Information Criteria for Outlier Detection Avoiding Arbitrary Significance Levels

Marco Riani[a], Anthony Curtis Atkinson[b,*], Aldo Corbellini[c], Alessio Farcomeni[d], Fabrizio Laurini[e]

[a]*University of Parma, Department of Economics and Management; Ro.S.A., Via J.F. Kennedy, 6, 43125, Parma, Italy*
[b]*London School of Economics, Houghton Street, London, WC2A 2AE, United Kingdom*
[c]*University of Parma, Department of Economics and Management; Ro.S.A., Via J.F. Kennedy, 6, 43125, Parma, Italy*
[d]*University of Rome "Tor Vergata", Department of Economics and Finance, Via Columbia, 2, 00133, Roma, Italy*
[e]*University of Parma, Department of Economics and Management; Ro.S.A., Via J.F. Kennedy, 6, 43125, Parma, Italy*

Abstract

Information criteria for model choice are extended to the detection of outliers in regression models. For deletion of observations (hard trimming) the family of models is generated by monitoring properties of the fitted models as the trimming level is varied. For soft trimming (downweighting of observations), some properties are monitored as the efficiency or breakdown point of the robust regression is varied. Least Trimmed Squares and the Forward Search are used to monitor hard trimming, with MM- and S-estimation the methods for soft trimming. Bayesian Information Criteria (BIC) for both scenarios are developed and results about their asymptotic properties provided. In agreement with the theory, simulations and data analyses show good performance for the hard trimming methods for outlier detection. Importantly, this is achieved very simply, without the need to specify either significance levels or decision rules for multiple outliers.

Keywords: automatic data analysis; Bayesian Information Criterion (BIC); Forward Search; Least Trimmed Squares; MM-estimation; S-estimation.

*Corresponding author.
Email addresses: marco.riani@unipr.it (Marco Riani),

1. Introduction

We extend information criteria for model choice to the detection of outliers in regression models. The resulting procedures are computationally straightforward and circumvent the construction of the complicated rules required for the detection of multiple outliers, the properties of which may be only approximately known. We develop criteria for adaptive hard trimming in least squares, together with related criteria for MM- and S-estimation.

Hard trimming in regression requires specification in advance of the proportion of observations to be trimmed. Likewise M-estimation in robust regression and its extensions, such as S- and MM-estimation, require advance specification of the breakdown point or efficiency desired for the estimation procedure. In Section 2 we describe the idea of monitoring that leads to a data dependent estimate of the trimming level or breakdown point. As a result, efficient estimates of the regression parameters are obtained that depend on the actual level of contamination in the data.

In §§3.1 - 3.3 we introduce the three major components of our procedure, respectively the BIC, customarily used in the choice of models, the mean shift outlier model and algebraic details of the forward search (FS). We combine these components in §3.4, using the mean shift outlier model and the ordering of observations from the forward search, to extend BIC to the choice of trimming level for outlier removal in least squares. We prove the consistency of the resulting outlier detection procedure and, in §3.5, provide a procedure for finite samples.

Results in §4.1 use the soft trimming of observations in M-estimation to apply the mean shift outlier model to the development of a form of BIC indicating the appropriate target asymptotic breakdown point or efficiency for specific robust regression analyses. Section 4.2 provides a proof of the difference between the asymptotic properties of BIC from soft and from hard trimming. Section 5 provides numerical procedure for outlier detection for both MM- and S-estimation.

Section 6 uses simulation to explore the relationship between the asymptotic results of §§3 and 4. Hard trimming with the forward search provides the clearest indication of the number of outliers, especially for a larger number of explanatory variables.

A.C.Atkinson@lse.ac.uk (Anthony Curtis Atkinson),
aldo.corbellini@unipr.it (Aldo Corbellini),
alessio.farcomeni@uniroma2.it (Alessio Farcomeni),
fabrizio.laurini@unipr.it (Fabrizio Laurini)

Section 7 applies these methods to three regression examples: one small and straightforward, one small but with a high proportion of outliers and one with 1,405 observations and several outliers. The soft trimming analyses use Tukey's biweight ρ function for both MM- and S-estimation, for which we introduce a new method of estimation of error variance that is appropriate for outlier detection. We arbitrarily classify as outliers those observations with a soft trimming weight below a specified threshold. In the appendix we derive two further forms of BIC for soft trimming. Section 8 summarises the comparative performance of these various forms of BIC on the three data examples.

In §9 we mention the potential extension of our work to model selection in the presence of outliers. The role of statistical significance testing in outlier detection is touched upon. Our overall conclusion is that monitored hard trimming methods provide the sharpest removal of outliers and so the most efficient robust parameter estimates. Of these, the computationally simplest is this paper's version of the forward search. This successfully detects outliers without requiring either specification of the expected contamination level in the data or arbitrary significance levels in, perhaps, arbitrary outlier identification rules.

2. Hard and Soft Trimming

2.1. Three Classes of Estimators for Robust Regression

It is helpful to divide methods of robust regression into three classes (Hampel et al., 1986; Atkinson et al., 2004; Farcomeni and Greco, 2015).

1. Hard $\{0,1\}$ Trimming. In Least Trimmed Squares (LTS: Hampel, 1975; Rousseeuw, 1984) the amount of trimming of n observations when the linear model has p parameters is determined by the choice of the trimming parameter h, $[n/2] + [(p+1)/2] \leq h \leq n$, which is specified in advance. The LTS estimate is intended to minimize the sum of squares of the residuals of h observations. For least squares, $h = n$.
2. Adaptive Hard Trimming. In the Forward Search , the observations are again hard trimmed, but the value of h is determined by the data, being found adaptively by the search (Riani et al., 2014a). Algebraic details are in §3.3.
3. Soft trimming (downweighting). M-estimation and derived methods, depending upon the way in which the residual variance σ^2 is estimated. The intention is that observations near the regression plane retain their value, but the ρ function (§4.1) ensures that increasingly remote observations have a weight that decreases with distance from the plane. The desired

value of either the asymptotic breakdown point (bdp) or of the efficiency has to be specified. This efficiency is that of estimation when the method is applied to a sample from the normal distribution.

The FS starts from a small subset of h_0 robustly chosen observations, by default found using least median of squares (Rousseeuw, 1984). The purpose is ensure that h_0 is outlier free. The search then moves forward incrementing the subset size by one until the final least squares fit is reached, when $h = n$. In this way parameter estimates are obtained for a range of values of h - typically interest is in $n/2 < h \leq n$. We avoid having to prespecify the value of h for LTS by monitoring the fit over a similar range of values.

We consider two derivatives of M-estimation. In S-estimation (Rousseeuw and Yohai, 1984) the estimate of σ^2 is found from a robust estimating equation with specified bdp. The associated estimate of the vector of regression coefficients is called an S-estimator because it is derived from a scale statistic, although in an implicit way.

The asymptotic relationship between the breakdown point and efficiency of S-estimators is that as one increases, the other decreases. In an attempt to break out of this relationship, Yohai (1987) introduced MM-estimation, which extends S-estimation. In the first stage the breakdown point of the scale estimate is set at 0.5, thus providing a high breakdown point. This fixed estimate of residual scale is then used in the estimation of β with a specified high theoretical efficiency.

For both MM- and S-estimation, we again monitor the performance of our outlier detection procedure over a range of values of the settings of the robust method. For S-estimation we monitor values of bdp from 0.5 (the value giving highest trimming) to a value of 0.01, whereas for MM-estimation we monitor over values of the nominal efficiency of estimation of β, for which Maronna et al. (2006, p. 126) recommend a value of 0.85, from 0.5 to 0.99. We observe in some data analyses in §7, as did Riani et al. (2014a), that specification of too high a value for this efficiency can lead to a failure of robustness and to a least squares fit. For all four methods of robust regression we monitor the values of information criteria and the values of residuals or the weights of observations as we move from very robust regression to least squares.

3. Information Criteria

3.1. BIC

There is a large literature on the use of a variety of information criteria in choosing the best model for a set of data. Claeskens and Hjort (2008) provide a treatment with a nice combination of mathematics and data analysis.

Let $L(\theta)$ be the loglikelihood of the n observations y_i, with the parameter vector θ of length p. With $\hat{\theta}$ the maximum likelihood estimate of θ, a general form of information criterion is $IC = 2L(\hat{\theta}) - k(p, n)$, where $k(p, n)$ is a function that penalizes more complicated models. For the Bayesian Information Criterion (BIC) introduced by Schwarz (1978), $k(p, n) = p \log n$ so that the penalty increases with sample size. That model is selected for which BIC is largest.

For the linear regression model with univariate response and independent normal errors of constant variance σ^2, where $\hat{\beta}$ is the least squares estimate of the p parameters β of the linear model and $R(\hat{\beta})$ is the residual sum of squares of the y_i,

$$BIC = -n \log\{R(\hat{\beta})/n\} - p \log n, \tag{1}$$

after constants irrelevant to the comparison of models are ignored.

We recall that use of BIC provides consistent selection of the true model, if that is included in the set of models under consideration. Justifications of the word 'Bayesian' in the name BIC are given, amongst others by Claeskens and Hjort (2008, p.78) and Bhat and Kumar (2010), expanding the original presentation of Schwarz (1978).

In the use of BIC in the choice of a regression model, the comparison is between models with different terms included or removed. As a preliminary to the results of §3.4 we consider BIC for nested regression models. Let the true model be the linear model with $p \times 1$ parameter β_p and $n \times p$ matrix of explanatory variables X_p. A model with $q \times 1$ parameter $\beta_q, q < p$ will be called ***false*** and a model with r parameters, $r > p$ is called *correct*, but is not minimal. For asymptotic results we require

Condition 1. $X_p^T X_p / n \stackrel{n}{\to} M_X$ with det$(M_X) \neq 0$.

We first test a false model. The likelihood ratio test for β_q against β_p has an asymptotic non-central χ^2_{p-q} distribution with non-centrality parameter λ, which from Condition 1, increases as n. The BIC penalty for the comparison of these two models is $(p-q) \log n$, increasing more slowly with n, so that the true model will be chosen as n increases.

For a correct model the likelihood ratio test for β_p against β_r has asymptotically a central χ^2_{r-p} distribution and the BIC penalty is $(r - p) \log n$. Thus, for large n the true model will be preferred. Putting these two together demonstrates the consistency of the BIC.

3.2. Mean Shift Outlier Model

Use of the forward search or least trimmed squares to provide robustness against outliers leads to the comparison of fitted models with differing numbers

of observations. We render outlier detection and deletion compatible with BIC through use of the mean shift outlier model in which deleted observations are each fitted with an individual parameter, so having a zero residual.

Let there be h observations remaining in the fitted model. Then $n-h$ observations will have been deleted. This can be expressed by writing the regression model as

$$y = X\beta + D\phi + \epsilon, \tag{2}$$

where the errors ϵ have constant variance σ^2. Here D is an $n \times (n-h)$ matrix with a single one in each of its columns and in $n-h$ rows, all other entries being zero. These entries specify the observations that are to have individual parameters or, equivalently, are to be deleted (Cook and Weisberg, 1982, p.20; Insolia et al., 2020).

For deletion of the single observation i, D becomes a vector. The likelihood ratio test for $\phi_i = 0$ is the deletion residual r_i^* which compares the observed value of y_i with the prediction $\hat{y}_{(i)} = x_i^T \hat{\beta}_{(i)}$, where $\hat{\beta}_{(i)}$ is the least squares estimate of β when observation i is deleted. Results from the theory of regression diagnostics (Cook and Weisberg, 1982, p.21; Atkinson, 1985, p.23) show that the null distribution of r_i^* is Student's t on $n-p-1$ degrees of freedom. This statistic can then be used to test if y_i, for a specified $i \in \{1, \ldots, n\}$, is an outlier. Sometimes interest is in whether there is an outlier in the data. Since the deletion residuals are correlated, the distribution of the maximum absolute order statistic for a sample of size n from the t_{n-p-1} distribution does not give the distribution of the maximum r_i^* for a sample. That can be approximated by use of the Bonferroni inequality when $\max|r_i^*|$ is tested using $t_{\alpha/n,n-p-1}$, although the power of such a test may be poor. Usually interest is in the more general question as to whether there are some outliers in the data, the number being unspecified. Buja and Rolke (2003) illustrate the large effect on test size arising from such simultaneous tests.

3.3. The Forward Search

The FS fits subsets of observations of size h to the data, with $h_0 \leq h \leq n$. Let $S^*(h)$ be the subset of size h found by the FS, for which the matrix of regressors is $X(h)$. Least squares on this subset of observations yields parameter estimates $\hat{\beta}(h)$. Residuals can be calculated for all observations including those not in $S^*(h)$. The search moves forward with the augmented subset $S^*(h+1)$ consisting of the observations with the $h+1$ smallest absolute values of the residuals. The outliers, if any, enter $S^*(h)$ towards the end of the search.

Above a threshold value on h, often $0.6n$, a test is performed for the presence of outliers before each incrementation of $S^*(h)$. Because of the multiple testing

involved, Riani et al. (2009) propose a complicated rule based on quantiles of order statistics which is intended to have a simultaneous size of 1% for samples with n up to around 1,000. Examples of the use of this rule in monitoring regression are in Riani et al. (2014a).

The consistency of the FS estimator when the data contain no outliers is proved by Cerioli et al. (2014) for multivariate data and by Johansen and Nielsen (2016) for univariate regression. We now allow outliers in the data generating distribution.

Let the uncontaminated observations belong to the set $\mathcal{H}$, of cardinality h^* and let the outliers belong to $\mathcal{H}^o$, with $\mathcal{H} \cup \mathcal{H}^o$ containing all observations. The number of outliers is then $n - h^*$, neither this number, nor their identity being known. Asymptotically, we consider a fraction of contaminated observations $\gamma = 1 - h^*/n,\ 0 \leq \gamma < 0.5$. The FS progresses by ordering the squared residuals $e_i^2(h); i \in \{1, \ldots, n\}$.

Asymptotic distribution of squared residuals. The small-sample distribution of the residuals $e_i^2(h)$ depends on the leverage $l_i = x_i^T \{X(h)^T X(h)\}^{-1} x_i$. From Condition 1, $l_i \xrightarrow{n} 0$. Then, in the absence of outliers in $S^*(h)$, $e_i^2(h) \xrightarrow{\text{a.s.}} \sigma^2 \chi_1^2$. The individual outliers have asymptotically the noncentral chi-squared distribution $\sigma^2 \bar{\chi}_1^2(\lambda_{hi})$. For asymptotic results about robustness we assume

Condition 2. For all $i \in \mathcal{H}^o$, $\lambda_{hi} = o(n)$.

Condition 2 is a rather strong, although standard, separation condition which requires outliers to be increasingly far from the clean observations as n grows, in such a way that the non-centrality parameter of the resulting chi-squared distribution grows faster than n. This condition is nevertheless slightly less stringent than some similar ones already considered in the literature, e.g., the separation condition in Cerioli et al. (2014) where it is required that the probability mass for outliers is asymptotically concentrated exponentially fast in the tails of the distribution of the clean observations. Here we simply assume that the the non-centrality parameter λ_{hi} grows as the square of the distance between the clean data centroid and the centroid of the outlier generating distribution.

We call a *correct* ordering of the observations one in which $S^*(h^*) = \mathcal{H}$; that is that at step h^* there are no outlying observations in $S^*(h^*)$.

3.4. Extended BIC for Outlier detection

To incorporate deletion of observations in BIC (1), let the residual sum of squares for a parameter estimate b when $n-h$ observations are deleted be $R_h(b)$,

To allow for the additional parameters in (2), BIC (1) is accordingly replaced by

$$\text{BICH} = -n \log\{R_h(\hat{\beta}_h)/h\} - (p + n - h) \log n. \tag{3}$$

The rationale in (3) is that under the model all observations are still included in the estimation set (hence the use of n), but only h are used for computation of the residual sum of squares. Finally, $(p + n - h)$ is the number of parameters.

Theorem 1. *Assume Conditions 1 and 2. Let h_n^* denote the cardinality of uncontaminated observations for a sample of size n, and assume $\lim_n h_n^*/n = 1 - \gamma$ for some $\gamma < 0.5$. The initial estimation set, of cardinality h_{0n}, is outlier free; we assume $\lim_n h_{0n}/n > 0$ and $h_{0n}/h_n^* \leq 1$ for all n. Then, for hard trimming,*

$$\lim_n \frac{\arg\max_h\{-n \log(R_h(\hat{\beta}_h)/h) - (p + n - h) \log n\}}{n} \leq 1 - \gamma,$$

that is, as n grows BICH is a maximum at an estimation set which does not include outliers.

Proof of Theorem 1 Since the initial set of h_{0n} observations is outlier free by assumption, and $\lim_n h_{0n}/n > 0$, $\hat{\beta}_{h0n}$ is a strongly consistent estimate of β. Then,

$$\lim_n \log\{R_{h_{0n}}(\hat{\beta}_{h_{0n}})/h_{0n}\} = \log(\sigma^2(h_{0n})\sigma^2), \tag{4}$$

where $\sigma^2(h)$ is a correction factor discussed in equation (6) of Section 3.5. Let $\arg\max_h\{-n \log(R_h(\hat{\beta}_h)/h) - (p + n - h) \log n\} = \tilde{h}_n$.

We proceed by contradiction. Suppose that there is at least one outlier in the estimation set based on $\tilde{h}_n$ observations. By assumption, this implies that $\tilde{h}_n > h_{0n}$, since the initial set is outlier free. Due to Condition 2,

$$\lim_n \log\{R_{\tilde{h}_n}(\hat{\beta}_{\tilde{h}_n})/(\tilde{h}_n)\} = o(\log n). \tag{5}$$

Now compare the initial estimation set with the optimal set in terms of BICH difference, divided by n. Let

$$\begin{aligned} K = &- \log\left(\frac{R_{\tilde{h}_n}(\hat{\beta}_{\tilde{h}_n})}{\tilde{h}_n}\right) - n^{-1}(p + n - \tilde{h}_n) \log n + \log\left(\frac{R_{h_{0n}}(\hat{\beta}_{h_{0n}})}{h_{0n}}\right) \\ &+ n^{-1}(p + n - h_{0n}) \log n. \end{aligned}$$

Simplifying and collecting terms we obtain

$$\begin{aligned} K &= -\log\left(\frac{R_{\tilde{h}_n}(\hat{\beta}_{\tilde{h}_n})h_{0n}}{R_{h_{0n}}(\hat{\beta}_{h_{0n}})\tilde{h}_n}\right) + n^{-1}(\tilde{h}_n - h_{0n})\log n \\ &\leq -\log\left(\frac{R_{\tilde{h}_n}(\hat{\beta}_{\tilde{h}_n})h_{0n}}{R_{h_{0n}}(\hat{\beta}_{h_{0n}})\tilde{h}_n}\right) + n^{-1}(n - h_{0n})\log n. \end{aligned}$$

Combining (4) and (5) the first summand is seen to diverge (negatively) at a faster rate than the second one, which, since by assumption $0 < \lim_n h_{0n}/n < 1$, diverges (positively) at the rate $O(\log n)$. There will then exist $\bar{n}$ such that, for $n \geq \bar{n}$, BICH associated with the initial set will be larger than BICH associated with $\tilde{h}_n$. This contradicts the definition of $\tilde{h}_n$ as the maximum of BICH. Asymptotically, the optimal estimation set in terms of BICH must then be outlier free. Since there are at most h_n^* uncontaminated observations,

$$\lim_n \frac{\arg\max_h\{-n\log(R_h(\hat{\beta}_h)/h) - (p+n-h)\log n\}}{n} \leq 1-\gamma,$$

which completes the proof. □

Let $h^\dagger$ be the size of the subset for which the rescaled value of BIC (7) is maximized. The theorem proves that asymptotically $S^*(h^\dagger)$ will not include any outliers. However, the asymptotic conditions may not be satisfied for finite sample sizes and small separation of the outliers, as in the simulations of Figures 3 and 4. Then $S^*(h^\dagger)$ may contain some outliers and miss some non-outlying observations.

Least Trimmed Squares. We note that the theorem also applies to monitored LTS, in which the best subset $S^*(h)$ is found for a range of values of h, rather than for a single specified trimming value as in the original proposal (Rousseeuw, 1984). The difference is then in the algorithms used for calculating $S^*(h), h_l \leq h < n$, where h_l is the lower limit of subset size that is of interest. If both algorithms result in the same value of $S^*(h)$ over the range of h, the residual sums of squares will be identical as will be the values of BICH. To the best of our knowledge formal conditions under which the two algorithms give the same $S^*(h)$ have not yet been studied.

3.5. Finite Sample Extended BIC for Outlier Detection

We now consider two further points arising from the application of BICH to finite samples.

3.5.1. Estimation of σ^2

In (3) σ^2 is estimated by $\{R_h(\hat{\beta}_h)/h\}$. To find the $n-h$ outlying observations, the FS or LTS deletes the observations most remote from the fitted model. If there are no outliers the value of $R_n(\hat{\beta}_n)$ leads to an asymptotically unbiased estimate of σ^2. Then the deletion of the $n-h$ most remote observations yields the parameter estimate $\hat{\beta}_h$ and the residual sum of squares $R_h(\hat{\beta}_h)$, which provides a too small estimate of σ^2 since it is calculated from the central h residuals. The variance of the truncated normal distribution containing the central h/n portion of the full distribution is.

$$\sigma^2(h) = 1 - \frac{2n}{h}\Phi^{-1}\left(\frac{n+h}{2n}\right)\phi\left\{\Phi^{-1}\left(\frac{n+h}{2n}\right)\right\}, \tag{6}$$

where $\phi(.)$ and $\Phi(.)$ are respectively the standard normal density and c.d.f. See, for example, Johnson et al. (1994, pp. 156-162). We scale up the value of $R_h(\hat{\beta}_h)$ to obtain the corrected BIC

$$-n\log[R_h(\hat{\beta}_h)/\{h\sigma^2(h)\}] - (p+n-h)\log n. \tag{7}$$

This consistency correction is standard in robust regression (Rousseeuw and Leroy, 1987, p.130). The correction $\sigma^2(h)$ in (6) is the one-dimensional case of the general result in Tallis (1963) on elliptical truncation in the multivariate normal distribution.

For hard trimming (7) can be rewritten with weights $w_i = 0, i \in n-h$ and one otherwise as

$$\text{BICW} = -n\log\left[R_h(\hat{\beta}_h)/\{\sigma^2(h)\sum_{i=1}^{n} w_i\}\right] - \left\{p + \sum_{i=1}^{n}(1-w_i)\right\}\log n, \tag{8}$$

where $\sum_n w_i = h$.

In §4.1 we rewrite BICH (3) in a weighted form for soft trimming.

3.5.2. Masking

The proof of Theorem 1 relies on Condition 2 for all $y_i \in \mathcal{H}^o$, particularly through the distribution of the residuals. In the analysis of data the outliers may not be very large and outlying observations will appear less so as the FS progresses and the parameter estimates are corrupted once outliers are included in $S^*(h)$. Although the value of $R_h(\hat{\beta}_h)$ will continue to increase with h it may not do so sufficiently fast to outweigh the decrease in the penalty of $-h\log n$. As a consequence, the value of BICW may increase for $h > h^\dagger + 1$. We see an example in §7.2.

4. Soft Trimming

4.1. BIC for Soft Trimming

We now extend the idea of downweighting in BICW (8) to include soft trimming such as S- and MM-estimation. The S-estimator of the regression parameters is defined as

$$\hat{\beta}_S = \min_{\beta} \sum_{i=1}^{n} \rho\left(\frac{e_i}{\hat{\sigma}_S}\right), \tag{9}$$

where $e_i = y_i - \hat{\beta}^T x_i$ is the i-th unscaled residual and $\hat{\sigma}_S$ is the robust estimate of σ found by minimizing the dispersion of the residuals as defined by Rousseeuw and Yohai (1984). A consistency factor is applied to $\hat{\sigma}_S$. S- and MM-estimation differ in the choice of $\hat{\sigma}$.

Let $\psi(u) = d\rho(u)/du$. Then, in the iterative least squares algorithm for M-estimation, the weights are taken as $\tilde{w}(u) = \psi(u)/u$, that is $\tilde{w}_i = \tilde{w}(u_i) = \tilde{w}(e_i/\hat{\sigma})$. It is required that $w_i = 1$ for observations that are not to be downweighted. Since for least squares the bdp is zero, no observations are downweighted and all have the same weight $w(0)$. We can then simply rescale and set

$$w_i = \tilde{w}_i / w(0). \tag{10}$$

We can now adapt the BICH (3) for hard trimming to soft trimming. Each observation will have a weight $w_i \in [0, 1]$, smoothly varying from 1 to 0 as the observation becomes more outlying. We obtain

$$\text{BICW}_\rho = -n \log\left\{R(\hat{\beta}_\rho)/\sum w_i\right\} - \left\{p + \sum_{i=1}^{n}(1 - w_i)\right\} \log n, \tag{11}$$

where $R(\hat{\beta}_\rho)$ is the weighted sum of squared residuals e_i. The expression for BICW (8) specifically shows the consistency correction $\sigma^2(h)$, whereas, in (11), $R(\hat{\beta}_\rho)$ has already been corrected to provide a consistent estimate.

4.2. BIC for Soft Trimming versus Hard Trimming

For hard trimming the bdp $d = 1 - h/n$. Theorem 1 shows that, eventually, use of BIC for the FS produced a value of $d^*_{\text{HT}} \geq 1 - h^*/n$. This section provides a proof that the bdp from use of the function $\rho(u)$ (9) to remove outliers leads to the same result under the separation condition. We also argue why we expect in general downweighting to lead to a value of bdp greater than $1 - h^*/n$, and so to parameter estimates of reduced efficiency.

Condition 3. Let $w(u)$ be such that

1. For any $|u| > 0, w(u) \leq w(0)$;
2. For $|u| > c, w(u) = 0$.

Tukey's biweight (Beaton and Tukey, 1974), which we use in our numerical examples, satisfies these conditions.

Theorem 2. *Let the estimate of d maximizing BICW$_\rho$ for soft downweighting be d^*_{nS}. Then, under Conditions 1 - 3,* $\lim_n d^*_{nS} = \lim_n d^*_{nHT}$.

Proof of Theorem 2 For least squares $d = 0$ and, in Condition 3, $c = \infty$. As d is increased, c decreases. Combining Condition 2 with Condition 3 it can be seen that there exists n^* such that for $n > n^*$ $w_i = 0$ for $i \in \mathcal{H}^o$. Hence, for $n > n^*$, FS and downweighting are equivalent. The proof is then equivalent to that of Theorem 1. □

When we relax Condition 2 we often can expect $\lim_n d^*_{n\mathrm{S}} > \gamma$. Indeed, when the outliers are less extreme, the weights (10) of the outliers may not all be zero, even though they may be the most appreciably downweighted observations. The monitoring plots of weights in §7 illustrate this point. Selection of an M-estimator with unnecessarily high bdp, leads to a loss of efficiency in estimation. Figure 5 and 6 of Riani et al. (2020) plot the relationship between breakdown point and efficiency for several well-known forms for $\rho(u)$.

5. Implementation of MM- and S-estimation

In our numerical examples we use Tukey's biweight (Beaton and Tukey, 1974) in which the boundary of the central region of the ρ function is defined by the parameter c. As $c \to \infty, \rho(.)$ approaches a quadratic and the fitted model becomes that from least squares: the bdp $d \to 0$ and the efficiency *eff* $\to 1$. We monitor the behaviour of BICW$_\rho$ as the values of these parameters change. Riani et al. (2014b) give computationally fast procedures for determining the value of c for Tukey's biweight which yields specified values of d or of *eff*. Use of MM-estimation provides estimates of β for a fixed, although data dependent, value of σ as *eff* varies. On the other hand, S-estimation (Rousseeuw and Yohai, 1984) provides simultaneous robust estimates of β and σ^2 as d varies.

In our robust calculations for MM we typically take 50 values of *eff* over the range 0.5 to 0.99. The value maximizing BICW$_\rho$ is denoted $eff^\dagger$. For S-estimation we vary d over a set $\mathcal{D}$ such that bdp varies from 0.5 to 0.01, giving the maximizing value $d^\dagger$. In both cases we also include LS ($d = 0$ or *eff* = 1.) to

cover the absence of any outliers. The procedure for MM-estimation is straightforward because, throughout we use a very robust estimate of σ^2. However, for S-estimation, each $d \in \mathcal{D}$ yields parameter estimates $\hat{\beta}_d$ and $\hat{\sigma}_d^2$, leading to raw residuals $e_{di} = y_i - x_i^T \hat{\beta}_d$. But, if d is too small, the estimate of σ^2 may be inflated due to contamination by outliers, even if $w_i < 1$ for $i \in \mathcal{H}^0$. To obtain a consistent estimate of σ^2 in the presence of contamination we use the very robust estimate $\hat{\sigma}_0^2$, which is the value of $\hat{\sigma}_d^2$ for $d = 0.5$. We then rescale the residuals used in calculating BICW$_\rho$ to obtain $r_{0di} = e_{di}/\hat{\sigma}_0$, so avoiding the effect of too large an estimate of σ on the scaled residuals. Monitoring the value of BIC over $\mathcal{D}$ leads to the maximizing value $d^\dagger$. However, because of the use of $\hat{\sigma}_0^2$, this procedure does not completely reflect the downweighting of outliers.

We follow the reasoning of Rousseeuw and Yohai (1984) in their derivation of S-estimation, but allow the data to determine the estimate of σ^2. At the value $d^\dagger$ let the residuals scaled by $\hat{\sigma}_0$ be $r_{0i}^\dagger$. To model the effect of downweighting outliers we find the residuals $\hat{r}_{di}$ from S-estimation, that is using $\hat{\sigma}_d$ as an estimate of σ and calculate their distance from $r_{0i}^\dagger$. The minimum value of this distance over $\mathcal{D}$ gives us a new estimated maximizing bdp d^*, depending on the S-estimates of β and σ calculated with the same value of d. We found a useful measure of distance to be the weighted sum of squares

$$SSD(d) = \sum_{i=1}^{n} w_{0i}^\dagger (\hat{r}_{di} - r_{0i}^\dagger)^2, \tag{12}$$

where the $w_{0i}^\dagger$ are the weights w_i (10) associated with the residuals $r_{0i}^\dagger$. Then

$$d^* = \arg\max_{d \in \mathcal{D}} SSD(d). \tag{13}$$

6. Simulations

We now use numerical simulation to illustrate some of the properties of our procedure for small samples. We are interested in behaviour as efficiency increases. For the FS we monitor performance as the value of h increases, one observation at a time. The other three methods were originally defined either by specifying efficiency or bdp, and we monitor them for 50 values of these. Table 1 presents expressions for the two weighted forms of the BIC, namely BICW and BICW$_\rho$ that are important in our simulations.

We begin in Figure 1 with the distribution of trajectories of the values of BIC for four forms of outlier detection. The simulations are for $n = 200$, with four explanatory variables ($p = 5$) simulated from standard normal distributions (the

Table 1: Two different forms of BIC

Criterion	Equation	Estimation	Formula
BICW	(8)	LS	$-n[\log R_h(\hat{\beta}_h)/\{\sigma^2(h)\sum_{i=1}^n w_i\}]$ $-\{p+\sum_{i=1}^n(1-w_i)\}\log n$
BICW$_\rho$	(11)	MM & S	$-n\log\left\{R(\hat{\beta}_\rho)/\sum w_i\right\}$ $-\{p+\sum_{i=1}^n(1-w_i)\}\log n$

procedures are invariant to the values of the regression parameters, so these are set to zero). The observational errors are independent standard normal with a shift of $\delta = 5$ added to 20 observations, so that there is 10% contamination, that is $\gamma = 0.1$. There are 200 simulations; we plot the 1, 50 and 99% quantiles of the estimated values of BICW. The top row of the figure shows the trajectories of BICW for the two hard trimming methods, monitored LTS and the FS, with BICW$_\rho$ for the two soft trimming procedures, S and MM, in the lower row. All curves have a similar shape, at first increasing almost linearly to a maximum before decreasing more or less sharply.

Since we monitor from $n/2$ to n, when $n > 100$ we evaluate at fewer values of h for LTS than we do for the FS. We also start LTS estimation anew for each h. We could modify monitored LTS by evaluating at each value of $h \geq n/2$ and using the parameter estimates from $S^*(h)$ to provide starting values for estimation for $S^*(h+1)$. These modifications would reduce some of the differences our simulations show between LTS and the FS. We choose not to do this, as we are not intending to develop new hybrid algorithms, but rather to compare those already in the literature.

The curves for hard trimming have a wider range of trajectory values than those for soft trimming, but the feature of main interest is the position of the maximum of the curves and so the indicated degree of trimming. For the hard trimming methods the maxima are close to 0.1, correctly indicating 10% contamination. For S-estimation the maximum is near a bdp of 0.2, whereas for MM-estimation the maximum is close to a high efficiency of 0.93. In all cases the presence of some outliers is indicated. In results not shown here we ran simulations for a smaller proportion of outliers and when there were none. In this latter case the trajectories for all four procedures increase linearly with the maximum at non-robust least squares estimation.

Figure 2 shows boxplots of the position of the maxima for the trajectories of the four versions of BIC plotted in Figure 1. We see that the boxplots for the

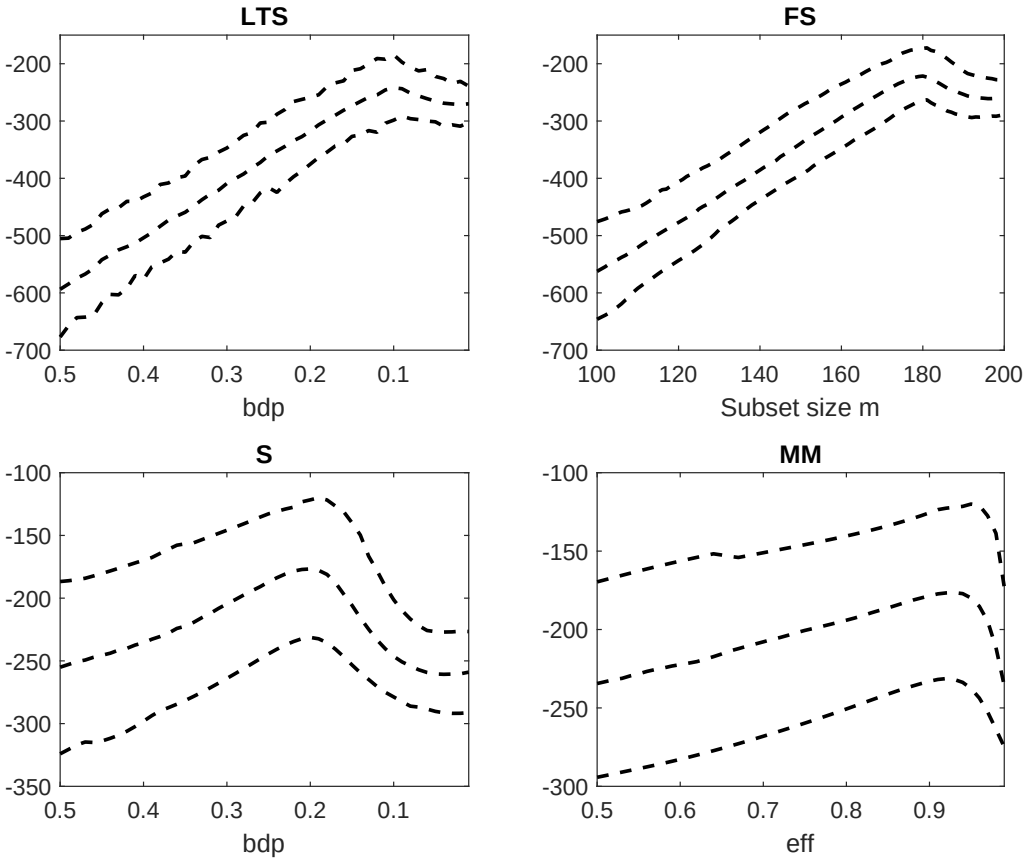

Figure 1: Distribution of trajectories of BIC for outlier detection when $p = 5$; 10% contamination. Upper row, BICW for hard trimming; lower row BICW$_\rho$ for soft trimming. 200 simulations, $n = 200$, 1, 50 and 99% points of the distribution

two hard trimming methods in the upper row are both centered around detecting 10% of outliers, but that the variability for LTS is greater than that for the FS. The scatter for both soft trimming methods is greater than that for the FS. As noted by a referee, the S-estimator is always computed with a bdp greater than the actual contamination level, which makes it more resistant to contamination. For the remainder of this simulation section we focus on hard trimming methods.

The outlier detection properties of the two hard trimming methods depend on the numerical details of the algorithms we have used. For each simulated set of 200 observations when using LTS we calculated BICW for 50 values of h from $0.5n$ to $0.99n$. The calculations for the different values of h are independent, using 1,000 elemental subsets with concentration steps (Rousseeuw and Van Driessen, 1999) to find the estimates of the regression parameters for each

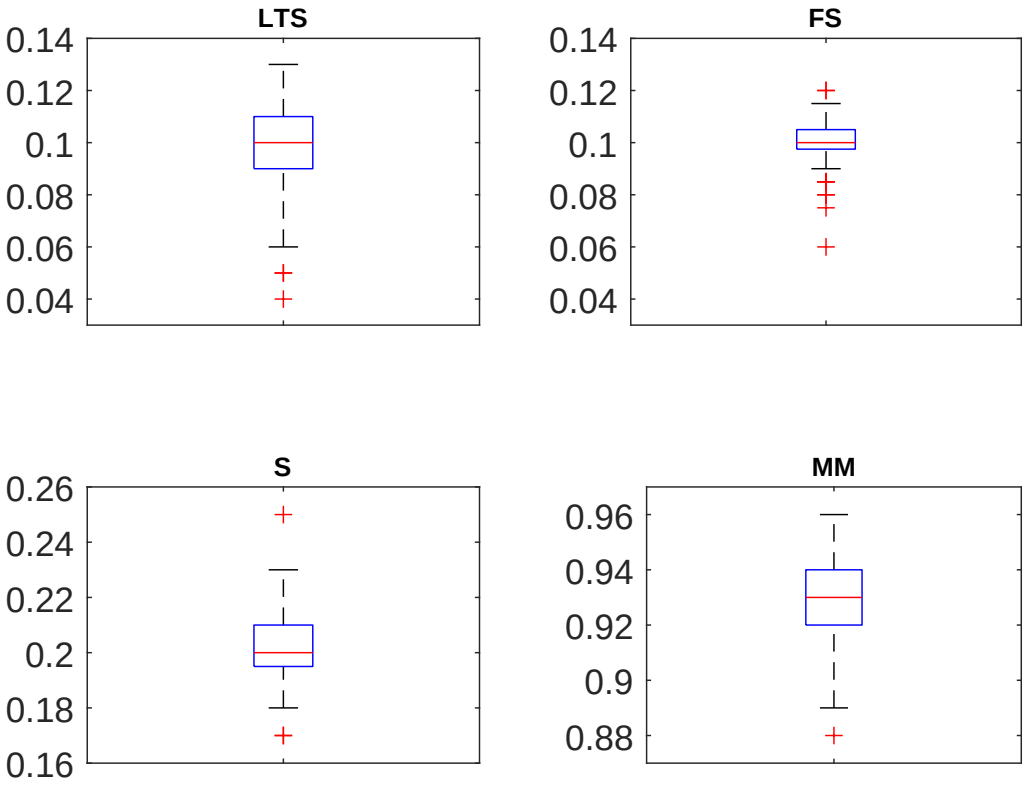

Figure 2: Boxplots of position of maxima of BIC outlier detection trajectories plotted in Figure 1; $p = 5$, 10% contamination. Upper row, BICW for hard trimming; lower row BICW$_\rho$ for soft trimming. 200 simulations, $n = 200$

value of h. When $n = 200$ the number of observations in h changes in steps of two, so that the proportion of outliers found can only change in steps of 0.01. For the FS we take 1,000 elemental subsets of all n observations, calculate the value of the LTS criterion with bdp 50% for each subset and take as the initial subset for the FS that yielding the minimum of the LTS criterion. Use of LTS, rather than the default LMS, at this point has no effect on the numerical results.

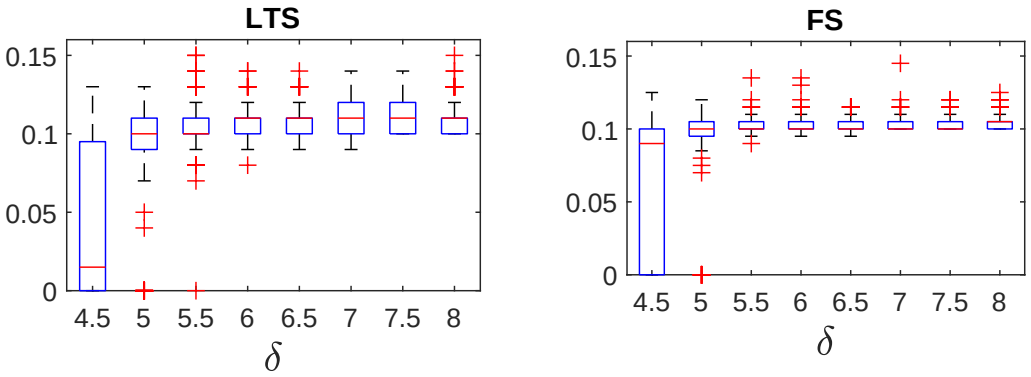

Figure 3: Boxplots of the proportion of observations declared as outliers as the outlier shift δ increases; $p = 5$, 10% contamination. Left-hand panel: LTS, right-hand panel FS; $n = 200$

We now look at the proportion of observations declared as outliers as the shift δ in the twenty outliers increases. Boxplots for these results are in Figure 3. The median number of outliers for each shift is shown, in the online version, as a red horizontal line. This shows that when $\delta = 4.5$, the median number of outliers declared by LTS is around 0.01, whereas it is a little less than 0.1 for the FS. In general, as δ increases from 4.5 to 8.0, the proportion of outliers detected for both LTS and the FS is around 0.1. The figure illustrates in three ways the improvement in using the FS compared to monitored LTS: the detection of outliers occurs for a lower value of δ, the mean results are more stable for larger δ as indicated by the width of the boxplots and there are fewer simulations leading to false declarations of extra outliers. The plots also show the effect for LTS of increments in the proportion of outliers detected in steps of 0.01, as opposed to 0.005 for the FS.

We now finally, for these simulations with $n = 200$, look at the number of good observations declared as outliers and the number of outliers correctly detected, again as the shift in the outliers increases. The average results over 200 simulations for both $n = 200$ and 500 are in Table 2. As the outlier shift δ increases from 4.5 to 8.0 both LTS and the FS detect all outliers, with the FS

detecting more for lower values, especially 4.5 and 5 when $n = 500$. As δ increases, the number of false declarations also increases, although the average number is smaller for the FS. When $n = 500$ the value of h for LTS is incremented by 5 observations. It is interesting that when $\delta = 8$ the average number of false declarations for the FS is 0.9, whereas it is 3.4 = 0.9 + 5/2 for LTS, 5/2 being half of the increment size for the values of h in this simulation.

Table 2: Average number of correct and incorrect declarations of outliers for LTS and the FS as a function of outlier shift δ and sample size; $p = 5$, 10% contamination

	$n = 200$				$n = 500$			
δ	LTS		FS		LTS		FS	
	Outliers	Good	Outliers	Good	Outliers	Good	Outliers	Good
4.5	8.84	0.58	13.31	0.42	3.10	0.10	13.60	0.14
5	16.38	1.25	18.50	0.63	32.6	1.3	45.50	0.73
5.5	19.42	1.50	19.84	0.63	47.57	2.80	49.40	0.85
6	19.85	1.63	19.96	0.62	49.54	3.16	49.84	0.89
6.5	19.94	1.66	19.99	1.46	49.81	3.24	49.96	0.90
7	19.99	1.95	20	0.66	49.94	3.34	50.00	0.90
7.5	20	1.99	20	0.72	49.98	3.32	50	0.90
8	20	1.84	20	0.72	50	3.4	50	0.90

Figure 3 and the related Table 2 provide a nice illustration of the result of Theorem 1. As both δ and n increase all the outliers are identified, both by LTS and the FS. However, there is always a small number of good observations that are mistakenly declared to be outliers. The results so far do show that the FS performs better on all measures than LTS, although the difference is not large provided $\delta > 5$. We also considered the performance of the two methods when the number of explanatory variables increases, both for $n = 200$ and 500, with p increased to 15. In summary, for this increased number of explanatory variables, LTS detected a maximum of four outliers until $\delta = 7.5$. For $\delta = 8$ too many were detected. The unsatisfactory behaviour of LTS when $p = 15$ was unexpected. We accordingly repeated the simulation with initial LTS subsets of 10,000, rather than 1,000, observations. There was little improvement.

In all these simulations the outlying responses have been within the range of the explanatory variables. As a final extension of our investigation, we consider outliers at leverage points. The 14 explanatory variables were simulated independently but were generated to have an average value of $R^2 = 0.8$ before contamination. We kept n at 500, including 50 outliers so that γ remained at 0.1.

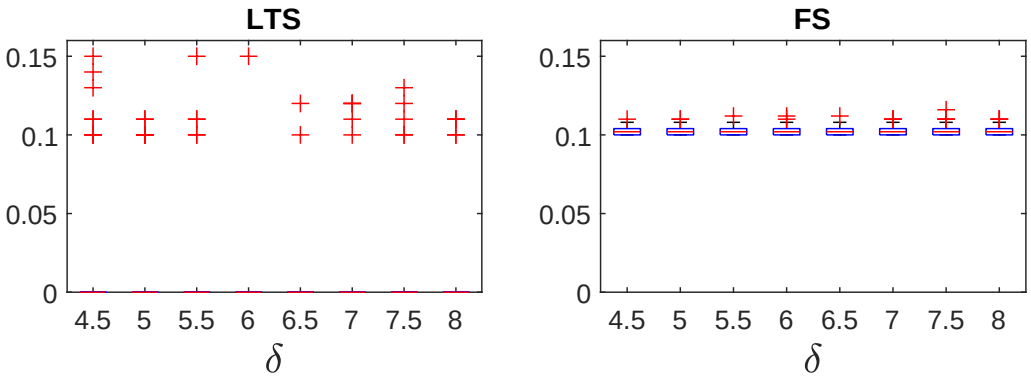

Figure 4: The effect of outliers at leverage points, combined with a larger number of explanatory variables. Boxplots of the proportion of observations declared as outliers as the outlier shift and explanatory variable remoteness δ increases when $n = 500$ and $p = 15$; 10% contamination. Left-hand panel: LTS, initial subsets of 1,000 observations; right-hand panel FS

For each value of δ not only was this value added to the 50 outlying responses but also to all explanatory variables. We thus generated outliers at extreme leverage points. Boxplots for the results are in Figure 4. The left-hand panel of the figure shows that LTS completely fails to detect the many outliers; the medians of the boxplots are all at zero, as shown in the online version of the plot by the red lines at this value. On the other hand, the FS reveals all the outliers, with a few extra, for all values of δ. This figure leads to the same conclusions as those for the two simulations with $p = 15$ mentioned above, in which the outliers were not at leverage points and the initial subsets for LTS were of size either 1,000 or 10,000.

We have no explanation for the surprisingly poor behaviour of LTS when $p = 15$. As Olive (2020) stresses, a complex estimator, such as many of those used in robustness, depends not only on the mathematical formulation and theoretical properties of the estimator, but also on the details of the algorithm used to provide numerical values of estimators. The good performance of the FS when $p = 15$ suggests one approach to an improved LTS algorithm.

7. Examples

We now illustrate the finite-sample properties of the procedure through the analysis of three distinct examples. We compare LTS and the FS with MM- and S-estimation. Since the purpose of our paper is to provide a method of outlier

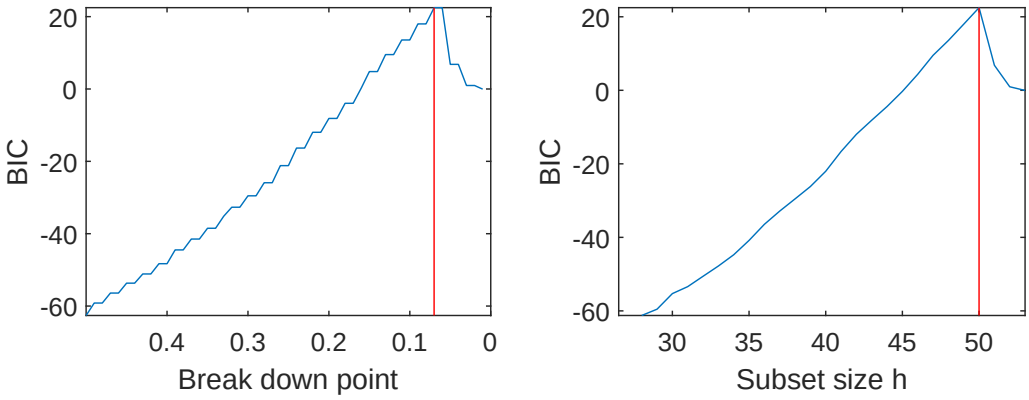

Figure 5: Mental illness data. Monitoring plots of BICW. Left-hand panel, Least trimmed squares (LTS); right-hand panel, Forward Search (FS). The stepping for LTS arise from a search over 50 values of h

detection we need to define an outlier in soft trimming, which we arbitrarily take as an observation for which $w_i < 0.01$.We then count the number of outliers at the BIC maximizing values $h^\dagger$, $eff^\dagger$, $d^\dagger$ and d^*. In our examples we use monitoring plots of residuals and weights to illuminate the procedure. But, for automatic outlier detection, we are not initially interested in monitoring over a series of grids, but in selecting a single value for further investigation.

In the first example there are data on 53 patients, the response of three of which are contaminated. The remaining data follow a normal distribution. In the second example, the 'Stars' data, there are only 47 observations, but the structure of outliers is more complicated; monitoring plots of residuals show a clear switch from a robust to a non-robust fit when the target bdp is too low. The third data set has 1,405 observations. The responses in two of these examples require transformation; in all cases we work with a suitable response transformation found outside this paper. The two forms of BIC are listed in Table 1, with a summary of the outliers detected for the three examples in Table 3.

7.1. Example 1: Mental Illness Data

Kleinbaum and Kupper (1978, p.148) describe observational data on the assessment of mental illness of 53 patients. The data come from a psychiatrist's assessment of mental retardation and degree of distrust of doctors in newly hospitalized patients. After six months of treatment, a value is assigned for the degree of illness of each patient. Atkinson et al. (2021) showed that when degree of illness is regressed on the two initial assessments, there is strong evidence for transformation of the response. The Box-Cox transformation indicates the log

transformation. After transformation the data are well behaved. We study the effect of outliers by modifying three of the smallest observations (17, 30 and 53), setting them equal to one. This contamination causes the log transformation to be rejected.

We start with least squares analysis of the logged contaminated data. The left-hand panel of Figure 5 shows the plot of BICW for LTS estimation and the right-hand panel that for FS. Both show the almost linear increase as more observations are included in the fit, as indicated by Theorem 1, until a peak for h near n, after which there is a sharp decline. For the FS the peak is at $h = 50$. For LTS the bdp at the maximum is 0.07, which corresponds to $h = 50$. Here both LTS and the FS have correctly identified the 3 outliers without the need for a complicated outlier detection rule such as that of Riani et al. (2009) and without the need of specifying subjective confidence bands. The stepping of LTS arises because, for breakdown point d, h is found as $\lceil n(1-d) \rceil$.

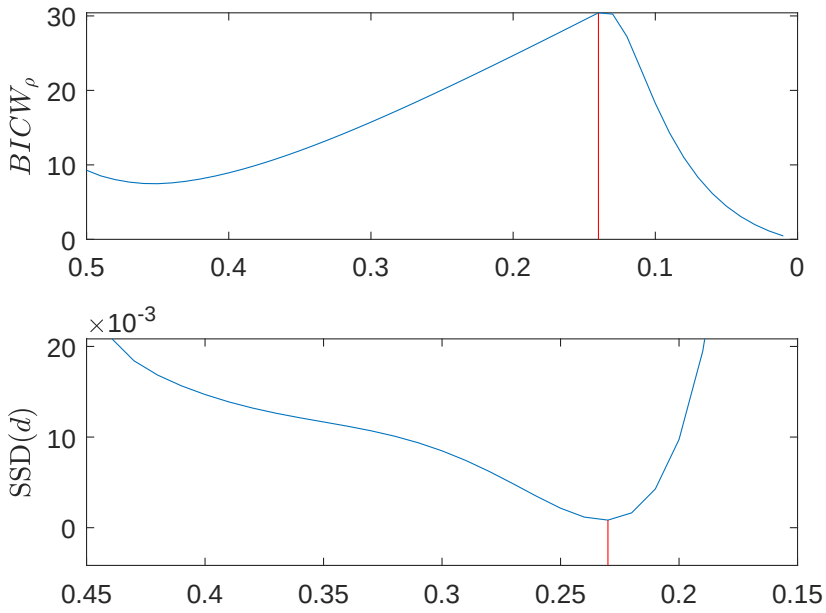

Figure 6: Logged mental illness data. Upper panel: Monitoring plot of BIC_ρ for S-estimation using $\hat{\sigma}_0$; $d^\dagger = 0.14$; lower panel: weighted sum of squares $\text{SSD}(d)$ of the differences of the residuals $\hat{r}_{di}$ and r_{0di}; $d^* = 0.23$.

The monitoring plots of BIC from MM- and S-estimation are similar in shape to that for the FS in Figure 5. The maximum for MM is $eff^\dagger = 0.96$. For S-

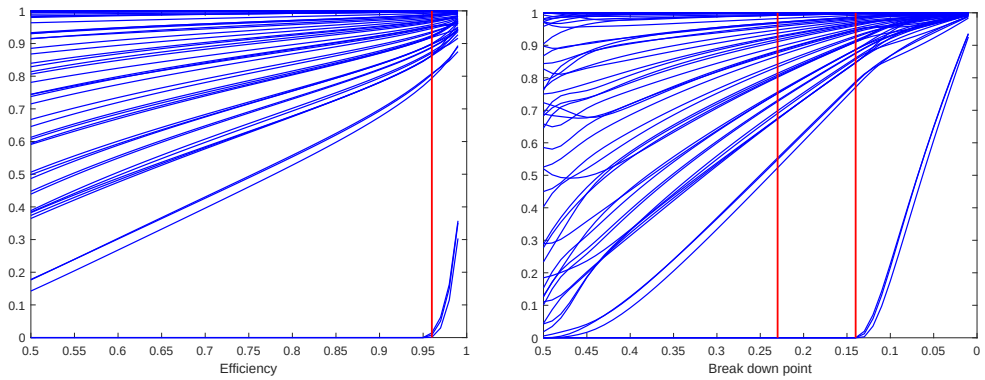

Figure 7: Logged mental illness data. Monitoring plot of weights $\hat{w}_i$. Left-hand panel MM-estimation; $\mathit{eff}^\dagger = 0.96$. Right-hand panel S-estimation with scale estimate $\hat{\sigma}_d$; $d^* = 0.23, d^\dagger = 0.14$

estimation the maximizing value when σ is estimated by σ_0 is $d^\dagger$ = 0.14. The trajectory for BIC$_\rho$ using this estimate is plotted in the upper panel of Figure 6. The lower panel of the figure shows the trajectory of the weighted sum of squared residuals SSD(d) (12), the minimum of which gives the estimated optimum bdp d* = 0.23, greater than the value of $d^\dagger$. For LTS and the FS the efficiency when the three outliers are deleted is 0.943, so that the bdp is 0.057. These results for S-estimation are in line with the comment following Theorem 2 that often $d^\dagger > \gamma = 0.057$. For small outlier displacements δ and small sample sizes, some soft trimming methods may, as in the other examples of this section, fail to detect all the outliers. Then the inequality may not hold. Similar remarks, with sign reversed, hold for comparisons of the values of eff.

To determine the outliers from soft trimming and to interpret the results we look at monitoring plots of the weights w_i over the range of values of ***eff*** or d. The left-hand panel of Figure 7, for MM-estimation shows the small weights for the three outliers. For $\mathit{eff} = 0.95$, three outliers are detected but at $\mathit{eff}^\dagger$ (= 0.96) only two are found. The right-hand panel of the figure shows the weights for S-estimation. The weights of the three outliers are smaller than 0.01 by a bdp of 0.14, which is the value $d^\dagger$. The three outliers, only, are also detected at d^*.

7.2. Example 2: Stars Data

We continue with a small example with a more complicated structure than that of §7.1. The data are taken from Rousseeuw and Leroy (1987, p.27) and have been much used to illustrate the properties of various forms of robust regression. They form part of a Hertzsprung-Russell diagram of stars. This log-log plot has the effective surface temperature of the star as the explanatory variable

and (logged) light intensity as the response. A typical plot has around 30,000 stars which fall into groups including “the main sequence”, “white dwarves” and “giants” of several kinds. However, in our example, there are only 47 observations.

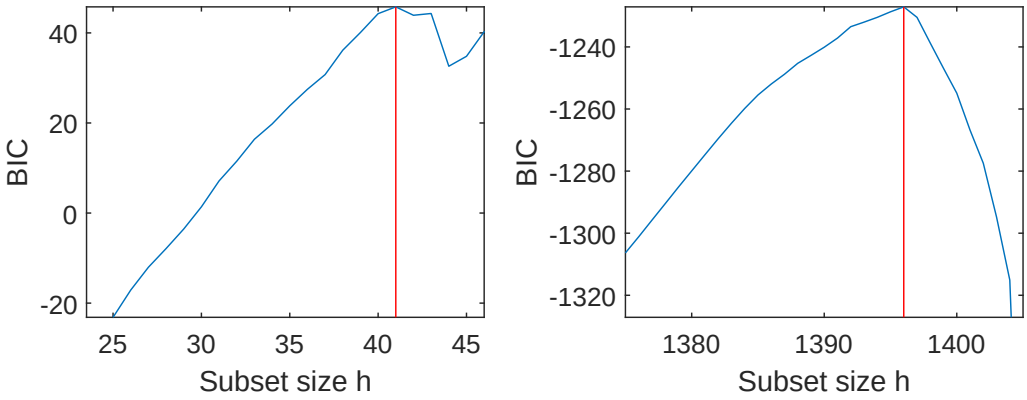

Figure 8: Monitoring plots of BICW using the FS. Left-hand panel, stars data; right-hand panel, transformed balance sheet data

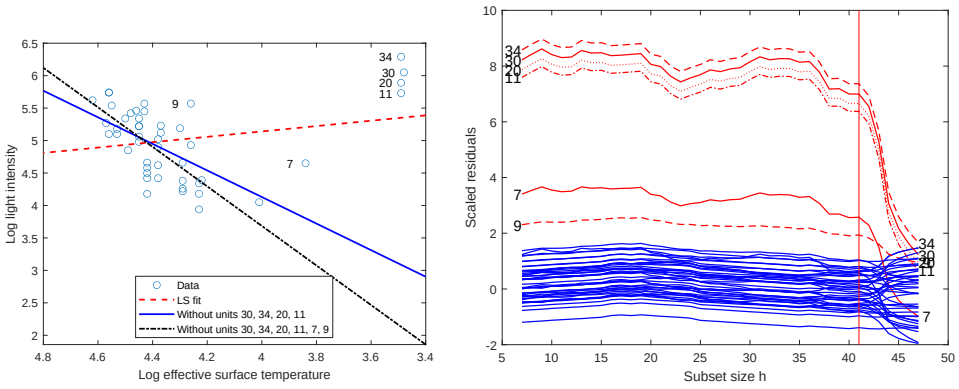

Figure 9: Stars data. Left-hand panel; least squares regression lines: dashed line, all observations; continuous line, four extreme outliers deleted; fine-dashed line, six observations deleted. Right-hand panel, monitoring plot of scaled FS residuals; $h^{\dagger} = 41$

The left-hand panel of Figure 8 shows that the monitoring plot of BICW for the FS again increases almost linearly with h, with a peak at $h = 41$. When LTS is monitored in steps of h corresponding to the addition of a single observation, the peak is at a bdp of 0.14, when the same six observations are deleted. In both plots the peak is well defined but is followed, as bdp decreases, by a small

decline and then an increase, a more complicated shape than we have seen before, which is caused by masking. The structure of the data and the nature of the outliers is clarified by the FS analysis. The left-hand panel of Figure 9 shows a scatterplot of the data and three different least squares regression line. There are four obvious outliers at x values remote from the rest of the data. These four observations cause the regression line to have a slight positive gradient. When they are deleted the slope of the regression line is negative, becoming more so when two further outliers are deleted. This is the regression line produced by LTS and the FS combined with monitoring BICW. The right-hand panel of the figure is a forward plot of the scaled FS residuals. The deletion of 6 outliers follows from the value of 41 for $h^\dagger$, at which point the groups of extreme and intermediate outliers are well separated. The transition between the regions of robust and non-robust estimation is clear.

We now turn to MM-estimation. The monitoring plot of BICW_ρ gives a value of 0.82 for $eff^\dagger$ (it is 0.86 for the two hard-trimming methods). For S-estimation the value of $d^\dagger$ is 0.25, appreciably greater than that from hard trimming. The monitoring plots of BICW_ρ for both MM and S-estimation show a similar shape to that for the FS; a peak followed by a decline and then an increase. To interpret these maximum values we look at monitoring plots of the weights. Those for MM-estimation are in the left-hand panel of Figure 10. This shows that the four extreme outliers are detected at an efficiency of 0.98. However, the weights for the two intermediate outliers decrease very slowly as the efficiency decreases, all six outliers receiving sufficiently small weights to be detected only when *eff* = 0.64. Just four outliers are detected at $eff^\dagger$.

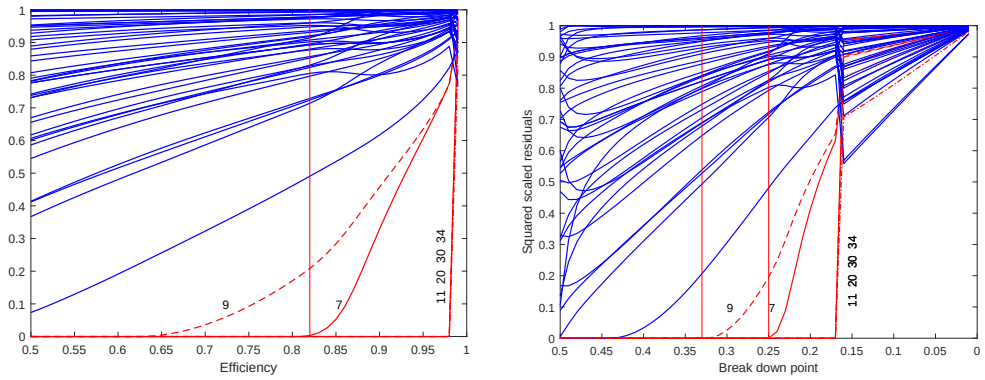

Figure 10: Stars data. Monitoring plot of weights $\hat{w}_i$. Left-hand panel MM-estimation; $eff^\dagger = 0.82$. Right-hand panel S-estimation with scale estimate $\hat{\sigma}_d$; $d^* = 0.33, d^\dagger = 0.25$

For S-estimation $d^\dagger$ = 0.25 and $d^* = 0.33$. The monitoring plot of the weights

for S-estimation in the right-hand panel of Figure 10 shows that the four extreme outliers have a weight ≤ 0.01 from a bdp of 0.17. At 0.25 ($d^{\dagger}$) five observations have small weights and for $d \leq 0.32$ six observations are detected as outlying. So d^{*} finds the same outliers as the hard trimming methods.

7.3. Example 3: Balance Sheet Data

Our final example has more observations and several explanatory variables. The data are taken from a larger set giving balance sheet information on limited liability companies. The response is profitability of individual firms in Italy. There are 998 observations with positive response and 407 with negative response, making 1,405 observations in all. There are five explanatory variables which are measures of financial properties of the firms, the two most important being the ratio of labour cost to value added and the ratio of tangible fixed assets to value added. The aim is to explain the profitability by regression on the five explanatory variables.

The data were introduced by Atkinson et al. (2021) who give further details. They show that the data need to be transformed to achieve approximate normality. Since 407 of the observations are negative, they used an extension of the transformation of Yeo and Johnson (2000). Atkinson et al. (2021) found that the positive observations should have a power transformation with parameter value 0.5, whereas the negative observations required a value of 1.5. We work throughout with this transformation.

The right-hand panel of Figure 8 shows the monitoring plot of BICW from the FS. This is similar in shape to those in Figure 5 for the two smaller sets of data. In this case there is slight curvature as the value of BICW increases with h, with a sharp peak. The maximum occurs when $h = 1396$. For LTS, again not shown, the maximum is at $d = 0.007$ when again $h = 1396$ (above $d = 0.01$ a finer grid of values in steps of 0.001 was used for monitoring). The indication is that there are only nine outliers, that is less than 1%. Figure 14 of Atkinson et al. (2021) has scatter plots of the residuals and shows the effect of outliers on the estimated parameters of the linear model.

For MM-estimation monitored over efficiency steps of 0.01, $\mathit{eff}^{\dagger} = 0.99$. Monitoring over a finer grid shows that the value of BIC_{ρ} continues to increase to *eff* = 0.999 before decreasing sharply when *eff* = 1. The left-hand panel of Figure 11 shows that only three observations have zero weight at $\mathit{eff}^{\dagger} = 0.99$. Although the figure suggests a group of nine residuals with small weights that seems separate from the other residuals, all nine do not have zero weight until *eff* has decreased to 0.95.

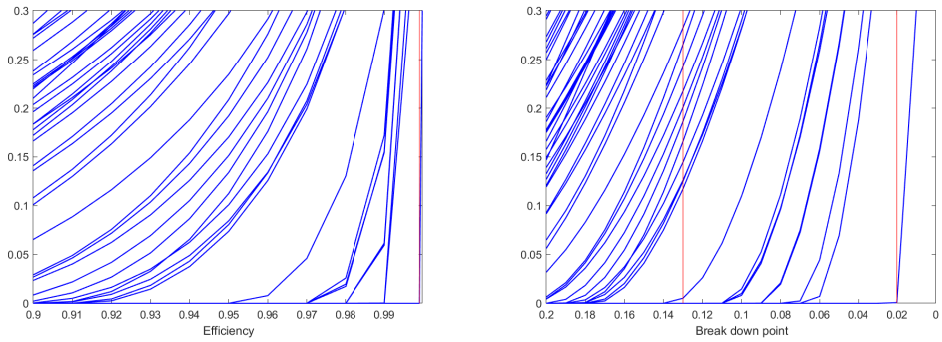

Figure 11: Balance sheet data. Zooms of monitoring plot of weights $\hat{w}_i$. Left-hand panel MM-estimation; $eff^\dagger = 0.99$. Right-hand panel S-estimation with scale estimate $\hat{\sigma}_d$; $d^* = 0.13, d^\dagger = 0.02$

The results for S-estimation are in the right-hand panel of the figure. For these data $d^\dagger = 0.02$. The figure shows that at this value of bdp only one observations has weight ≤ 0.01. The value of d^* is a much larger 0.13. At this value all the nine outliers are the only observations with small weights, so, for this example, S-estimation of this form agrees with hard trimming.

7.4. Summary of Analysis of Examples

Table 3: Number of outliers detected by five methods for the three examples of §7

Example		Illness	Stars	Balance Sheet
Method	Estimate			
LTS/FS	$n - h^\dagger$	3	6	9
MM	$eff^\dagger$	2	4	3
S	$d^\dagger$	3	4	1
S	d^*	3	6	9

Table 3 shows the number of outliers detected by the hard and soft trimming procedures of our paper when applied to the examples in this section. All, apart from MM, find the three introduced outliers in the contaminated illness data. S-estimation using the value d^* (13) agrees with the two hard trimming methods, although the parameter estimates are found with more trimming, that is a higher bdp, than those of the hard trimming methods. Both MM-estimation with $eff^\dagger$ and S-estimation using $d^\dagger$ appear, on this evidence, less reliable than hard trimming.

8. Comparisons and Extensions

In the appendix we describe two further forms of BIC for outlier detection using soft trimming. The starting point is the robust criterion for regression model selection of Maronna et al. (2019). This uses the quantity $2p$ of Akaike's AIC (Akaike, 1974) to penalize increasingly complex models. To form a similar BIC for outlier detection we replace this penalty with that of §3.1, that is $k(p, n) = p \log n$. We call the resulting robust criterion BICR and define it in equation (A.4). An expression for this criterion that is closer in form to BICW$_\rho$ is found by Taylor series linearisation of BICR. In (A.6) this is called BICL.

Table 4 provides a comparison of the performance of these two further forms of BIC when they are used with S-estimation in the analysis of the three examples of §7. The table gives the maximizing values of both $d^\dagger$ and d^*, together with, for reference, the corresponding results from hard trimming.

Table 4: Comparison of values of the bdp d maximizing three forms of BIC for analysis of the examples of §7 using S-estimation

Measure	Estimated Maximum	Contaminated Illness Data	Stars Data	Balance Sheet Data
BICW	LTS	0.07 ($h = 50$)	0.14 ($h = 41$)	0.007 ($h = 1396$)
BICW	FS	$h = 50$	$h = 41$	$h = 1396$
BICW$_\rho$	$d^\dagger$	0.14	0.25	0.02
BICL	$d^\dagger$	0.15	0.24	0.02
BICR	$d^\dagger$	0.24	0.33	0.03
BICW$_\rho$	d^*	0.23	0.33	0.128
BICL	d^*	0.23	0.33	0.128
BICR	d^*	0.23	0.35	0.133

The purpose of the data analyses in this paper is to identify outliers and to provide efficient parameter estimates for the non-outlying data. In the case of LTS and S-estimation this leads directly to finding the smallest bdp value at which the outliers are excluded from the analysis. The first two lines of Table 4 show that LTS and the FS find values of bdp smaller than those from the soft trimming procedures based on S-estimation. Within the soft trimming methods, the three forms of BIC provide similar values of $d^\dagger$ and d^*, except for the higher value of $d^\dagger$ from BICR for the contaminated illness data. We therefore, if soft trimming is required, suggest the use of BICW$_\rho$, the properties of which have been more thoroughly explored in this paper.

We also tested robustness to the choice of ρ function. As an alternative to the Tukey biweight we used the power divergence ρ function $\rho(u) = 1 - \exp(-\alpha u^2/2)$, where α is the parameter controlling breakdown point and efficiency. Plots such as Figure 5 of Riani et al. (2020) indicate that there is little difference in asymptotic efficiency and breakdown point between this ρ function and the Tukey biweight. The simulation results showed that BICW$_\rho$ continued to perform best although the results gave slightly larger values of $d^\dagger$ and d^* than those of Table 4. Tukey's biweight is to be preferred for soft trimming, but flexible hard trimming, either from the FS or monitoring LTS, is to be preferred.

9. Discussion

Flexible hard trimming combined with the information criterion BICW provides clear identification of outliers. For soft trimming our results indicate that BICW$_\rho$ is the preferred form of information criterion when S-estimation is used, which is to be preferred to MM-estimation.

The outliers identified by applying BICW to the results of LTS or the FS agree with those in our previous analyses using significance levels and the properties of order statistics. For the three examples the FS analyses are respectively given by Atkinson et al. (2021), Riani et al. (2014b) and Atkinson et al. (2020). Here, in contrast, we run a single search through the data and monitor the value of the appropriate BIC, the maximum indicating which observations are outliers. The complicated rule for determining significance of potential outliers is circumvented.

The calculations for the forward search only require updating a regression that starts from a small number of observations. This is computationally much simpler than the implementation of LTS we have used (see §6) in which an optimum solution has to be found for each h, as it also is than monitoring MM- or S-estimation, where numerical optimizations are required for each value of eff or d. For large data sets we can use the results of Torti et al. (2021) which extend the FS to moving forward by adding batches of $k > 1$ observations.

We have restricted our attention to the classical case of the Tukey-Huber contamination model with fewer than 50% contaminated observations. It is important that the FS can be adapted also to the case of more than 50% contamination, when this is physically meaningful, for example in clustering (Cerioli et al., 2019).

The forms of BIC we have developed for the automatic detection of outliers do not include significance tests. However, significance testing may be important once the outliers, if any, have been detected (Cerioli and Farcomeni, 2011). The

inclusion of least squares with $h = n$ in the monitoring ensures that the models we consider include one in which outliers are not deleted. Both diagnostic plots, such as those of §7, and significance testing will be part of the determination of the importance of any outliers; they may be random, their effect being to reduce the accuracy of conclusions drawn from the data or they might indicate, in a medical context, a group of patients that responds differently to treatment. Cox (2020) dissects forms of statistical significance appropriate to a variety of data analytical tasks. Finally, we see a place for our results in automatic model selection as an extension to the use of conventional BIC to cover the presence of outliers.

Acknowledgements and Code

We are very grateful to the referees for their engagement with our proposal and for their thorough and insightful reports. Revisions in the light of their comments and suggestions have greatly improved our paper.

This research benefits from the High Performance Computing (HPC) facility of the University of Parma. We also acknowledge financial support from the "Statistics for fraud detection, with applications to trade data and financial statements" project of the University of Parma. All the calculations in this paper have used the Flexible Statistics and Data Analysis (FSDA) MATLAB toolbox, which is freely downloadable from the file exchange of Mathworks at the web address
https://www.mathworks.com/matlabcentral/fileexchange/72999-fsda
or from github at the web address https://uniprjrc.github.io/FSDA/. More specifically, the monitoring of S residuals and weights to obtain, for example, Figure 7 uses function *Sregeda.m*. For MM-estimation we use the function *MMregeda.m*. The associated HTML documentation can be found after installing the toolbox or directly from the web address
http://rosa.unipr.it/FSDA/Sregeda.html. The monitoring of LTS residuals to obtain, for example, the left-hand panel of Figure 5 is based on the repeated call of function *LXS.m* (http://rosa.unipr.it/FSDA/LXS.html), while the monitoring of least squares residuals to obtain the right-hand panel of Figure 5 is based on the use of function *FSRbsb.m* (http://rosa.unipr.it/FSDA/FSRbsb.html). Finally, all the datasets described in this paper are contained in the regression dataset section of FSDA or in folder
https://github.com/UniprJRC/FSDA/tree/master/datasets/regression of the associated github repository.

Appendix A. Appendix: Other Forms of Robust BIC for Soft Trimming

Maronna et al. (2019, p.137) derive a robust form of C_p for model selection. We now extend their approach to provide two further versions of BIC for soft downweighting.

Akaike's non-robust Final Prediction Error, FPE, for least squares regression is written in (5.37) of Maronna et al. (2019) as

$$\text{FPE} = \frac{1}{n}\sum_{i=1}^{n} e_i^2 \left(1 + \frac{2p}{n}\right), \tag{A.1}$$

where $e_i = y_i - x_i^T\hat{\beta}$. Then $R(\hat{\beta}) = \sum e_i^2$ and FPE $= \{R(\hat{\beta})/n\}(1 + 2p/n)$.

Maronna et al. (2019) obtain a robust criterion for regression model selection starting from (A.1) in which least squares is replaced, in our case, by S-estimation. The corresponding robust version of FPE (A.1) is

$$\text{RFPE} = \frac{1}{n}\sum \rho\left(\frac{e_i}{\hat{\sigma}}\right) + \frac{p}{n}\frac{\hat{A}}{\hat{B}}, \tag{A.2}$$

where

$$\hat{A} = \frac{1}{n}\sum_{i=1}^{n}\left\{\psi\left(\frac{e_i}{\hat{\sigma}}\right)\right\}^2 \quad \text{and} \quad \hat{B} = \frac{1}{n}\sum_{i=1}^{n}\psi'\left(\frac{e_i}{\hat{\sigma}}\right),$$

with $\psi(u) = \rho'(u)$.

The RFPE criterion (A.2) is for the comparison of regression models. We take the form of the linear model as given and are interested in monitoring the behaviour of BIC as the value d of the breakdown point of the robust method varies. To develop this novel form of BIC we rewrite (A.2) with the BIC penalty and obtain

$$\frac{1}{n}\sum \rho\left(\frac{e_i}{\hat{\sigma}}\right) + \frac{\hat{A}}{2\hat{B}}\frac{p\log n}{n}. \tag{A.3}$$

Multiplication by $-n$ yields the robust BIC

$$\text{BICR} = -\sum \rho\left(\frac{e_i}{\hat{\sigma}}\right) - \frac{\hat{A}}{2\hat{B}}\{p + \sum_{i=1}^{n}(1 - w_i)\}\log n. \tag{A.4}$$

To see the relationship between BICR and the BIC for hard trimming in (7) we rewrite (A.3) as

$$\overline{\rho}\left\{1 + \frac{\hat{A}}{\hat{B}}\frac{1}{2\overline{\rho}}\frac{p\log n}{n}\right\}, \tag{A.5}$$

where $\overline{\rho} = \sum \rho(e_i/\hat{\sigma})/n$. Taylor expansion of the logarithm of (A.5) yields

$$\log \overline{\rho} + \frac{\hat{A}}{\hat{B}} \frac{1}{2\overline{\rho}} \frac{p \log n}{n}.$$

Multiplication by $-n$ and inclusion of the penalty for downweighting used in (A.4) gives the approximate BIC

$$\text{BICL} = -n \log \overline{\rho} - \frac{0.5}{\overline{\rho}} \frac{\hat{A}}{\hat{B}} \left\{ p + \sum_{i=1}^{n} (1 - w_i) \right\} \log n. \tag{A.6}$$

The relationship between BICL and BICR depends on the Taylor expansion. For asymptotic equivalence we require that the quantity $(\hat{A}/\hat{B})\{(p \log n)/n\}$ decrease with n. Since the asymptotic value of $\hat{A}/\hat{B}$ is finite (Maronna et al., 2019, Chapter 10), we only require that the limit of $(p \log n)/n \to 0$ with n, which it does for models in which the number of parameters is not a function of n. Observations which come in blocks of fixed sizes, each of which introduces a new parameter, such as mixed pairs (Cox and Hinkley, 1974, p.17), require special treatment. A model with random effects for blocks is one possibility.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19, 716–723.

Atkinson, A.C., 1985. Plots, Transformations, and Regression. Oxford University Press, Oxford.

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

Atkinson, A.C., Riani, M., Corbellini, A., 2020. The analysis of transformations for profit-and-loss data. Applied Statistics 69, 251–275. DOI: https://doi.org/10.1111/rssc.12389.

Atkinson, A.C., Riani, M., Corbellini, A., 2021. The Box-Cox transformation: review and extensions. Statistical Science 36, 239–255. DOI: 10.1214/20-STS778.

Beaton, A.E., Tukey, J.W., 1974. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. Technometrics 16, 147–185.

Bhat, H.S., Kumar, N., 2010. On the derivation of the Bayesian Information Criterion. Technical Report. University of California. Merced CA 95343.

Buja, A., Rolke, W., 2003. Calibration for Simultaneity: (Re)Sampling Methods for simultaneous inference with applications to function estimation and Functional data. Technical Report. The Wharton School. University of Pennsylvania.

Cerioli, A., Farcomeni, A., 2011. Error rates for multivariate outlier detection. Computational Statistics and Data Analysis 55, 544–553.

Cerioli, A., Farcomeni, A., Riani, M., 2014. Strong consistency and robustness of the Forward Search estimator of multivariate location and scatter. Journal of Multivariate Analysis 126, 167–183.

Cerioli, A., Farcomeni, A., Riani, M., 2019. Wild adaptive trimming for robust estimation and cluster analysis. Scandinavian Journal of Statistics 46, 235—256.

Claeskens, G., Hjort, N.L., 2008. Model Selection and Model Averaging. Cambridge University Press, Cambridge.

Cook, R.D., Weisberg, S., 1982. Residuals and Influence in Regression. Chapman and Hall, London.

Cox, D.R., 2020. Statistical significance. Annual Review of Statistics and Its Application 7, 1–10. Doi: 10.1146/annurev-statistics-031219-041051.

Cox, D.R., Hinkley, D.V., 1974. Theoretical Statistics. Chapman and Hall, London.

Farcomeni, A., Greco, L., 2015. Robust Methods for Data Reduction. Chapman & Hall/CRC Press, Boca Raton, FL.

Hampel, F., Ronchetti, E.M., Rousseeuw, P., Stahel, W.A., 1986. Robust Statistics. Wiley, New York.

Hampel, F.R., 1975. Beyond location parameters: robust concepts and methods. Bulletin of the International Statistical Institute 46, 375–382.

Insolia, L., Kenney, A., Chiaromonte, F., Felici, G., 2020. Simultaneous Feature Selection and Outlier Detection with Optimality Guarantees. Technical Report. Scuola Normale Superiore di Pisa. Https://www.researchgate.net/publication/342915458.

Johansen, S., Nielsen, B., 2016. Analysis of the Forward Search using some new results for martingales and empirical processes. Bernoulli 21, 1131–1183.

Johnson, N.L., Kotz, S., Balakrishnan, N., 1994. Continuous Univariate Distributions - 1, 2nd edition. Wiley, New York.

Kleinbaum, D.G., Kupper, L., 1978. Applied Regression Analysis and Other Multivariable Methods. Duxbury, Boston, Mass.

Maronna, R.A., Martin, R.D., Yohai, V.J., 2006. Robust Statistics: Theory and Methods. Wiley, Chichester.

Maronna, R.A., Martin, R.D., Yohai, V.J., 2019. Robust Statistics: Theory and Methods (with R), 2nd edn. Wiley, Chichester.

Olive, D.J., 2020. Robust statistics. Manuscript not really ready. Revisions are ongoing. Online text available at `http://parker.ad.siu.edu/Olive/robbook.htm`.

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., Atkinson, A.C., Corbellini, A., Perrotta, D., 2020. Robust regression with density power divergence: theory, comparisons and data analysis. Entropy 22. Doi:10.3390/e22040399.

Riani, M., Cerioli, A., Atkinson, A.C., Perrotta, D., 2014a. Monitoring robust regression. Electronic Journal of Statistics 8, 642–673.

Riani, M., Cerioli, A., Torti, F., 2014b. On consistency factors and efficiency of robust S-estimators. TEST 23, 356–387.

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. Wiley, New York.

Rousseeuw, P.J., Van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212–223.

Rousseeuw, P.J., Yohai, V.J., 1984. Robust regression by means of S-estimators, in: Franke, J.,

Härdle, W., Martin, R.D. (Eds.), Robust and Nonlinear Time Series Analysis: Lecture Notes in Statistics 26. Springer Verlag, New York, pp. 256–272.
Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6, 461–464.
Tallis, G.M., 1963. Elliptical and radial truncation in normal samples. Annals of Mathematical Statistics 34, 940–944.
Torti, F., Corbellini, A., Atkinson, A.C., 2021. fsdaSAS: a package for robust regression for very large datasets including the Batch Forward Search. Stats 4, 327–347.
Yeo, I.K., Johnson, R.A., 2000. A new family of power transformations to improve normality or symmetry. Biometrika 87, 954–959.
Yohai, V.J., 1987. High breakdown-point and high efficiency estimates for regression. The Annals of Statistics 15, 642–656.