

Robust Model Selection Using Fast and Robust Bootstrap

Matias Salibian-Barrera¹

University of British Columbia, Department of Statistics, 333 - 3656 Agricultural Road, Vancouver, BC, V6T 1Z4, Canada.

Stefan Van Aelst^{2,*}

Ghent University, Department of Applied Mathematics and Computer Science, Krijgslaan 281 S9, B-9000 Gent, Belgium.

Abstract

Robust model selection procedures control the undue influence that outliers can have on the selection criteria by using both robust point estimators and a bounded loss function when measuring either the goodness-of-fit or the expected prediction error of each model. Furthermore, to avoid favoring over-fitting models, these two measures can be combined with a penalty term for the size of the model. The expected prediction error conditional on the observed data may be estimated using the bootstrap. However, bootstrapping robust estimators becomes extremely time consuming on moderate to high dimensional data sets. It is shown that the expected prediction error can be estimated using a very fast and robust bootstrap method, and that this approach yields a consistent model selection method that is computationally feasible even for a relatively large number of covariates. Moreover, as opposed to other bootstrap methods, this proposal avoids the numerical problems associated with the small bootstrap samples required to obtain consistent model

selection criteria. The finite-sample performance of the fast and robust bootstrap model selection method is investigated through a simulation study while its feasibility and good performance on moderately large regression models are illustrated on several real data examples.

Key words: Model selection, robustness, bootstrap.

1991 MSC: 62J05, 62F35

1 Introduction

Model selection consists of choosing a model from a set of competing ones. The procedure generally involves fitting the different models, and then comparing these using a numerical summary of their goodness-of-fit, prediction properties, or a combination of both. To avoid selecting a model that over-fits the data, a term that penalizes models with a larger number of parameters is sometimes included as well. We are particularly concerned with the case where data quality (or the assumptions regarding error distributions required by the estimation and model selection procedures) might be questionable, and thus it is of interest to use robust estimators. Therefore, in this paper we focus on model selection for linear models using robust regression estimators.

It is well known that model selection methods which rely on least squares

* Corresponding author. Tel: +32-9-264-49-08; fax: +32-9-264-49-95.

Email address: Stefan.VanAelst@UGent.be (Stefan Van Aelst).

¹ Research supported by a Discovery Research Grant from NSERC

² Research supported by a grant of the Fund for Scientific Research-Flanders (FWO-Vlaanderen) and by IAP research network grant nr. P6/03 of the Belgian government (Belgian Science Policy)

or other likelihood-type approaches (e.g. AIC (Akaike, 1970), Mallows' C_p (Mallows, 1973), and BIC (Schwarz, 1978)) may be severely affected by a small proportion of atypical observations in the data. These “outliers” need not be “extreme” (i.e. have “large” values), but they do not follow the model that applies to the majority of the sample. Other selection methods are based on finding the model that minimizes the expected squared prediction loss, which needs to be estimated generally by cross-validation (Shao, 1993) or the bootstrap (Shao, 1996). However, when the prediction loss is measured with an unbounded loss function, these criteria are also susceptible to the potentially damaging effect of slight departures from the model under which they were derived. Moreover, this can happen even when robust estimators are used to fit each model (see e.g. Ronchetti and Staudte, 1994; Wisnowski, Simpson, Montgomery and Runger, 2003).

Robust model selection methods for linear regression have received some attention in the literature recently (see, e.g. Ronchetti 1985, 1997; Ronchetti and Staudte, 1994; Sommer and Staudte, 1995; Ronchetti, Field and Blanchart, 1997; Qian and Künsch, 1998; Agostinelli, 2002; and Maronna, Martin and Yohai, 2006). These proposals are based on robust versions of classical selection criteria (e.g. robust C_p , robust final prediction error, etc.). Müller and Welsh (2005) proposed a selection criterion that, for each model, combines a measure of goodness-of-fit, a penalty term for the number of parameters and the expected prediction error, conditional on the observed sample. To obtain a model selection criterion that is robust against outliers both the goodness-of-fit and the expected prediction error are computed using a bounded loss function (as in Ronchetti and Staudte, 1994). Intuitively, replacing the squared loss by a bounded function limits the “cost” of not adjusting outlying obser-

vations, and hence, a fit that approximates well most of the observations but poorly a small fraction of them may be preferred over one that only produces a mediocre fit to all points. As in Shao (1996), the conditional expected prediction error is estimated using the bootstrap (Efron, 1979). Specifically, the robust linear regression estimators are re-computed on several hundred bootstrap samples, and the average loss is calculated for each such a sample. These mean losses are then averaged over all the bootstrap samples.

Unfortunately, bootstrapping robust estimators when the data may contain outliers presents some difficulties. One of them is the important negative effect of bootstrap samples where the proportion of outliers is much higher than in the original data set. Another one is the high computational complexity of regression estimators with good robustness and efficiency properties. Müller and Welsh (2005) address the first problem by proposing to use a stratified bootstrap. In this approach, bootstrap samples are constructed so that the distribution of the residuals in each bootstrap sample reflects the one observed in the original data set. This strategy seems to avoid the first problem well in practice. However, the high computational cost of re-computing robust regression estimators still limits this approach in practice as it rapidly becomes infeasible for moderately large models. To illustrate this, in Table 5 we list the CPU time needed to compute the stratified bootstrap estimator of the expected prediction error based on 1000 bootstrap samples of size 200 for different numbers of covariates p . From this table it can be seen that performing a complete model selection analysis on a dataset like the Ozone data with $p = 45$ (discussed in Section 6.1) would take more than 15 days of continuous computing time.

Furthermore, to obtain a consistent selection criterion (in the sense that the

probability of selecting the correct model tends to 1 as the sample size increases) Shao (1996) showed that the size of the bootstrap samples used to evaluate the expected prediction error should grow slower than the sample size. More specifically, if n denotes the number of observations in the data, one should use bootstrap samples of size $m = o(n)$. Müller and Welsh (2005) require even smaller bootstrap samples $m = o(\sqrt{n})$, mainly because the loss function in their selection criterion need not be the one associated with the robust estimator used to fit the models. Shao (1996) already indicates that, when using least-squares regression estimators, the practical choice of $m = o(n)$ needs to take into consideration the increased variability of the re-computed (bootstrapped) estimator, which may induce extra uncertainty in the estimation of the expected prediction error. This problem is even more delicate in the case of robust regression estimators. Serious computational problems can arise when the sample size is relatively small with respect to the number of covariates. The main difficulty is that the complex optimization problems that need to be solved involve randomly drawing a large number of sub-samples of size $p + 1$ with nonsingular design matrices. If the size of the bootstrap sample m is close to p , then it may be difficult to find a reasonable number of nonsingular sub-samples of size $p + 1$. Moreover, note that some of these m data points may actually be repetitions of a single observation in the original data set, which exacerbates this problem even further. We illustrate this with a simple example. Suppose that $n = 50$ and $p = 5$. The selection criterion of Müller and Welsh (2005) requires $m < \sqrt{50}$. However, a high breakdown point efficient regression estimator such as the MM-estimator (suggested by Müller and Welsh) would be extremely difficult to compute with a sample of size $m = 6$, say, and $p = 5$, let alone one where some points may be repeated.

In this paper we show that robust and feasible model selection methods for linear regression estimators can be obtained by using the fast and robust bootstrap of Salibian-Barrera and Zamar (2002). In particular, we show that using the fast and robust bootstrap to estimate the conditional expected prediction loss in the selection criteria proposed by either Shao (1996) or Müller and Welsh (2005) provides consistent model selection criteria (in the sense of Shao (1996)). We also show that if the same loss function is used in the selection criterion and in the estimation step, we can take larger bootstrap samples of order $m = o(n)$, which naturally provide more stable re-computed estimators. Moreover, note that the fast and robust bootstrap does not require the full recalculation of the estimator, so we expect the estimated conditional expected prediction error to be more accurate. The gain in speed can be illustrated by noting that the full model selection analysis mentioned above that would have taken approximately 15 days of CPU time with the stratified bootstrap, took just over 4 hours when using the fast and robust bootstrap. This difference is more pronounced for larger values of p (see the examples in Section 6).

Given a set of p covariates, there exist efficient algorithms to compare all possible sub-models of these p covariates if the criterion is based on the residual sum of squares (RSS), see e.g. Furnival and Wilson (1974); Gatu and Kontoghiorghes (2006); Hofmann, Gatu, and Kontoghiorghes (2007). However, for robust regression, unless p is small, it is prohibitively time consuming to compare all possible sub-models of k covariates with $1 \leq k \leq p$. One commonly used strategy is backward elimination: (i) fit all models of size $p - 1$ and select the one with the best value of the selection criterion; (ii) fit all submodels of size $p - 2$ that are subsets of the above optimal model of size $p - 1$, and select the best one according to the selection criterion; (iii) continue in this way until

$p = 0$ (a model with only an intercept term). Finally, from these $p + 1$ “optimal” models, select the one with the smallest selection criterion. We applied the robust model selection criteria using the fast and robust bootstrap to relatively large real datasets (models with 14, 45, and 65 explanatory variables) and found it to have both a good performance and a reasonable computation time. For really high dimensional problems (e.g. $p > n$), a forward selection approach can be used as in (Khan, Van Aelst, and Zamar 2007a,b).

The rest of the paper is organized as follows. Section 2 contains the definition of MM-estimators for regression, which are the robust estimators used in this paper, and the criteria we use for robust model selection. The fast and robust bootstrap procedure of Salibián-Barrera and Zamar (2002) is reviewed briefly in Section 3 while the consistency of the robust model selection procedure is examined in Section 4. The results of our simulation studies are presented in Section 5 while Section 6 illustrates the robustness and feasibility of the method on some real data examples. Finally, Section 7 summarizes our conclusions while the Appendix contains the proofs.

2 Definitions

Let $(y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n)$ be n independent observations, where $y_i \in \mathbb{R}$ and $\mathbf{x}_i \in \mathbb{R}^p$. Let α denote a subset of p_α indices from the set $\{1, 2, \dots, p\}$ and let $\mathbf{x}_{\alpha i} \in \mathbb{R}^{p_\alpha}$ be the corresponding coordinates of \mathbf{x}_i , $i = 1, \dots, n$. The linear model corresponding to model α is

$$y_i = \mathbf{x}'_{\alpha i} \boldsymbol{\beta}_\alpha + \sigma_\alpha \epsilon_{\alpha i}, \quad i = 1, \dots, n, \quad (1)$$

where $\boldsymbol{\beta}_\alpha \in \mathbb{R}^{p_\alpha}$, $\sigma_\alpha > 0$ and the errors $\epsilon_{\alpha i}$ are assumed to have location zero and spread (scale) one.

Given a collection \mathcal{A} of candidate models, we are interested in selecting one of them based on the properties of the corresponding fit. Note that we only consider models with intercept. To fit the models, we use robust MM-estimators for linear regression (Yohai, 1987) which combine good robustness properties with high efficiency if there are no outliers present in the data. These estimators are based on two loss functions ρ_0 and ρ_1 , which determine the breakdown point (see e.g. Maronna *et al.* 2006) and the efficiency of the estimator, respectively. More precisely, for the full linear model (1) with $\alpha = \{1, 2, \dots, p\}$, the MM-estimator $\hat{\boldsymbol{\beta}}_n$ satisfies

$$\frac{1}{n} \sum_{i=1}^n \rho_1' \left(\frac{y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}_n}{\hat{\sigma}_n} \right) \mathbf{x}_i = \mathbf{0}, \quad (2)$$

where $\rho_1'(u)$ is the derivative of the loss function ρ_1 and $\hat{\sigma}_n$ is an S-estimate of scale (Rousseeuw and Yohai, 1984). Hence, $\hat{\sigma}_n$ minimizes the M-scale $\hat{\sigma}_n(\boldsymbol{\beta})$ which is implicitly defined for each $\boldsymbol{\beta} \in \mathbb{R}^p$ by

$$\frac{1}{n} \sum_{i=1}^n \rho_0 \left(\frac{y_i - \mathbf{x}_i' \boldsymbol{\beta}}{\hat{\sigma}_n(\boldsymbol{\beta})} \right) = b, \quad (3)$$

where $b \in (0, 1)$ is a tuning constant that determines the breakdown point of $\hat{\sigma}_n$ which equals $\min(b, 1 - b)$. The associated regression S-estimate $\tilde{\boldsymbol{\beta}}_n$ is the solution

$$\tilde{\boldsymbol{\beta}}_n = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \hat{\sigma}_n(\boldsymbol{\beta}), \quad (4)$$

and is used as an initial value for the iterations that determine $\hat{\boldsymbol{\beta}}_n$ in (2).

A widely used family of loss functions is Tukey's biweight family

$$\rho_c(t) = \begin{cases} 3(t/c)^2 - 3(t/c)^4 + (t/c)^6 & \text{if } |t| \leq c \\ 1 & \text{if } |t| > c, \end{cases} \quad (5)$$

where $c > 0$ is a fixed tuning constant. For the S-estimator, the choice $c = 1.54764$ in the loss function ρ_0 together with $b = 1/2$ in (3) yields a 50% breakdown-point consistent scale-estimator for normally distributed errors. For the M-estimator, the choice $c = 4.685061$ in the loss function ρ_1 yields a 95%-efficient regression estimator when the errors follow a normal distribution.

Model selection usually involves comparing estimates obtained by fitting different models. We assume that all models $\alpha \in \mathcal{A}$ in the comparison are sub-models of a "full" model which can be used to obtain a valid estimate of the error scale. In what follows, $\hat{\sigma}_n$ will denote the S-scale estimate computed with the "full" model as described above. For each model $\alpha \in \mathcal{A}$, the regression estimator $\hat{\beta}_{\alpha,n}$ solves

$$\frac{1}{n} \sum_{i=1}^n \rho_1' \left(\frac{y_i - \mathbf{x}_{\alpha i}' \hat{\beta}_{\alpha,n}}{\hat{\sigma}_n} \right) \mathbf{x}_{\alpha i} = \mathbf{0}. \quad (6)$$

The solution to this equation can be found using an iterative re-weighted least squares algorithm starting from the S-regression estimate $\tilde{\beta}_n$ computed for the full model.

A good model should fit the data reasonably well and at the same time be able to predict future observations accurately. Let $\mathbf{y} = (y_1, \dots, y_n)'$ denote the vector of responses and let $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)'$ be the design matrix. For a given loss function $\rho : \mathbb{R} \rightarrow \mathbb{R}_+$, the expected prediction error of a model α

conditional on the observed data can be measured by

$$M^{\text{pe}}(\alpha) = \frac{\sigma^2}{n} E \left[\sum_{i=1}^n \rho \left(\frac{z_i - \mathbf{x}'_{\alpha i} \hat{\boldsymbol{\beta}}_{\alpha}}{\sigma} \right) \middle| \mathbf{y}, \mathbf{X} \right], \quad (7)$$

where $\mathbf{z} = (z_1, \dots, z_n)'$ is a vector of future responses at \mathbf{X} , independent of \mathbf{y} , and σ is the scale of the error distribution at the “full” model. This measure of prediction error was first considered as a selection criterion by Shao (1996) in the context of least squares regression using the loss function $\rho(t) = t^2/2$.

Similarly, the goodness of fit of a particular model α can be measured by

$$\frac{\sigma^2}{n} E \left[\sum_{i=1}^n \rho \left(\frac{y_i - \mathbf{x}'_{\alpha i} \hat{\boldsymbol{\beta}}_{\alpha}}{\sigma} \right) \right].$$

Since, additionally, parsimonious models are typically preferred over more complex ones, Müller and Welsh (2005) introduced the following model selection criterion:

$$M^{\text{ppe}}(\alpha) = \frac{\sigma^2}{n} \left\{ E \left[\sum_{i=1}^n \rho \left(\frac{y_i - \mathbf{x}'_{\alpha i} \hat{\boldsymbol{\beta}}_{\alpha}}{\sigma} \right) \right] + \delta(n) p_{\alpha} \right\} + M^{\text{pe}}(\alpha), \quad (8)$$

where $\delta(n) \rightarrow \infty$ as $n \rightarrow \infty$ and $\delta(n)/n \rightarrow 0$ as $n \rightarrow \infty$. These conditions are satisfied by the choice $\delta(n) = \log(n)$. Note that the first term on the right-hand side measures the quality of the fit for the observed sample data, the second term penalizes complexity which expresses a preference for smaller, simpler models, and the last term measures the expected prediction error as before.

Among the models α being considered, we wish to select the one that minimizes either $M^{\text{pe}}(\alpha)$ or $M^{\text{ppe}}(\alpha)$ in (7) and (8) respectively. Since both selection criteria involve the unknown distribution of the data, they are estimated by

$$M_{m,n}^{\text{pe}}(\alpha) = \frac{\hat{\sigma}_n^2}{n} E_* \left[\sum_{i=1}^n \rho \left(\frac{y_i - \mathbf{x}'_{\alpha i} \hat{\boldsymbol{\beta}}_{\alpha,n}}{\hat{\sigma}_n} \right) \middle| \mathbf{y}, \mathbf{X} \right], \quad (9)$$

$$M_{m,n}^{\text{ppe}}(\alpha) = \frac{\hat{\sigma}_n^2}{n} \left\{ \sum_{i=1}^n \rho \left(\frac{y_i - \mathbf{x}'_{\alpha i} \hat{\boldsymbol{\beta}}_{\alpha,n}}{\hat{\sigma}_n} \right) + \delta(n) p_\alpha \right\} + M_{m,n}^{\text{pe}}(\alpha), \quad (10)$$

respectively, where E_* is the bootstrap estimator of the corresponding expected value in (7). Based on a sample of size n , and using bootstrap samples of size m for E_* , we select the model $\alpha \in \mathcal{A}$ that minimizes $M_{m,n}^{\text{pe}}(\alpha)$ or $M_{m,n}^{\text{ppe}}(\alpha)$, i.e.

$$\hat{\alpha}_{m,n}^{\text{pe}} = \arg \min_{\alpha \in \mathcal{A}} M_{m,n}^{\text{pe}}(\alpha), \quad (11)$$

$$\hat{\alpha}_{m,n}^{\text{ppe}} = \arg \min_{\alpha \in \mathcal{A}} M_{m,n}^{\text{ppe}}(\alpha). \quad (12)$$

3 Fast and robust bootstrap

A practical problem with the estimators $M_{m,n}^{\text{pe}}(\alpha)$ and $M_{m,n}^{\text{ppe}}(\alpha)$ in (9)-(10) is that they involve bootstrapping a robust regression estimator. It is easy to see that bootstrapping robust estimators with potentially contaminated data may present some practical difficulties. In particular, two problems that can arise are: (a) robust estimates with good properties are computationally very demanding, especially in moderate to large dimensions, even if efficient algorithms such as the fast S-algorithm of Salibián-Barrera and Yohai (2006) are used; and (b) an unduly large proportion of outliers might enter a significant number of bootstrap samples, upsetting the tails of our distribution estimate. Note that problem (b) is present regardless of the robustness properties of the estimator being bootstrapped.

To solve problem (b) above, Müller and Welsh (2005) proposed to use a stratified bootstrap in order to have the bootstrap samples reflect more closely

the occurrence of outliers in the original sample. Unfortunately, this approach does not address the heavy computational cost of high-breakdown point, robust regression estimators, and hence limits in practice the accuracy that can be achieved with the estimator $M_{m,n}^{\text{pe}}(\alpha)$ or $M_{m,n}^{\text{ppe}}(\alpha)$.

Recently, Salibian-Barrera and Zamar (2002) proposed a bootstrap method to estimate the distribution of regression MM-estimates. This method, which we call Fast and Robust Bootstrap (FRB), is easy to compute and resistant to the presence of outliers in the sample. In other words, it solves both problems (a) and (b) above. The FRB procedure has also been used successfully in multivariate models (Van Aelst and Willems, 2005; Salibian-Barrera, Van Aelst, and Willems 2006, 2008). In this paper we investigate the performance of selection procedures (11) and (12) when the expected value E_* is estimated by FRB based on bootstrap samples of size m .

We first explain in detail the FRB procedure for the estimator $\hat{\beta}_{\alpha,n}$ in (6) based on bootstrap samples of size $m \leq n$. Note that $\hat{\beta}_{\alpha,n}$ can be represented as a weighted least squares fit. Define the weights $\omega_{\alpha i}$ as

$$\omega_{\alpha i} = \rho'_1(r_{\alpha i}/\hat{\sigma}_n)/r_{\alpha i}, \quad (13)$$

for $i = 1, \dots, n$, where $r_{\alpha i} = y_i - \mathbf{x}'_{\alpha i} \hat{\beta}_{\alpha,n}$. Then, the solution to (6) can be rewritten as

$$\hat{\beta}_{\alpha,n} = \left[\sum_{i=1}^n \omega_{\alpha i} \mathbf{x}_{\alpha i} \mathbf{x}'_{\alpha i} \right]^{-1} \sum_{i=1}^n \omega_{\alpha i} \mathbf{x}_{\alpha i} y_i, \quad (14)$$

Let (y_i^*, \mathbf{x}_i^*) , $i = 1, \dots, m$ be a bootstrap sample of size $m \leq n$ from the observations. Define the random vector $\hat{\beta}_{\alpha,m}^*$ by

$$\hat{\beta}_{\alpha,m}^* = \left[\sum_{i=1}^m \omega_{\alpha i}^* \mathbf{x}_{\alpha i}^* \mathbf{x}'_{\alpha i} \right]^{-1} \sum_{i=1}^m \omega_{\alpha i}^* \mathbf{x}_{\alpha i}^* y_i^*, \quad (15)$$

where $\omega_{\alpha i}^* = \rho_1'(r_{\alpha i}^*/\hat{\sigma}_n)/r_{\alpha i}^*$, $r_{\alpha i}^* = y_i^* - \mathbf{x}_{\alpha i}^{*'}\hat{\boldsymbol{\beta}}_{\alpha, n}$, for $1 \leq i \leq m$. Note that $\hat{\boldsymbol{\beta}}_{\alpha, n}$ and $\hat{\sigma}_n$ are not re-calculated from each bootstrap sample $\{(y_i^*, \mathbf{x}_i^*); i = 1, \dots, m\}$.

We now apply a linear correction to the estimates obtained in (15). Intuitively, the correction is needed to account for the loss in variability due to the fixed weights. Let

$$\mathbf{K}_{\alpha, n} = \hat{\sigma}_n \left[\sum_{i=1}^n \rho_1''(r_{\alpha i}/\hat{\sigma}_n, \mathbf{x}_{\alpha i}) \mathbf{x}_{\alpha i} \mathbf{x}_{\alpha i}' \right]^{-1} \sum_{i=1}^n \omega_{\alpha i} \mathbf{x}_{\alpha i} \mathbf{x}_{\alpha i}', \quad (16)$$

Note that $\mathbf{K}_{\alpha, n}$ is only computed once with the full sample and not bootstrapped. The Fast and Robust Bootstrap estimates $\hat{\boldsymbol{\beta}}_{\alpha, m}^{R*}$ are now given by

$$\hat{\boldsymbol{\beta}}_{\alpha, m}^{R*} = \hat{\boldsymbol{\beta}}_{\alpha, n} + \mathbf{K}_{\alpha, n} (\hat{\boldsymbol{\beta}}_{\alpha, m}^* - \hat{\boldsymbol{\beta}}_{\alpha, n}). \quad (17)$$

Salibian-Barrera and Zamar (2002) showed that if $\hat{\boldsymbol{\beta}}_{\alpha, n} \rightarrow \boldsymbol{\beta}_\alpha$ then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\alpha, n}^{R*} - \hat{\boldsymbol{\beta}}_{\alpha, n})$ has the same asymptotic distribution as $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\alpha, n} - \boldsymbol{\beta}_\alpha)$, and that the breakdown point of the quantile estimates is higher than those obtained with the classical bootstrap.

4 Consistency

In this section we study the asymptotic behaviour of the robust selection procedures (11) and (12). For this purpose, we consider the usual linear model $y_i = \mathbf{x}_i' \boldsymbol{\beta} + \sigma \epsilon_i$, $i = 1, \dots, n$, with $\boldsymbol{\beta} \in \mathbb{R}^p$, $\sigma > 0$ and independent errors ϵ_i with zero location and scale equal to one. Let $\mathcal{A}_c \subset \mathcal{A}$ be the set of models α such that their associated vector of regression coefficients $\boldsymbol{\beta}_\alpha$ in (1) contain all non-zero components of $\boldsymbol{\beta}$. In what follows we will assume that \mathcal{A}_c is not empty.

The smallest model in \mathcal{A}_c will be called the “true” model α_0 (i.e. $p_{\alpha_0} < p_\alpha$ for all $\alpha \in \mathcal{A}_c$, $\alpha \neq \alpha_0$). Note that if $\alpha \in \mathcal{A}_c$ then $\mathbf{x}_{\alpha i}'\boldsymbol{\beta}_\alpha = \mathbf{x}'_{\alpha_0 i}\boldsymbol{\beta}_{\alpha_0} = \mathbf{x}'_i\boldsymbol{\beta}$ and the errors $\epsilon_{\alpha i}$ in (1) satisfy $\epsilon_{\alpha i} = \epsilon_{\alpha_0 i} = \epsilon_i$.

Following Shao (1996) we will say that a model selection criterion is consistent if the probability of selecting the true model α_0 converges to 1 as the sample size $n \rightarrow \infty$. In other words if $\lim_{n \rightarrow \infty} P(\hat{\alpha}_n = \alpha_0) = 1$. Theorem 1 below proves the consistency of the robust selection procedures (11) and (12) under the condition that $\alpha_0 \in \mathcal{A}$ exists.

Theorem 1 *Consider MM-regression estimators as defined by (2) and associated S-scale estimators as in (4). Assume that*

(A1) *for all models α we have $n^{-1} \sum \mathbf{x}_{\alpha i}\mathbf{x}'_{\alpha i} \rightarrow \Gamma_\alpha$ and $n^{-1} \sum \omega_{\alpha i}\mathbf{x}_{\alpha i}\mathbf{x}'_{\alpha i} \rightarrow \Gamma_{\omega_\alpha}$ where Γ_α and Γ_{ω_α} are of full rank, and $n^{-1} \sum \|\mathbf{x}_{\alpha i}\|^4 < \infty$, where the weights $\omega_{\alpha i}$ are given in (13);*

(A2) *$\delta(n) = o(n/m)$ and $m = o(n)$;*

(A3) *for all models α , $\hat{\boldsymbol{\beta}}_{\alpha, n}$ satisfies $\sum_{i=1}^n \rho'_1(r_i(\hat{\boldsymbol{\beta}}_{\alpha, n})/\hat{\sigma}_n)\mathbf{x}_{\alpha i} = \mathbf{0}$, where $r_i(\hat{\boldsymbol{\beta}}_{\alpha, n}) = y_i - \hat{\boldsymbol{\beta}}'_{\alpha, n}\mathbf{x}_{\alpha i}$, $i = 1, \dots, n$;*

(A4) *$\hat{\sigma}_n - \sigma = O_p(1/\sqrt{n})$, and for all models α , $\hat{\boldsymbol{\beta}}_{\alpha, n} - \boldsymbol{\beta}_\alpha = O_p(1/\sqrt{n})$;*

(A5) *ρ'_1 and ρ''_1 are uniformly continuous, $\text{var}(\rho'_1(\epsilon_{\alpha_0}))$, and $\text{var}(\rho''_1(\epsilon_{\alpha_0}))$ are finite and $E(\rho''_1(\epsilon_{\alpha_0})) > 0$; and*

(A6) *for any $\alpha \notin \mathcal{A}_c$, $\text{var}(\rho'_1(\epsilon_\alpha)) < \infty$ and with probability one*

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \rho_1(r_i(\hat{\boldsymbol{\beta}}_\alpha)/\hat{\sigma}_n) > \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \rho_1(r_i(\hat{\boldsymbol{\beta}}_{\alpha_0, n})/\hat{\sigma}_n).$$

Then the selection procedures (11) and (12) are consistent when E_ in (9) denotes the expected value of the fast and robust bootstrap. In other words, we*

have

$$\lim_{n \rightarrow \infty} P(\hat{\alpha}_{m,n}^{ppe} = \alpha_0) = \lim_{n \rightarrow \infty} P(\hat{\alpha}_{m,n}^{pe} = \alpha_0) = 1.$$

Remark 1 *Note that, as in Shao (1996), to obtain a consistent selection criterion we need to use bootstrap samples of smaller size than the original data set, specifically: $m = o(n)$. Shao (1996) indicates that these small bootstrap samples may result in an increase of the variability of the bootstrap estimator of the conditional expected loss. The use of computationally complex robust estimators creates additional practical problems. In particular, the condition $m = o(\sqrt{n})$ of Müller and Welsh (2005) together with the full re-calculation of the robust estimator for each bootstrap sample may present serious computational difficulties, particularly when n/p is not too large. More specifically, we found that if n/p is moderately small, it may be hard to calculate the initial S -estimator based on a bootstrap sample of size m with $m \ll n$, particularly when some observations are repeated. The use of the fast and robust bootstrap avoids this problem because it does not require the re-calculation of the robust estimator on the bootstrap samples.*

5 Simulations

We ran two simulation studies to investigate the finite-sample performance of the robust selection procedures based on FRB.

First, to compare with published results of other proposals in the literature, we considered the solid waste data of Gunst and Mason (1980), which were already used by Shao (1993, 1996, 1997), Wisnowski *et al.* (2003) and Müller

and Welsh (2005). The model is

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \epsilon_i, \quad i = 1, \dots, 40. \quad (18)$$

Following Shao (1993), we generated 1000 samples from this model for different values of $\boldsymbol{\beta} = (\beta_1, \dots, \beta_5)'$, with $\epsilon_i \sim \mathcal{N}(0, 1)$. We then applied the different model selection procedures to choose a model from a pre-specified list. The models were fit using MM-estimators with loss functions from the Tukey bi-weight family (5) tuned to have maximal breakdown point ($\approx 50\%$) and 95% efficiency for Gaussian errors. For the model selection criteria both the prediction error criterion $\hat{\alpha}_{m,n}^{\text{pe}}$ and its penalized version $\hat{\alpha}_{m,n}^{\text{ppe}}$ with $\delta(n) = \log(n)$ were used. For the scale estimate $\hat{\sigma}_n$ we used the S-scale of the largest model considered.

The loss function ρ in (7)-(8) was taken equal to the loss function ρ_1 of the estimator in (2). In principle, by using a different ρ function in the model selection criterion one can construct a measure of prediction error that is unrelated to any particular estimation method, and thus could be used to compare models that have been fit with different estimators. However, this may produce contradictions between the observations down-weighted by the estimation procedure and those down-weighted in the prediction error estimate, which can be hard to interpret. Furthermore, as shown in Müller and Welsh (2005), using a different loss function requires much smaller bootstrap samples, which may affect the precision of the estimated expected prediction loss.

We used 100 bootstrap samples of size $m = 20$ and considered model selection for the MM-estimator based on the fast robust bootstrap [FRB], the stratified bootstrap [SB] and the standard full bootstrap [FB]. We also used the model selection procedures using squared error loss and the LS estimator with the

usual bootstrap [LS]. Finally, to compare these methods with computationally less demanding approaches, we included Akaike's AIC, Mallows's C_p , and the Robust future prediction error criterion (RFPE) (Maronna *et al.* 2006).

Tables 1 and 2 contain the results for penalty criteria $\hat{\alpha}_{m,n}^{\text{pe}}$ and $\hat{\alpha}_{m,n}^{\text{ppe}}$, respectively. For the penalty term in (10) we used $\delta(n) = \log(n)$. In each table there are three cases: one where the true model is the full model containing all predictors, another where there is one zero predictor, and in the last case the true model is the smallest model considered. From Table 1 we see that even in this contamination free case, model selection based on least squares does not always outperform the other methods, although it is better than both AIC and C_p . Not surprisingly, the RFPE criterion exhibits the worst performance in this case. Moreover, for the MM-estimator, the FRB, SB and FB behaved similarly. Only in the last case, where the true model is the smallest model considered, the FRB seems to work better than SB or FB. Comparing the results for the prediction error criterion (11) in Table 1 with the results for its penalized alternative (12) in Table 2, we see that for all procedures penalization is helpful in all cases, even when the true model is the full model.

To illustrate the performance of the selection procedures when there are outliers in the data, we follow the simulation scheme of Müller and Welsh (2005).

We generated 1000 samples of size $n = 64$ from the following model:

$$y = 2 + 2x_1 + 0x_2 + \epsilon. \quad (19)$$

The covariates were generated from a $U(0, 1)$ distribution and kept fixed throughout the study. The errors ϵ were generated from the following six distributions:

Table 1

Proportion of times each model was chosen by the selection criterion $\hat{\alpha}_{m,n}^{\text{pe}}$. Results are based on 1000 independent samples following model (18). FRB = Fast and robust Bootstrap, FB = Full (classical) Bootstrap, SB = Stratified Bootstrap, LS = Least Squares and full bootstrap.

True beta	Method	Model				
		1,4,5	1,2,3,5	1,2,4,5	1,3,4,5	1,2,3,4,5
<hr/>						
(2,9,6,4,8)						
	FRB	0.000	0.021	0.005	0.003	0.971
	FB	0.000	0.003	0.000	0.000	0.997
	SB	0.000	0.000	0.000	0.000	1.000
	LS	0.000	0.000	0.000	0.000	1.000
	AIC	0.000	0.000	0.000	0.000	1.000
	C_p	0.000	0.000	0.000	0.000	1.000
	RFPE	0.079	0.005	0.024	0.026	0.866
<hr/>						
(2,9,0,4,8)						
	FRB	0.000	0.001	0.870	0.000	0.129
	FB	0.000	0.000	0.832	0.000	0.168
	SB	0.000	0.000	0.784	0.000	0.216
	LS	0.000	0.000	0.923	0.000	0.077
	AIC	0.000	0.000	0.811	0.000	0.189
	C_p	0.000	0.000	0.847	0.000	0.153
	RFPE	0.097	0.005	0.604	0.027	0.267
<hr/>						
(2,0,0,4,8)						
	FRB	0.838	0.001	0.063	0.074	0.024
	FB	0.680	0.001	0.127	0.150	0.042
	SB	0.635	0.001	0.141	0.163	0.060
	LS	0.853	0.000	0.065	0.076	0.006
	AIC	0.668	0.000	0.136	0.134	0.062
	C_p	0.695	0.000	0.134	0.122	0.049
	RFPE	0.541	0.005	0.165	0.178	0.111

Table 2

Proportion of times each model was chosen by the selection criterion $\hat{\alpha}_{m,n}^{\text{ppe}}$. Results are based on 1000 independent samples following model (18) with penalty term $\delta(n) = \log(n)$. FRB = Fast and robust Bootstrap, FB = Full (classical) Bootstrap, SB = Stratified Bootstrap, LS = Least Squares and full bootstrap.

True beta	Method	Model				
		1,4,5	1,2,3,5	1,2,4,5	1,3,4,5	1,2,3,4,5
<hr/>						
(2,9,6,4,8)						
	FRB	0.000	0.010	0.000	0.001	0.989
	FB	0.000	0.011	0.000	0.000	0.989
	SB	0.000	0.000	0.000	0.000	1.000
	LS	0.000	0.000	0.000	0.000	1.000
<hr/>						
(2,9,0,4,8)						
	FRB	0.000	0.000	0.921	0.000	0.079
	FB	0.000	0.000	0.927	0.000	0.073
	SB	0.000	0.000	0.922	0.000	0.078
	LS	0.000	0.000	0.929	0.000	0.071
<hr/>						
(2,0,0,4,8)						
	FRB	0.907	0.009	0.036	0.047	0.010
	FB	0.865	0.001	0.050	0.070	0.014
	SB	0.859	0.001	0.056	0.069	0.015
	LS	0.880	0.000	0.052	0.063	0.005
<hr/>						

- (i) $\epsilon \sim 5/8 N(0, 1) + 3/8 N(30 - 2 - 2x_1, 1)$,
- (ii) $\epsilon \sim 3/4 N(0, 1) + 1/4 N(30 - 2 - 2x_1, 1)$,
- (iii) $\epsilon \sim 7/8 N(0, 1) + 1/8 N(30 - 2 - 2x_1, 1)$,
- (iv) $\epsilon \sim \mathcal{N}(0, 1)$,
- (v) $\epsilon \sim V/W$ where $V \sim N(0, 1)$ and $W \sim U(0, 1)$, V and W independent,
and
- (vi) $\epsilon \sim \text{Cauchy}$.

We used 100 bootstrap samples of size $m = 24$ and considered the same selection criteria as in the previous simulation. The results are shown in Tables 3 and 4. We considered all possible models: intercept only (1), intercept and x_1 (1,2), intercept and x_2 (1,3) and the full model (1,2,3). We see that RFPE tends to select larger models than the bootstrap based selection procedures. The selection procedures based on least squares and the AIC and C_p (not shown) do not perform well except in the uncontaminated normal errors case. For the MM-estimators combined with any of the three bootstrap procedures, the selection criterion solely based on prediction error shows robust behavior at the contaminated normal error distributions. The FRB performs marginally better than the SB and FB. In this simulation setting, the penalized selection criterion only improves performance in the uncontaminated case. For all other error distribution it selects too often the intercept only model. None of the methods performs very well in the difficult settings with the Cauchy and Slash error distributions with FRB being somewhat worse than FB and SB. Note that to keep the simulation study feasible, we considered only a few low dimensional models. In higher dimensional model selection problems, the penalty term in the selection criterion can become beneficial also in contaminated cases. This will be further illustrated with the examples in the next

section. In general, model selection based on FRB behaves quite similar as model selection based on SB. However, FRB has the additional advantage of being much faster and thus feasible for larger scale problems as will be illustrated in the next section.

6 Examples

We now consider robust model selection for three real data examples available in R. For each of the examples we fitted models using regression MM-estimators with Tukey biweight loss function tuned to have maximal breakdown point and 95% efficiency for Gaussian errors.

We applied the commonly used backward elimination strategy as explained in the Introduction which implies that we need to fit $(p + 1)p/2$ models. As is common practice in model selection, the predictors were standardized. We used the median and the mad as center and scale estimators, respectively. As selection criteria we used the prediction error criterion $\hat{\alpha}_{m,n}^{\text{pe}}$ and its penalized version $\hat{\alpha}_{m,n}^{\text{ppe}}$, both based on FRB with $B = 1000$ bootstrap samples. The penalty term in (10) was $\delta(n) = k \log(n)$ with $k = 1$ or $k = 2$. As before, the scale estimate $\hat{\sigma}_n$ in these criteria was the S-scale of the full model and the loss function ρ was equal to ρ_1 in (2). For comparison we also performed the model selection procedure based on the RFPE criterion.

As explained before, the goal of the selection procedure is to find a parsimonious model that fits the data reasonably well and at the same time is able to predict future observations accurately. To investigate the quality of the fits selected by the different selection procedures, we calculate an adjusted robust

Table 3

Proportion of times each model is selected by the different selection criteria for samples without outliers [N] and samples with 12.5% ([1/8]), and 25% ([1/4]) of outliers. Results are based on 1000 samples following model (19) with penalty term $\delta(n) = \log(n)$. FRB = Fast and robust Bootstrap, FB = Full (classical) Bootstrap, SB = Stratified Bootstrap, LS = Least Squares and full bootstrap, RFPE = Robust future prediction error.

Errors	Model	Method								
		FB		SB		FRB		LS	RFPE	
		$\hat{\alpha}_{m,n}^{\text{pe}}$	$\hat{\alpha}_{m,n}^{\text{ppe}}$	$\hat{\alpha}_{m,n}^{\text{pe}}$	$\hat{\alpha}_{m,n}^{\text{ppe}}$	$\hat{\alpha}_{m,n}^{\text{pe}}$	$\hat{\alpha}_{m,n}^{\text{ppe}}$	$\hat{\alpha}_{m,n}^{\text{pe}}$	$\hat{\alpha}_{m,n}^{\text{ppe}}$	
[N]	1	0.007	0.027	0.008	0.028	0.036	0.054	0.004	0.007	0.002
	1,2	0.887	0.950	0.877	0.948	0.905	0.930	0.911	0.935	0.801
	1,3	0.000	0.000	0.000	0.000	0.004	0.003	0.000	0.000	0.001
	1,2,3	0.106	0.020	0.115	0.021	0.050	0.013	0.085	0.058	0.196
[1/8]	1	0.007	0.084	0.011	0.088	0.026	0.104	0.971	1.000	0.007
	1,2	0.910	0.909	0.885	0.904	0.932	0.889	0.029	0.000	0.801
	1,3	0.001	0.001	0.001	0.001	0.001	0.001	0.000	0.000	0.001
	1,2,3	0.082	0.006	0.103	0.007	0.041	0.006	0.000	0.000	0.191
[1/4]	1	0.007	0.230	0.008	0.234	0.015	0.245	1.000	1.000	0.005
	1,2	0.913	0.767	0.900	0.763	0.933	0.752	0.000	0.000	0.782
	1,3	0.001	0.000	0.001	0.000	0.002	0.000	0.000	0.000	0.002
	1,2,3	0.079	0.003	0.091	0.003	0.050	0.003	0.000	0.000	0.211

Table 4

Proportion of times each model is selected by the different selection criteria for samples with errors generated from contaminated normal with 37.5% ([3/8]) of outliers, Slash [S] and Cauchy [C] distributions. Results are based on 1000 samples following model (19) with penalty term $\delta(n) = \log(n)$. FRB = Fast and robust Bootstrap, FB = Full (classical) Bootstrap, SB = Stratified Bootstrap, LS = Least Squares and full bootstrap, RFPE = Robust future prediction error.

Errors	Model	Method								
		FB		SB		FRB		LS	RFPE	
		$\hat{\alpha}_{m,n}^{\text{pe}}$	$\hat{\alpha}_{m,n}^{\text{ppe}}$	$\hat{\alpha}_{m,n}^{\text{pe}}$	$\hat{\alpha}_{m,n}^{\text{ppe}}$	$\hat{\alpha}_{m,n}^{\text{pe}}$	$\hat{\alpha}_{m,n}^{\text{ppe}}$	$\hat{\alpha}_{m,n}^{\text{pe}}$	$\hat{\alpha}_{m,n}^{\text{ppe}}$	
[3/8]	1	0.032	0.815	0.032	0.825	0.045	0.817	1.000	1.000	0.013
	1,2	0.896	0.184	0.886	0.174	0.894	0.182	0.000	0.000	0.751
	1,3	0.007	0.000	0.007	0.000	0.007	0.000	0.000	0.000	0.002
	1,2,3	0.065	0.001	0.075	0.001	0.054	0.001	0.000	0.000	0.234
[S]	1	0.451	0.751	0.382	0.742	0.582	0.790	0.820	0.893	0.258
	1,2	0.476	0.236	0.529	0.242	0.373	0.199	0.114	0.071	0.549
	1,3	0.043	0.012	0.046	0.013	0.034	0.009	0.054	0.032	0.055
	1,2,3	0.030	0.001	0.043	0.003	0.011	0.002	0.012	0.004	0.138
[C]	1	0.219	0.463	0.169	0.457	0.373	0.552	0.780	0.856	0.078
	1,2	0.709	0.523	0.736	0.526	0.580	0.436	0.155	0.109	0.713
	1,3	0.019	0.008	0.025	0.009	0.020	0.008	0.057	0.029	0.030
	1,2,3	0.053	0.006	0.070	0.008	0.027	0.004	0.008	0.006	0.179

R-squared (see Maronna *et al.* 2006), which for a model of size p_α is given by

$$RR_{\text{adj}}^2 = \frac{\hat{\sigma}_{1,n}^2/(n-1) - \hat{\sigma}_{\alpha,n}^2/(n-p_\alpha)}{\hat{\sigma}_{1,n}^2/(n-1)},$$

where $\hat{\sigma}_{1,n}$ is the robust residual scale of the “null” model that only includes the intercept as predictor. We also compare plots of standardized residuals versus fitted values.

To compare the predictive power of the models we ran 5-fold cross-validation on each of the models. More specifically, consider a model α and let \mathcal{A}_j , $j = 1, \dots, 5$ be disjoint subsets such that $\bigcup_{j=1}^5 \mathcal{A}_j = \{1, \dots, n\}$. For each j , let $\hat{\boldsymbol{\beta}}_{\alpha,n}^{(-j)}$ be the estimated vector of regression parameters without using the observations with indices in the set \mathcal{A}_j . We report the following two prediction error criteria:

$$\text{TMSE} = \left[\sum_{j=1}^5 \text{ave}_{i \in \mathcal{A}_j}^{(\gamma)} \left\{ (y_i - \mathbf{x}'_{\alpha i} \hat{\boldsymbol{\beta}}_{\alpha,n}^{(j)})^2 \right\} \right] / 5, \quad (20)$$

and

$$\bar{\rho} = \hat{\sigma}_n^2 \left[\sum_{j=1}^5 \text{ave}_{i \in \mathcal{A}_j} \left\{ \rho \left((y_i - \mathbf{x}'_{\alpha i} \hat{\boldsymbol{\beta}}_{\alpha,n}^{(j)}) / \hat{\sigma}_n \right) \right\} \right] / 5,$$

where $\hat{\sigma}_n$ is the error scale estimate, and $\text{ave}_{i \in A}^{(\gamma)} \{t_i\}$ and $\text{ave}_{i \in A} \{t_i\}$ denote the γ 100% upper trimmed mean and the usual sample mean of the t_i 's with $i \in A$, respectively. Note that TMSE is a trimmed mean squared error where the trimming reflects that we allow a fraction of outliers to be predicted not well. We used $\gamma = 0.05$ and 0.10 .

6.1 Example – Ozone

The Los Angeles Ozone Pollution Data contains 366 daily observations on 9 variables (see Breiman and Friedman, 1985). The response variable is Daily maximum one-hour-average ozone reading. The measured explanatory variables are temperature (degrees F) measured at Sandburg, CA, inversion base height (feet) at LAX, pressure gradient (mm Hg) from LAX to Daggett, CA, visibility (miles) measured at LAX, 500 millibar pressure height (m) measured at Vandenberg AFB, humidity (%) at LAX, inversion base temperature (degrees F) at LAX, and wind speed (mph) at Los Angeles International Airport (LAX), respectively. The “full” model includes all second order interactions and quadratic terms which yields a model with $p = 45$ predictors.

We applied the backward selection procedure based on $\hat{\alpha}_{m,n}^{\text{pe}}$ and $\hat{\alpha}_{m,n}^{\text{ppe}}$ with FRB using 1000 bootstrap samples of size $m = 100$ to this dataset. This took 4 hours and ten minutes on an Intel Xeon CPU with 2GB of RAM running at 3GHz. To illustrate that it would not be feasible to use the stratified bootstrap instead of the fast and robust bootstrap in the backward selection procedure, Table 5 compares the average time needed to run 1000 bootstrap samples of size $m = 200$ for models of different sizes p . These times are CPU seconds on the same Intel Xeon CPU machine mentioned above. The computing times in Table 5 suggest that it takes at least 100 times longer to run the selection procedure using the stratified bootstrap. With the stratified bootstrap it would thus take more than 15 days to obtain the result for the Ozone data instead of just over 4 hours for the FRB.

Backward selection based on $\hat{\alpha}_{m,n}^{\text{pe}}$ selected a model with 10 predictor variables.

Table 5

Average time (CPU seconds) needed to run 1000 bootstrap samples of size $m = 200$ for models of different number of covariates (p) using the fast and robust bootstrap (FRB) and the stratified bootstrap (SB).

p	25	35	45
FRB	8	28	35
SB	1955	4300	10700

The penalized criterion $\hat{\alpha}_{m,n}^{\text{ppe}}$ with $k = 1$ selected a model with 7 predictors. Using $k = 2$ resulted in a model with 6 predictors which was a submodel of the one obtained with $k = 1$. Backward selection using RFPE selected a model with $p = 23$ predictor variables.

Although there are large differences in the number of predictors of the selected models, residual plots (not shown) did not reveal any appreciable differences between these models. The RR_{adj}^2 coefficients for the models selected with $\hat{\alpha}_{m,n}^{\text{pe}}$, $\hat{\alpha}_{m,n}^{\text{ppe}}$ ($k = 2$), RFPE and the full model were 0.7583, 0.7643, 0.8174 and 0.8660, respectively. While the difference in size between the RFPE model and the small models selected by $\hat{\alpha}_{m,n}^{\text{pe}}$ and $\hat{\alpha}_{m,n}^{\text{ppe}}$ ($k = 2$) is 13 and 16 respectively, the difference in adjusted robust R-squared is only 7% and 6% respectively. Compared with the full model, there is a very large size difference, but only a very small difference in adjusted R-squared. This confirms that the small models selected by the $\hat{\alpha}_{m,n}^{\text{pe}}$ and $\hat{\alpha}_{m,n}^{\text{ppe}}$ ($k = 2$) criteria produce good fits with an important reduction in the number of covariates. Moreover, Table 6 shows that the small models selected by $\hat{\alpha}_{m,n}^{\text{pe}}$ and $\hat{\alpha}_{m,n}^{\text{ppe}}$ ($k = 1$) also perform very well in terms of prediction error.

Table 6

5-fold CV prediction error estimators for the Ozone data. The parameter γ is the trimming fraction in the trimmed mean squared error (TMSE).

		$\hat{\alpha}_{m,n}^{\text{pe}}$		$\hat{\alpha}_{m,n}^{\text{ppe}}$		RFPE		Full model	
		$p = 10$		$p = 7$		$p = 23$		$p = 45$	
γ	TMSE	$\bar{\rho}$	TMSE	$\bar{\rho}$	TMSE	$\bar{\rho}$	TMSE	$\bar{\rho}$	
0.05	11.67	5.36	10.45	5.03	10.66	4.98	10.78	5.03	
0.10	9.18		8.35		8.18		8.33		

6.2 Example – Boston

The well-known Boston housing data contains 506 observations on 14 variables (see e.g. Belsley, Kuh, and Welsch, 1980). The response variable is the median value of occupied homes (in \$1000's). There are 13 measured predictors which leads to a full model of size $p = 14$.

Selection based on $\hat{\alpha}_{m,n}^{\text{pe}}$ and $\hat{\alpha}_{m,n}^{\text{ppe}}$ with $k = 1$ or $k = 2$ in the penalty term, using FRB and bootstrap samples of size $m = 150$ yielded models with 4 predictor variables while the selection based on RFPE produced a model of size 13. For $\hat{\alpha}_{m,n}^{\text{ppe}}$ the models found with $k = 1$ and $k = 2$ were identical, and it is different from the optimal model based on $\hat{\alpha}_{m,n}^{\text{pe}}$ in only one variable.

Figure 1 shows plots of standardized residuals versus fitted values for the optimal models based on $\hat{\alpha}_{m,n}^{\text{pe}}$, $\hat{\alpha}_{m,n}^{\text{ppe}}$ ($k = 2$), RFPE, and for the full model. We see that all models fit the data similarly, except for the model selected by $\hat{\alpha}_{m,n}^{\text{pe}}$. The RR_{adj}^2 coefficients for the models selected with $\hat{\alpha}_{m,n}^{\text{pe}}$, $\hat{\alpha}_{m,n}^{\text{ppe}}$ ($k = 2$), RFPE and the full model were 0.7244, 0.6815, 0.8264 and 0.8268, respectively.

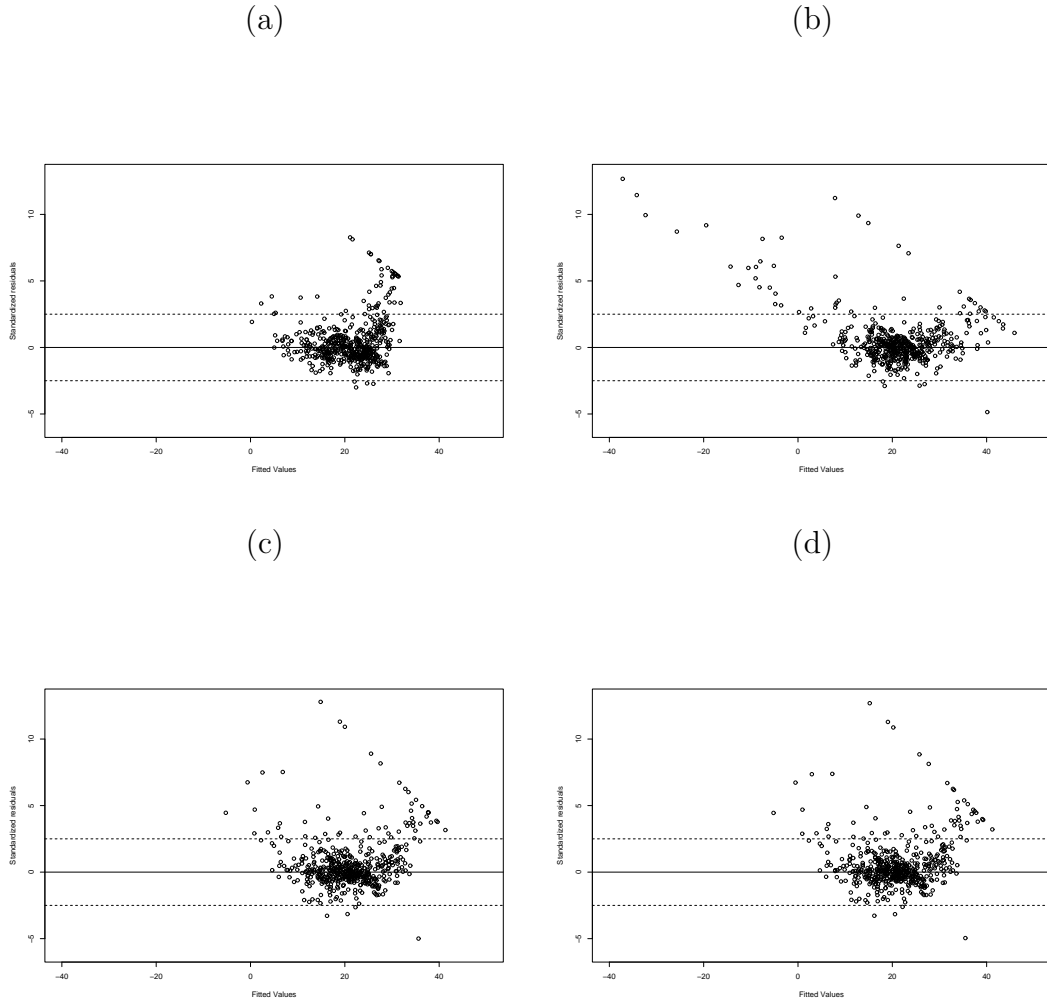


Fig. 1. Boston housing data: Plots of standardized residuals versus fitted values for the models selected by (a) $\hat{\alpha}_{m,n}^{\text{pe}}$; (b) $\hat{\alpha}_{m,n}^{\text{ppe}}$ ($k = 2$); (c) RFPE, and for (d) the full model.

Table 7 shows that the models selected by $\hat{\alpha}_{m,n}^{\text{pe}}$ and $\hat{\alpha}_{m,n}^{\text{ppe}}$ have a larger 5%-trimmed mean squared prediction error (TMSE) than the full and RFPE optimal models. However, both the more robust 10%-TMSE and the bounded-loss mean prediction error $\bar{\rho}$ show that the considerably smaller model selected by $\hat{\alpha}_{m,n}^{\text{pe}}$ performs very similarly to the full model. This example illustrates the distortion that outliers can produce on not sufficiently robust prediction error

Table 7

5-fold CV prediction error estimators for the Boston data. The parameter γ is the trimming fraction in the trimmed mean squared error (TMSE).

		$\hat{\alpha}_{m,n}^{\text{pe}}$		$\hat{\alpha}_{m,n}^{\text{ppe}}$		RFPE		Full model	
		$p = 4$		$p = 4$		$p = 13$		$p = 14$	
γ	TMSE	$\bar{\rho}$	TMSE	$\bar{\rho}$	TMSE	$\bar{\rho}$	TMSE	$\bar{\rho}$	
0.05	18.98	6.71	22.42	7.22	15.63	5.81	16.25	5.99	
0.10	10.63		13.21		9.61		10.16		

measures.

6.3 Example – Diabetes

These data contain $n = 442$ observations of ten baseline variables, age, sex, body mass index, average blood pressure, and six blood serum measurements. The response is a measure of disease progression one year after baseline. A quadratic model ($p = 65$) including some interactions of possible interest is fitted (see Efron *et al.*, 2004).

In this case, selection based on $\hat{\alpha}_{m,n}^{\text{pe}}$ with FRB using bootstrap samples of size $m = 110$ yields a model with 11 predictor variables while the penalized criteria $\hat{\alpha}_{m,n}^{\text{ppe}}$ with either $k = 1$ or $k = 2$ yield the same model with 7 predictors. As before, selection based on RFPE returns the largest model, with 16 predictors.

Residual plots (not shown) indicated that the selected models fit the data similarly and produce better fits than the full model. The RR_{adj}^2 coefficients for the models selected with $\hat{\alpha}_{m,n}^{\text{pe}}$, $\hat{\alpha}_{m,n}^{\text{ppe}}$, RFPE and the full model were 0.5127,

Table 8

5-fold CV prediction error estimators for the Diabetes data. The parameter γ is the trimming fraction in the trimmed mean squared error (TMSE).

		$\hat{\alpha}_{m,n}^{\text{pe}}$		$\hat{\alpha}_{m,n}^{\text{ppe}}$		RFPE		Full model	
		$p = 11$		$p = 7$		$p = 16$		$p = 65$	
γ	TMSE	$\bar{\rho}$	TMSE	$\bar{\rho}$	TMSE	$\bar{\rho}$	TMSE	$\bar{\rho}$	
0.05	2657.2	1220.6	2497.0	1178.3	2231.2	1079.8	4988.1	1741.8	
0.10	2199.1		2135.9		1898.3		4109.4		

0.5302, 0.6045 and 0.7731, respectively. The difference in adjusted robust R-squared between the selected models and the full model is small compared to the large difference in size. The models selected by $\hat{\alpha}_{m,n}^{\text{pe}}$ and $\hat{\alpha}_{m,n}^{\text{ppe}}$ have a considerably smaller adjusted R-squared than the model selected by RFPE, but also the size difference is considerable. Hence, in this example the $\hat{\alpha}_{m,n}^{\text{pe}}$ and $\hat{\alpha}_{m,n}^{\text{ppe}}$ criteria yield parsimonious models that reveal the effect of a few important covariates while the RFPE criterion yields a larger model that fits the data better but is more difficult to interpret due to its larger size.

Table 8 shows that the three selected models have a much better prediction accuracy than the full model. Moreover, there is only a small difference in accuracy between the parsimonious models selected by $\hat{\alpha}_{m,n}^{\text{pe}}$ and $\hat{\alpha}_{m,n}^{\text{ppe}}$ compared to the larger model selected by RFPE.

7 Conclusions

In this paper we investigated robust model selection for linear regression models. In particular, we considered selection criteria that estimate the prediction error of each model by using the fast and robust bootstrap (FRB). By using the FRB, the selection criteria can be calculated much faster than with alternative bootstrap approaches, which makes robust model selection feasible in higher dimensional problems than before. We proved the consistency of the selection criteria computed with the FRB if a true model exists and it is contained in the collection of candidate models under consideration. The simulations confirm the satisfactory behavior of the FRB selection procedures for finite samples and the analysis of three real data sets illustrates the capability of this approach to select parsimonious models that fit the data well. Although we have focused our presentation on MM-estimators, note that the FRB may in principle be applied to any estimator that can be written as the solution of a “smooth” fixed-point equation. These include other highly efficient and robust regression estimators such as the τ -estimators (Yohai and Zamar, 1988) and CM-estimators (Mendes and Tyler, 1996).

8 Appendix

Proof of theorem 1 First we will show that for all models $\alpha \in \mathcal{A}$ the bootstrap estimator $\hat{\beta}_{\alpha,m}^*$ satisfies

$$\hat{\beta}_{\alpha,m}^{R*} - \beta_{\alpha} = o_p(1), \quad (21)$$

and

$$E_*[\hat{\boldsymbol{\beta}}_{\alpha,m}^{R*}] - \hat{\boldsymbol{\beta}}_{\alpha,n} = o_p(1). \quad (22)$$

Furthermore, for correct models $\alpha \in \mathcal{A}_c$ we need

$$m \operatorname{var}_*(\hat{\boldsymbol{\beta}}_{\alpha,m}^{R*}) = n a \operatorname{var}(\hat{\boldsymbol{\beta}}_{\alpha,n}) + o_p(1), \quad (23)$$

for some $a > 0$. Finally (to simplify the proofs) in what follows we will assume that for all correct models $\alpha \in \mathcal{A}_c$ it holds that $n \operatorname{var}(\hat{\boldsymbol{\beta}}_{\alpha,n}) = \tau \Gamma_\alpha^{-1} + o_p(1)$, where $\tau \in \mathbb{R}_+$. Note that this assumption implies symmetric error distributions, but we conjecture that the result is true provided that (23) holds.

From (15) we have

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\alpha,m}^* - \hat{\boldsymbol{\beta}}_{\alpha,n} &= \left[\frac{1}{m} \sum_{i=1}^m \omega_{\alpha i}^* \mathbf{x}_{\alpha i}^* \mathbf{x}_{\alpha i}^{*'} \right]^{-1} \left[\frac{1}{m} \sum_{i=1}^m \rho_1' \left(\frac{y_i^* - \mathbf{x}_{\alpha i}^{*'} \hat{\boldsymbol{\beta}}_{\alpha,n}}{\hat{\sigma}_n} \right) \mathbf{x}_{\alpha i}^{*'} \right], \\ &= \mathbf{A}_m^{*-1} \mathbf{v}_m^*. \end{aligned}$$

It is easy to see that $\mathbf{v}_m^* = o_p(1)$ and that $\mathbf{A}_m^{*-1} = O_p(1)$. Hence $\hat{\boldsymbol{\beta}}_{\alpha,m}^* - \boldsymbol{\beta}_\alpha = \hat{\boldsymbol{\beta}}_{\alpha,m}^* - \hat{\boldsymbol{\beta}}_{\alpha,n} + \hat{\boldsymbol{\beta}}_{\alpha,n} - \boldsymbol{\beta}_\alpha = o_p(1)$ and (21) holds.

To show (22) write

$$\mathbf{A}_n = \frac{1}{n} \sum_{i=1}^n \omega_{\alpha i} \mathbf{x}_{\alpha i} \mathbf{x}_{\alpha i}' \xrightarrow[n \rightarrow \infty]{P} \Gamma_\alpha^\omega > 0,$$

and

$$\mathbf{A}_m^{*-1} \mathbf{v}_m^* = \left[\mathbf{A}_m^{*-1} - (\Gamma_\alpha^\omega)^{-1} \right] \mathbf{v}_m^* + (\Gamma_\alpha^\omega)^{-1} \mathbf{v}_m^* \quad (24)$$

It is easy to see that $E_* \mathbf{v}_m^* = \sum_i^n \rho_1'(r_i(\hat{\boldsymbol{\beta}}_{\alpha,n})/\hat{\sigma}_n) \mathbf{x}_{\alpha i}/n = \mathbf{0}$. Also, note that if \mathbf{A} and $\boldsymbol{\Gamma}$ are non-singular and $\mathbf{A} = \boldsymbol{\Gamma} + (\mathbf{A} - \boldsymbol{\Gamma})$, we have

$$\mathbf{A}^{-1} = \boldsymbol{\Gamma}^{-1} - \boldsymbol{\Gamma}^{-1}(\mathbf{A} - \boldsymbol{\Gamma})(\mathbf{I} + \boldsymbol{\Gamma}^{-1}(\mathbf{A} - \boldsymbol{\Gamma}))^{-1} \boldsymbol{\Gamma}^{-1}, \quad (25)$$

(Seber, 1984, page 519), and thus if $\mathbf{A} - \mathbf{\Gamma} = o_p(1)$ we get $\mathbf{A}^{-1} = \mathbf{\Gamma}^{-1} + o_p(1)$. It follows that $\mathbf{A}_m^{*-1} - \mathbf{\Gamma}_\alpha^{\omega-1} = o_p(1)$ in (24). Furthermore, $E_*[\mathbf{A}_m^{*-1} - \mathbf{\Gamma}_\alpha^{\omega-1}] = \mathbf{A}_n^{-1} - \mathbf{\Gamma}_\alpha^{\omega-1} + o_p(1) = o_p(1)$ and thus $E_*(\hat{\boldsymbol{\beta}}_{\alpha,m}^* - \hat{\boldsymbol{\beta}}_{\alpha,n}) = o_p(1)$. Next, note that the correction matrix is bounded in probability, and thus will not affect the convergence rate. Hence (22) holds. Finally, for the fast and robust bootstrap, (23) holds with $a = 1$ (see Salibian-Barrera and Zamar, 2002).

We can now show that the selection criteria $M_{m,n}^{\text{pe}}$ and $M_{m,n}^{\text{ppe}}$ are consistent. Consider the bootstrap term in $M_{m,n}^{\text{ppe}}$ and for any $\boldsymbol{\beta}$ let $r_i(\boldsymbol{\beta}) = y_i - \mathbf{x}'_i \boldsymbol{\beta}$. We have

$$\begin{aligned} n^{-1} E_* \sum_{i=1}^n \rho(r_i(\hat{\boldsymbol{\beta}}_{\alpha,n}^{R*})/\hat{\sigma}_n) &= n^{-1} \sum_{i=1}^n \rho(r_i(\hat{\boldsymbol{\beta}}_{\alpha,n})/\hat{\sigma}_n) \\ &\quad + (\hat{\sigma}_n n)^{-1} E_* \sum_{i=1}^n \rho'(r_i(\hat{\boldsymbol{\beta}}_{\alpha,n})/\hat{\sigma}_n) \mathbf{x}_{\alpha i}' (\hat{\boldsymbol{\beta}}_{\alpha,n}^{R*} - \hat{\boldsymbol{\beta}}_{\alpha,n}) \\ &\quad + (2\hat{\sigma}_n^2 n)^{-1} E_* \sum_{i=1}^n \rho''(r_i(\hat{\boldsymbol{\beta}}_{\alpha,n}^*/\hat{\sigma}_n) \mathbf{x}_{\alpha i}' (\hat{\boldsymbol{\beta}}_{\alpha,n}^{R*} - \hat{\boldsymbol{\beta}}_{\alpha,n}) (\hat{\boldsymbol{\beta}}_{\alpha,n}^{R*} - \hat{\boldsymbol{\beta}}_{\alpha,n})' \mathbf{x}_{\alpha i}, \end{aligned} \quad (26)$$

here $\hat{\boldsymbol{\beta}}_{\alpha,n}^*$ is an intermediate point between $\hat{\boldsymbol{\beta}}_{\alpha,n}^{R*}$ and $\hat{\boldsymbol{\beta}}_{\alpha,n}$. First note that the second term on the right-hand side of (26) satisfies

$$n^{-1} E_* \sum_{i=1}^n \rho'(r_i(\hat{\boldsymbol{\beta}}_{\alpha,n})/\hat{\sigma}_n) \mathbf{x}_{\alpha i}' (\hat{\boldsymbol{\beta}}_{\alpha,n}^{R*} - \hat{\boldsymbol{\beta}}_{\alpha,n}) = \mathbf{0}. \quad (27)$$

Remark 2 Note that if we allow the estimator $\hat{\boldsymbol{\beta}}_{\alpha,n}$ to satisfy the estimating equations approximately, i.e.

$$\frac{1}{n} \sum_{i=1}^n \rho'(r_i(\hat{\boldsymbol{\beta}}_{\alpha,n})/\hat{\sigma}_n) \mathbf{x}_{\alpha i} = o_p(1/\sqrt{n}),$$

then this proof is still valid as long as $m = O(\sqrt{n})$, because in that case we have that the left-hand side of (27) is $o_p(1/m)$.

Now note that using (23) and calling $K_\rho = E[\rho''(r(\boldsymbol{\beta}_\alpha)/\sigma)]$ we have that the

third term on the right-hand side of (26) satisfies

$$\begin{aligned}
& \left| (2\hat{\sigma}_n^2 n)^{-1} E_* \sum_{i=1}^n \rho''(r_i(\hat{\boldsymbol{\beta}}_{\alpha,n}^{R*})/\hat{\sigma}_n) \mathbf{x}_{\alpha i}' (\hat{\boldsymbol{\beta}}_{\alpha,n}^{R*} - \hat{\boldsymbol{\beta}}_{\alpha,n}) (\hat{\boldsymbol{\beta}}_{\alpha,n}^{R*} - \hat{\boldsymbol{\beta}}_{\alpha,n})' \mathbf{x}_{\alpha i} \right. \\
& \quad \left. - \frac{1}{2m} p_\alpha a \tau K_\rho \right| \\
&= \left| (2\hat{\sigma}_n^2 n)^{-1} \sum_{i=1}^n E_* \mathbf{x}_{\alpha i}' (\hat{\boldsymbol{\beta}}_{\alpha,n}^{R*} - \hat{\boldsymbol{\beta}}_{\alpha,n}) (\hat{\boldsymbol{\beta}}_{\alpha,n}^{R*} - \hat{\boldsymbol{\beta}}_{\alpha,n})' \mathbf{x}_{\alpha i} \rho''(r_i(\hat{\boldsymbol{\beta}}_{\alpha,n}^*)/\hat{\sigma}_n) \right. \\
& \quad \left. - \frac{1}{2m} p_\alpha a \tau K_\rho \right| \\
&= \left| (2\hat{\sigma}_n^2 n)^{-1} \sum_{i=1}^n E_* \text{tr} \left[\mathbf{x}_{\alpha i}' (\hat{\boldsymbol{\beta}}_{\alpha,n}^{R*} - \hat{\boldsymbol{\beta}}_{\alpha,n}) (\hat{\boldsymbol{\beta}}_{\alpha,n}^{R*} - \hat{\boldsymbol{\beta}}_{\alpha,n})' \mathbf{x}_{\alpha i} \rho''(r_i(\hat{\boldsymbol{\beta}}_{\alpha,n}^*)/\hat{\sigma}_n) \right] \right. \\
& \quad \left. - \frac{1}{2m} p_\alpha a \tau K_\rho \right| \\
&= \left| (2\hat{\sigma}_n^2 n)^{-1} \sum_{i=1}^n \text{tr} \left[V_*(\hat{\boldsymbol{\beta}}_{\alpha,n}^{R*}) \mathbf{x}_{\alpha i} \mathbf{x}_{\alpha i}' \rho''(r_i(\hat{\boldsymbol{\beta}}_{\alpha,n}^*)/\hat{\sigma}_n) - \frac{1}{2m} p_\alpha a \tau K_\rho \right] \right|.
\end{aligned}$$

Since $V_*(\hat{\boldsymbol{\beta}}_{\alpha,n}^{R*}) = a \tau \Gamma^{-1}/m + o_p(1/m)$, and letting $\tilde{r}_i = r_i(\hat{\boldsymbol{\beta}}_{\alpha,n}^*)$ we have

$$\begin{aligned}
& \left| (2\hat{\sigma}_n^2 n)^{-1} E_* \sum_{i=1}^n \rho''(\tilde{r}_i/\hat{\sigma}_n) \mathbf{x}_{\alpha i}' (\hat{\boldsymbol{\beta}}_{\alpha,n}^{R*} - \hat{\boldsymbol{\beta}}_{\alpha,n}) (\hat{\boldsymbol{\beta}}_{\alpha,n}^{R*} - \hat{\boldsymbol{\beta}}_{\alpha,n})' \mathbf{x}_{\alpha i} - \frac{1}{2m} p_\alpha a K_\rho \right| \\
& \leq \left| \frac{a \tau}{m} \text{tr} \left[\Gamma^{-1} \frac{1}{n} \sum_{i=1}^n \mathbf{x}_{\alpha i} \mathbf{x}_{\alpha i}' \rho''(\tilde{r}_i/\hat{\sigma}_n) \right] - \frac{a \tau}{m} \text{tr} [\Gamma^{-1} \Gamma K_\rho] \right| + o_p(1/m) \\
& \leq \left| \frac{a \tau}{m} \text{tr} \left[\Gamma^{-1} \frac{1}{n} \sum_{i=1}^n \mathbf{x}_{\alpha i} \mathbf{x}_{\alpha i}' [\rho''(\tilde{r}_i/\hat{\sigma}_n) - \rho''(e_i)] \right] \right| \\
& + \left| \frac{a \tau}{m} \text{tr} \left[\Gamma^{-1} \left[\frac{1}{n} \sum_{i=1}^n \mathbf{x}_{\alpha i} \mathbf{x}_{\alpha i}' \rho''(e_i) - \Gamma K_\rho \right] \right] \right| + o_p(1/m) = o_p(1/m). \quad (28)
\end{aligned}$$

Therefore, for each correct model $\alpha \in \mathcal{A}_c$ we have

$$\begin{aligned}
M_{m,n}^{\text{PPE}}(\alpha) &= \frac{\hat{\sigma}_n^2}{n} \left[2 \sum_{i=1}^n \rho(r_i(\hat{\boldsymbol{\beta}}_{\alpha,n})/\hat{\sigma}_n) + \delta(n) p_\alpha + \frac{1}{2m} p_\alpha a \tau K_\rho \right] \\
&+ o_p(1/m) = \frac{2\hat{\sigma}_n^2}{n} \sum_{i=1}^n \rho(r_i(\hat{\boldsymbol{\beta}}_{\alpha,n})/\hat{\sigma}_n) + \hat{\sigma}_n^2 \delta(n) p_\alpha + \frac{\hat{\sigma}_n^2}{2m} p_\alpha a \tau K_\rho \\
&+ o_p(1/m).
\end{aligned}$$

Now note that

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^n \rho \left(r_i(\hat{\boldsymbol{\beta}}_{\alpha,n}) / \hat{\sigma}_n \right) \\
&= \frac{1}{n} \sum_{i=1}^n \rho \left(r_i(\boldsymbol{\beta}_\alpha) / \hat{\sigma}_n \right) - (\hat{\boldsymbol{\beta}}_{\alpha,n} - \boldsymbol{\beta}_\alpha)' \frac{1}{n} \sum_{i=1}^n \rho' \left(r_i(\boldsymbol{\beta}_\alpha) / \hat{\sigma}_n \right) \mathbf{x}_{\alpha i} / \hat{\sigma}_n \\
&\quad + \frac{1}{n \hat{\sigma}_n^2} \sum_{i=1}^n \rho'' \left(r_i(\boldsymbol{\beta}_\alpha) / \hat{\sigma}_n \right) \mathbf{x}_{\alpha i}' (\hat{\boldsymbol{\beta}}_{\alpha,n} - \boldsymbol{\beta}_\alpha) (\hat{\boldsymbol{\beta}}_{\alpha,n} - \boldsymbol{\beta}_\alpha)' \mathbf{x}_{\alpha i} \\
&= \frac{1}{n} \sum_{i=1}^n \rho \left(r_i(\boldsymbol{\beta}_\alpha) / \hat{\sigma}_n \right) + O_p(1/n). \quad (29)
\end{aligned}$$

It follows that for correct models $\alpha \in \mathcal{A}_c$ we have

$$\begin{aligned}
M_{m,n}^{\text{ppe}}(\alpha) &= \frac{2 \hat{\sigma}_n^2}{n} \sum_{i=1}^n \rho \left(r_i(\boldsymbol{\beta}_\alpha) / \hat{\sigma}_n \right) + \hat{\sigma}_n^2 \delta(n) p_\alpha + \frac{\hat{\sigma}_n^2}{2m} p_\alpha a \tau K_\rho \\
&\quad + o_p(1/m) + O_p(1/n),
\end{aligned}$$

Since $m O_p(1/n) = (m/n) n O_p(1/n) = o(1) O_p(1) = o_p(1)$ it follows that

$M_{m,n}^{\text{ppe}}(\alpha) - M_{m,n}^{\text{ppe}}(\alpha_0) > 0$ if and only if

$$\begin{aligned}
& (p_\alpha - p_{\alpha_0}) \left(\frac{\delta(n)}{n} \hat{\sigma}_n^2 + \frac{a \tau}{m} K_\rho \right) + o_p(1/m) > 0, \\
\Leftrightarrow & (p_\alpha - p_{\alpha_0}) \frac{\delta(n)}{n} \hat{\sigma}_n^2 + (p_\alpha - p_{\alpha_0}) \frac{a \tau}{m} K_\rho + o_p(1/m) > 0, \\
\Leftrightarrow & (p_\alpha - p_{\alpha_0}) \frac{m \delta(n)}{n} \hat{\sigma}_n^2 + (p_\alpha - p_{\alpha_0}) a \tau K_\rho + o_p(1) > 0, \\
& \Leftrightarrow (p_\alpha - p_{\alpha_0}) a \tau K_\rho + o_p(1) > 0,
\end{aligned}$$

because $m \delta(n) \hat{\sigma}_n^2 / n = o_p(1)$. Since $a \tau K_\rho > 0$ it follows that, for $\alpha \in \mathcal{A}_c$,

$$P(M_n(\alpha) > M_n(\alpha_0)) \rightarrow 1.$$

We now turn our attention to models $\alpha \notin \mathcal{A}_c$. Note that

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n \rho \left(r_i(\hat{\boldsymbol{\beta}}_{\alpha,n}^{R*}) / \hat{\sigma}_n \right) &= \frac{1}{n} \sum_{i=1}^n \rho \left(r_i(\hat{\boldsymbol{\beta}}_{\alpha,n}) / \hat{\sigma}_n \right) \\
&\quad - (\hat{\boldsymbol{\beta}}_{\alpha,n}^{R*} - \hat{\boldsymbol{\beta}}_{\alpha,n})' \frac{1}{n} \sum_{i=1}^n \rho' \left(r_i(\hat{\boldsymbol{\beta}}_{\alpha,n}) / \hat{\sigma}_n \right) \mathbf{x}_{\alpha i} / \hat{\sigma}_n,
\end{aligned}$$

and thus the bootstrap term of $M_{m,n}^{\text{ppe}}(\alpha)$ is

$$E_* \left[\frac{1}{n} \sum_{i=1}^n \rho \left(r_i(\hat{\boldsymbol{\beta}}_{\alpha,n}^{R*}) / \hat{\sigma}_n \right) \right] = \frac{1}{n} \sum_{i=1}^n \rho \left(r_i(\hat{\boldsymbol{\beta}}_{\alpha,n}) / \hat{\sigma}_n \right) + o_p(1). \quad (30)$$

Hence

$$\begin{aligned} P \left(M_{m,n}^{\text{ppe}}(\alpha) > M_{m,n}^{\text{ppe}}(\alpha_0) \right) &= \\ P \left(\frac{1}{n} \sum_{i=1}^n \rho \left(r_i(\hat{\boldsymbol{\beta}}_{\alpha,n}) / \hat{\sigma}_n \right) + o_p(1) > \frac{1}{n} \sum_{i=1}^n \rho \left(r_i(\hat{\boldsymbol{\beta}}_{\alpha_0,n}) / \hat{\sigma}_n \right) \right) &\rightarrow 1. \end{aligned} \quad (31)$$

The result for $M_{m,n}^{\text{pe}}(\alpha)$ follows by noting that, for correct models $\alpha \in \mathcal{A}_c$, from (26) and (28) we have

$$M_{m,n}^{\text{pe}}(\alpha) = \frac{\hat{\sigma}_n^2}{n} \sum_{i=1}^n \rho(r_i(\hat{\boldsymbol{\beta}}_{\alpha,n}) / \hat{\sigma}_n) + \frac{\hat{\sigma}_n^2}{2m} a \tau K_\rho + o_p(1/m),$$

which together with (29) yields, as before,

$$P \left(M_{m,n}^{\text{pe}}(\alpha) - M_{m,n}^{\text{pe}}(\alpha_0) > 0 \right) = P \left((p_\alpha - p_{\alpha_0}) a \tau K_\rho + o_p(1) > 0 \right) \rightarrow 1.$$

For $\alpha \notin \mathcal{A}_c$ note that (30) implies that

$$\begin{aligned} P \left(M_{m,n}^{\text{pe}}(\alpha) > M_{m,n}^{\text{pe}}(\alpha_0) \right) &= \\ P \left(\frac{1}{n} \sum_{i=1}^n \rho \left(r_i(\hat{\boldsymbol{\beta}}_{\alpha,n}) / \hat{\sigma}_n \right) + o_p(1) > \frac{1}{n} \sum_{i=1}^n \rho \left(r_i(\hat{\boldsymbol{\beta}}_{\alpha_0,n}) / \hat{\sigma}_n \right) \right) &\rightarrow 1. \end{aligned}$$

■

References

- Agostinelli, C. (2002). Robust model selection in regression via weighted likelihood methodology. *Statistics and Probability Letters*, **56**, 289-300.
- Akaike, H. (1970). Statistical predictor identification. *Annals of the Institute of Statistical Mathematics*, **22**, 203-217.

- Belsley D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics. Identifying Influential Data and Sources of Collinearity*. New York: Wiley.
- Breiman, L. and Friedman, J.H. (1985). Estimating optimal transformations for multiple regression and correlation, *Journal of the American Statistical Association*, **80**, 580-598.
- Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, **7**, 1-26.
- Efron, B., Johnston, I., Hastie, T. and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, **32**, 407-499.
- Furnival, G., and Wilson, R. (1974). Regression by leaps and bounds. *Technometrics*, **16**, 499-511.
- Gatu, C., and Kontoghiorghes, E.J. (2006). Branch-and-bound algorithms for computing the best subset regression models. *J. Comput. Graph. Statist.*, **15**, 139-156.
- Gunst, G.P. and Mason, R.L. (1980). *Regression Analysis and Its Applications*, New York: Marcel Dekker.
- Hofmann, M., Gatu, C., and Kontoghiorghes, E.J. (2007). Efficient algorithms for computing the best subset regression models for large-scale problems. *Comp. Statist. Data Anal.*, **52**, 16-29.
- Khan, J.A., Van Aelst, S., and Zamar, R.H. (2007a). Building a robust linear model with forward selection and stepwise procedures. *Comp. Statist. Data Anal.*, **52**, 239-248.
- Khan, J.A., Van Aelst, S., and Zamar, R.H. (2007b). Robust Linear Model Selection Based on Least Angle Regression. *J. Amer. Statist. Assoc.*, **102**, 1289-1299.
- Mallows, C.L. (1973). Some comments on C_p . *Technometrics*, **15**, 661-675.
- Maronna, R.A., Martin, D.R. and Yohai, V.J. (2006). *Robust Statistics: Theory*

- and Methods*. West Sussex, England, Wiley.
- Mendes, B. and Tyler, D.E. (1996). Constrained M-estimates for regression. In H. Rieder (Ed.), *Robust Statistics: Data Analysis and Computer Intensive Methods*, Lecture Notes in Statistics, Vol. 109, New York: Springer, pp. 299-320.
- Müller, S. and Welsh, A. H. (2005). Outlier robust model selection in linear regression. *Journal of the American Statistical Association*, **100**, 1297-1310.
- Qian, G. and Künsch, H.R. (1998). On model selection via stochastic complexity in robust linear regression. *Journal of Statistical Planning and Inference*, **75**, 91-116.
- Ronchetti, E. (1985). Robust model selection in regression. *Statistics and Probability Letters*, **3**, 21-23.
- Ronchetti, E. (1997). Robustness aspects of model choice. *Statistica Sinica*, **7**, 327-338.
- Ronchetti, E. and Staudte, R.G. (1994). A robust version of Mallows' C_p . *Journal of the American Statistical Association*, **89**, 550-559.
- Ronchetti, E., Field, C. and Blanchard, W. (1997). Robust linear model selection by cross-validation. *Journal of the American Statistical Association*, **92**, 1017-1023.
- Rousseeuw, P.J. and Yohai, V.J. (1984). Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series*. (J. Franke, W. Hardle and D. Martin, eds.). *Lecture Notes in Statist.*, **26** 256-272. Berlin: Springer-Verlag.
- Salibian-Barrera, M., Van Aelst, S. and Willems, G. (2006). PCA based on multivariate MM-estimators with fast and robust bootstrap. *Journal of the American Statistical Association*, **101**, 1198 - 1211.
- Salibian-Barrera, M., Van Aelst, S. and Willems, G. (2008). Fast and robust

- bootstrap. *Statistical Methods and Applications*, **17**, 41-71.
- Salibian-Barrera, M. and Yohai, V.J. (2006). A fast algorithm for S-regression estimators. *Journal of Computational and Graphical Statistics*, **15**, 414-427.
- Salibian-Barrera, M. and Zamar, R.H. (2002). Bootstrapping robust estimates of regression. *The Annals of Statistics*, **30**, 556-582.
- Schwartz, G. (1978). Estimating the dimensions of a model. *The Annals of Statistics*, **6**, 461-464.
- Seber, G.A.F. (1984). *Multivariate Observations*, New Jersey: Wiley.
- Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, **88**, 486-494.
- Shao, J. (1996). Bootstrap model selection. *Journal of the American Statistical Association*, **91**, 655-665.
- Shao, J. (1997). An asymptotic theory for linear model selection (with discussion). *Statistica Sinica*, **7**, 221-264.
- Sommer, S. and Staudte, R.G. (1995). Robust variable selection in regression in the presence of outliers and leverage points. *Australian Journal of Statistics*, **37**, 323-336.
- Van Aelst, S. and Willems, G. (2005). Multivariate regression S-estimators for robust estimation and inference. *Statistica Sinica*, **15**, 981-1001.
- Wisnowski, J.W., Simpson, J.R., Montgomery, D.C. and Runger, G.C. (2003). Resampling methods for variable selection in robust regression. *Computational Statistics and Data Analysis*, **43**, 341-355.
- Yohai, V.J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, **15**, 642-656.
- Yohai, V.J. and Zamar, R.H. (1988). High breakdown-point estimates of regression by means of the minimization of an efficient scale. *Journal of the American Statistical Association*, **83**, 406-413.