 257–270.