

Fast Robust Estimation of Prediction Error Based on Resampling

Jafar A. Khan^a, Stefan Van Aelst^{b,1}, Ruben H. Zamar^c

^a*Department of Statistics, University of Dhaka Dhaka-1000, Bangladesh*

^b*Dept. of Applied Mathematics and Computer Science, Ghent University, Belgium*

^c*Dept. of Statistics, University of British Columbia, Canada*

Abstract

Robust estimators of the prediction error of a linear model are proposed. The estimators are based on the resampling techniques cross-validation and bootstrap. The robustness of the prediction error estimators is obtained by robustly estimating the regression parameters of the linear model and by trimming the largest prediction errors. To avoid the recalculation of time-consuming robust regression estimates, fast approximations for the robust estimates of the resampled data are used. This leads to time efficient and robust estimators of prediction error.

Key words: Bootstrap, Cross-validation, Prediction error, Robustness

1. Introduction

We focus on the problem of estimating the expected prediction performance of linear models. The model or models under consideration may be obtained from existing theory or practice, or they may be the result of an initial model selection procedure. Once we have determined a moderate number of promising prediction models, it may be of interest to reliably estimate their prediction performance. Because it is not reasonable to attempt predicting the future outliers without knowledge of the underlying mechanism

Email addresses: jkhan66@gmail.com (Jafar A. Khan),

Stefan.VanAelst@UGent.be (Stefan Van Aelst), ruben@stat.ubc.ca (Ruben H. Zamar)

¹Corresponding author. Dept. of Applied Mathematics and Computer Science, Ghent University, Belgium. E-mail: Stefan.VanAelst@UGent.be, phone: +32 9 264 49 08, fax: +32 9 264 49 95.

that produces them, we focus on measuring how well the models predict the future non-outlying cases.

The most common approaches to measure prediction accuracy use resampling. The standard resampling method in this context is cross-validation while bootstrap can be used as an alternative (Efron, 1983). To obtain reliable estimates of the prediction error, quite extensive resampling is required in both the cross-validation and bootstrap procedures (see e.g. Kim, 2009). However, in large scale problems, recalculating a robust fit a large number of times becomes very time consuming. Therefore, our goal is to construct robust versions of cross-validation and bootstrap based measures of prediction error, that can be computed efficiently by avoiding the recalculation of robust estimates for each resample. Note that besides providing information on the prediction performance of models, the proposed estimates of prediction error can also be useful to compare competing linear models based on their prediction performance. The developments in this paper are based on initial results that appeared in the Ph.D. thesis of the first author (Khan, 2006). In a robust PCA context, a related time-efficient cross-validation method to select an optimal number of principal components has been proposed by Hubert and Engelen (2007).

Reliable estimates of prediction error can be useful in the broader context of selecting a stable prediction model when the number d of candidate predictors may be large. In such setting the selection strategy often proceeds in two steps. In the first step, time-efficient methods are applied to drastically reduce the number of candidate predictors and produce a small or moderate number of most promising models. In the second step, more refined techniques are used to select a prediction model from the reduced set of the most promising ones. A reliable estimate of the expected prediction error, such as the ones that we propose here, for each of the models emerging from the first step can be used to make the final selection in the second step.

Selection algorithms that are useful for the screening in the first step are discussed in e.g. Miller (2002); Gatu and Kontoghiorghes (2006); Hofmann, Gatu, and Kontoghiorghes (2007). Computationally efficient techniques inspired by machine learning can be found in e.g. Hastie, Tibshirani, and Friedman (2009). Unfortunately, these selection algorithms yield poor results when the data are contaminated because they try to select covariates that fit well all the cases (including the outliers), and often fail to select the model that would have been chosen if those outliers were not present in the data. Therefore, robust, time-efficient selection algorithms have been

developed by e.g. Khan, Van Aelst, and Zamar (2007a,b); Lutz, Kalisch and Bühlmann (2008); Morgenthaler, Welsch and Zenide (2003); McCann and Welsch (2007). Hence, nowadays several techniques exist to robustly select one or more promising prediction models that can then be investigated more thoroughly in the second step, e.g. by using the procedures proposed here.

A simple approach to robustly estimate the generalization error of competing models that may come to mind is the following. First, carry out a 'full-model' robust fit to obtain weights for the observations. Taking these case weights as fixed, one could then apply a fast weighted-LS cross validation approach to estimate the generalization error of the candidate submodels. A potential flaw of this simple approach is the possibility that, as the predictor subset changes, so does the identification of concordant and outlying cases. Cases that were outlying in the full-model fit might become concordant when some predictors are dropped; cases that were concordant in the full-model fit might become outliers when some predictors are dropped. This mislabeling could have a harmful impact on the estimation of the generalization error. Moreover, our procedure is not limited to situations where there is a natural full model that can be estimated reliably. For instance, we could have a setting where the total number of variables involved in the candidate submodels exceeds the number of observations.

The rest of the paper is organized as follows. Section 2 shortly reviews resampling based prediction measures based on cross-validation and bootstrap. Section 3 introduces time-efficient robust measures of prediction error based on cross-validation and bootstrap. Section 4 presents a Monte Carlo study that compares our robust prediction measures with each other and with the classical ones. Section 5 contains two examples and Section 6 concludes.

2. Resampling based prediction measures

Suppose that we have an $n \times p$ dataset $Z_n = \{z_i = (x_i, y_i), i = 1, 2, \dots, n\}$ randomly sampled from a p -dimensional distribution H , where x_i are the measurements for the $p-1$ predictors and y_i is the observed response for each observation. Following Efron (1983), we want to evaluate a linear prediction rule $\eta(x, \hat{\beta}(Z_n)) = x' \hat{\beta}$ that has been constructed based on the given dataset. Based on this prediction rule, the prediction of a new outcome y_0 corresponding to a given vector of predictor values x_0 is given by $\eta(x_0, \hat{\beta}(Z_n)) = x_0' \hat{\beta}$. Let $Q[y_0, \eta(x_0, \hat{\beta}(Z_n))]$ denote the error in predicting y_0 from x_0 . For example, we can consider the squared loss $Q_{L_2}[y_0, \eta(x_0, \hat{\beta}(Z_n))] = (y_0 - \eta(x_0, \hat{\beta}(Z_n)))^2$.

Then, the true error rate $\text{Err}(Z_n, H)$ of the prediction rule $\eta(x_0, \hat{\beta}(Z_n))$ can be defined as

$$\text{Err}(Z_n, H) = E_H \left(Q[Y_0, \eta(X_0, \hat{\beta}(Z_n))] \right), \quad (1)$$

where the expectation is taken over $(X_0, Y_0) \sim H$ with Z_n fixed at its observed value.

As the distribution H is unknown, in practice we can only estimate (1) from the observed sample Z_n . It is well known that the apparent error rate, given by

$$\overline{\text{err}}(Z_n) = \text{ave } Q[y_i, \eta(x_i, \hat{\beta}(Z_n))],$$

usually underestimates the true error rate $\text{Err}(Z_n, H)$ because the same data have been used both to construct and to evaluate the prediction rule $\eta(x, \hat{\beta}(Z_n))$. Resampling based estimates of the error rate try to alleviate the problem of underestimation of the true error rate.

2.1. Cross-validation

Cross-validation (CV) estimates of the error-rate of a prediction rule are obtained by splitting the n data points into a training sample of size n_t (used for fitting the prediction model) and a validation sample of size $n_v = n - n_t$ (used for assessing the model). Often, k -fold CV is used which means that the data set is split randomly in k blocks of approximately equal size. The training sample then consists of $k - 1$ blocks and the validation sample is given by the left-out block. Each block is left out once, so that a prediction is obtained for each of the observations in the sample. We calculate the average prediction error based on a number R of possible random k -fold splits of the data set, and use it as a criterion to evaluate a prediction model. If the total number of possible k -fold splits is small, then all possible splits can be considered. However, if it is large, then only a subset of random splits is used. Hence, the k -fold CV estimate of $\text{Err}(Z_n, H)$ is given by

$$\widehat{\text{Err}}^{(\text{CV}_k)} = \frac{1}{R} \sum_{r=1}^R \left\{ \frac{1}{n} \sum_{i=1}^n Q[y_i, \eta(x_i, \hat{\beta}(Z_n^{-k_r(i)}))] \right\}, \quad (2)$$

where for each observation i , $Z_n^{-k_r(i)}$ denotes the data set Z_n without the block containing observation i in the r th random run of k -fold CV. By far, the most often used version of cross-validation is the n -fold CV, that is, leave-one-out cross-validation. In this case R is always equal to 1 since the splits are not random anymore. Another common choice is $k = 5$, that is, random 5-fold CV.

2.2. Bootstrap

Efron (1983) considered bootstrap estimators of prediction error as an alternative to cross-validation. It turned out that a simple, well-performing estimator is the .632 bootstrap estimator of prediction error. This estimator is given by

$$\widehat{\text{Err}}^{(.632)} = 0.368 \overline{\text{err}}(Z_n) + 0.632 \widehat{\text{Err}}^{(\text{OOB})}, \quad (3)$$

where $\overline{\text{err}}(Z_n)$ is the apparent error rate as before and $\widehat{\text{Err}}^{(\text{OOB})}$ is the *out-of-bag estimator* of prediction error. The out-of-bag estimator considers for each bootstrap sample Z_n^* only the predictions for the observations of the original data set Z_n that do not appear in the bootstrap sample, i.e. the out-of-bag observations, and then takes the average of all these predictions over a large number of bootstrap samples. Hence, the out-of-bag estimate is given by

$$\widehat{\text{Err}}^{(\text{OOB})} = \frac{1}{B} \sum_{b=1}^B \frac{1}{|b_{\text{oob}}|} \sum_{j \in b_{\text{oob}}} \left(Q[y_j, \eta(x_j, \hat{\beta}(Z_n^{*,b}))] \right), \quad (4)$$

where b_{oob} indicates the indices of the observations not contained in the b th bootstrap sample $Z_n^{*,b}$. Due to the resampling with replacement in the bootstrap, the effective sample size of bootstrap samples can be shown to equal $0.632n$ on average. This smaller sample size affects the accuracy of the predictions for the out-of-bag observations. Hence, while the apparent error rate underestimates the true prediction error, the out-of-bag estimator overestimates the prediction error. The .632 bootstrap estimator (3) tries to balance these two opposite effects to obtain a good estimator of prediction error, see Efron (1983) for details.

3. Fast robust prediction measures

To obtain robust versions of the CV and bootstrap estimates of prediction error discussed in the previous section, the prediction rule $\eta(x, \hat{\beta}(Z_n))$ needs to be robust as well as the summary statistic calculated from the losses $Q[y_i, \eta(x_i, \hat{\beta}(Z'))]$, where Z' are resamples obtained from the original sample Z_n . The robustness of the prediction rule assures that a reliable prediction rule is obtained, even in the presence of outliers in the training sample. The robust summary statistic calculated from the prediction errors guarantees that the estimated error rate is not severely affected by the presence of outliers in the validation sample.

For the robust linear prediction rule $\eta(x, \hat{\beta}(Z_n)) = x' \hat{\beta}^R$, we use MM-estimates $\hat{\beta}^R$ based on Tukey biweight loss functions (Yohai, 1987). However, any other robust estimator based on a smooth loss function (e.g. S-estimates, tau-estimates, etc.) could also be used with straightforward and minimal modifications. We now give a short overview of the definition and properties of MM-estimators. More details can be found in e.g. Maronna, et al. (2006). MM-estimators are based on two score functions ρ_0 and ρ_1 , which determine the breakdown point and the efficiency of the estimator, respectively. More precisely, the MM-estimate $\hat{\beta}^R$ satisfies

$$\frac{1}{n} \sum_{i=1}^n \rho_1' \left(\frac{y_i - x_i' \hat{\beta}^R}{\hat{\sigma}} \right) x_i = \mathbf{0}, \quad (5)$$

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

$$\frac{1}{n} \sum_{i=1}^n \rho_0 \left(\frac{y_i - x_i' \beta}{\hat{\sigma}(\beta)} \right) = b, \quad (6)$$

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

$$\tilde{\beta} = \arg \min_{\beta \in \mathbb{R}^p} \hat{\sigma}(\beta), \quad (7)$$

and is used as an initial value for the iterations that determine $\hat{\beta}$ in (5).

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

$$\rho_c(t) = \begin{cases} 3(t/c)^2 - 3(t/c)^4 + (t/c)^6 & \text{if } |t| \leq c \\ 1 & \text{if } |t| > c, \end{cases} \quad (8)$$

where $c > 0$ is a fixed tuning constant. For the score function ρ_0 of the S-estimator, the choice $\rho_c(t)$ with $c = 1.54764$ together with $b = 1/2$ in (6) yields a 50% breakdown-point scale-estimator that is consistent for normally distributed errors. For the M-estimator, taking the score function $\rho_1(t) = \rho_c(t)$ with $c = 4.685061$ yields a 95%-efficient regression estimator when the errors follow a normal distribution.

Let $r_i = y_i - x_i' \hat{\beta}^R$ and $\tilde{r}_i = y_i - x_i' \tilde{\beta}$; $i = 1, \dots, n$ be the residuals corresponding to the MM-estimate $\hat{\beta}^R$ and the initial S-estimate $\tilde{\beta}$, respectively.

The MM-estimates $\hat{\beta}^R$ and $\hat{\sigma}$ then satisfy the following equations which will be used in the following subsections.

$$\hat{\beta}^R = \left(\sum_{i=1}^n w_i x_i x_i' \right)^{-1} \sum_{i=1}^n w_i x_i y_i, \quad (9)$$

and

$$\hat{\sigma} = \sum_{i=1}^n v_i (y_i - x_i' \hat{\beta}), \quad (10)$$

where the weights w_i and v_i are given by

$$w_i = \rho_1'(r_i/\hat{\sigma})/r_i, \quad i = 1, 2, \dots, n, \quad (11)$$

and

$$v_i = \frac{\hat{\sigma}}{nb} \rho_0(\tilde{r}_i/\hat{\sigma})/\tilde{r}_i, \quad i = 1, 2, \dots, n. \quad (12)$$

As robust summary statistic of the losses $Q[y_i, \eta(x_i, \hat{\beta}(Z'))]$ we use trimmed means, where only the largest losses are trimmed as in Serneels, Filzmoser, Croux and Van Espen (2005). Let $u_1 < u_2 < \dots < u_n$ be the ordered observations of a sample U , and put $k = [n(1 - \alpha)]$, where $[n(1 - \alpha)]$ is the integer part of $n(1 - \alpha)$. Then, the α -trimmed mean $m_\alpha(U)$ of U is the mean that is obtained after dropping the largest $100\alpha\%$ observations of U . That is,

$$m_\alpha(U) = \frac{1}{n - k} \sum_{j=1}^{n-k} u_j. \quad (13)$$

In our setting, α -trimmed means are particularly appealing measures of prediction error because by comparing the α -trimmed mean losses, we can identify the model(s) that can be expected to predict $100(1 - \alpha)\%$ of the future data better than other models. Although the median and quantiles are in general very popular robust summary measures, trimmed means are preferred here because of their higher accuracy and their direct interpretability as explained above.

The trimming level α is a tuning parameter in the calculation of the robust estimate of prediction error. If the training data is a representative random sample, then a reasonable trimming level can be selected by looking at the fraction of outliers in the training data. In some applications one may have a good idea about the (maximal) fraction of outliers that can be expected

in the future, and then this information determines the trimming level. In general, it can be very useful to calculate the robust estimate of prediction error for different trimming levels, so that it is clear how well the model(s) under consideration can predict e.g. 80%, 90%, 95%, and 99% of the future data. If the performance of several models is compared in this way, then this allows us to select the model that accurately predicts the largest fraction of future data.

3.1. Fast robust cross-validation prediction error

In k-fold CV we have to calculate the fit for each of the training samples Z' consisting of $k - 1$ blocks. To avoid the time-consuming recalculation of the robust estimate $\hat{\beta}^R(Z')$ for each of these training samples Z' , we use equations (9) and (11) to calculate an approximation of $\hat{\beta}^R(Z')$, starting from the fit $\hat{\beta}^R(Z_n)$ for the complete dataset.

An initial approximation of the regression MM-estimate $\hat{\beta}^R(Z')$ can be calculated as follows,

$$\hat{\beta}_0(Z') = \left(\sum_{j \in Z'}^n w_j x_j x_j' \right)^{-1} \sum_{j \in Z'}^n w_j x_j y_j, \quad (14)$$

where w_j are the weights of the observations in the complete dataset Z_n , defined in (11). Note that (14) is just a weighted least squares representation and thus no new robust estimate needs to be calculated to obtain $\hat{\beta}_0(Z')$.

The initial estimate $\hat{\beta}_0(Z')$ can be further improved by updating the weights w_j in (14). For the observations in Z' , let $r_j^0 = y_j - x_j' \hat{\beta}_0(Z')$; $j \in Z'$ be the residuals w.r.t. the fit $\hat{\beta}_0(Z')$. Then, using (11), the weights of the observations in the subsample can be updated as

$$w_j^1 = \rho_1'(r_j^0/\hat{\sigma})/r_j^0, \quad j \in Z'. \quad (15)$$

With these updated weights, the new approximation of $\hat{\beta}^R(Z')$ becomes

$$\hat{\beta}_1(Z') = \left(\sum_{j \in Z'}^n w_j^1 x_j x_j' \right)^{-1} \sum_{j \in Z'}^n w_j^1 x_j y_j. \quad (16)$$

With this new fit $\hat{\beta}_1(Z')$ the weights can now be updated again, which in turn leads to a new approximation, $\hat{\beta}_2(Z')$. This process can be iterated further,

but more than two steps do not seem necessary in practice. Note that updating the weights and calculating the new approximations $\hat{\beta}_l(Z')$; $l = 1, 2$ also does not require calculating any new robust estimates. Hence, calculating the successive approximations $\hat{\beta}_l(Z')$; $l = 0, 1, 2$ of $\hat{\beta}^R(Z')$ is computationally very efficient. Moreover, because the fits for the resamples are obtained by using the weights of the observations in the original sample, the fit in each resample will be as robust as the fit for the original, full sample. Indeed, an observation will be downweighted in a resample whenever it is also downweighted in the full sample. This robustness behavior is similar as the robustness of the fast and robust bootstrap of Salibián-Barrera and Zamar (2002).

The fast and robust counterparts of $\widehat{\text{Err}}^{(\text{CV}_k)}$ are now given by

$$\widehat{\text{Err}}^{(\text{FRCV}_k)} = \frac{1}{R} \sum_{r=1}^R \left\{ m_\alpha(Q[y_i, \eta(x_i, \hat{\beta}_l(Z_n^{-k_r(i)}))]) \right\} \quad l = 1, 2. \quad (17)$$

When the one-step approximation $\hat{\beta}_1$ is used in (17), then we call this procedure *fast robust one-step k-fold CV*, and when $\hat{\beta}_2$ is used we call it *fast robust two-step k-fold CV*.

Note that when updating the weights as in (15), we use the scale $\hat{\sigma}$ that was calculated for the original sample Z_n . Let $\hat{\sigma}_{Z'}$ be the S-scale based on the training sample Z' , then we have assumed that $\hat{\sigma}_{Z'} \approx \hat{\sigma}$ which seems reasonable if k is not too small. We have also considered adjusting the scale by using (10) and (12) before updating the weights. However, simulation results (not shown) have indicated that for fast and robust leave-one-out and 5-fold CV, the performance is better without scale adjustment.

3.2. Robust bootstrap prediction error

To obtain a robust version of the .632 bootstrap estimate of the prediction error in (3), we need robust versions of both the apparent error rate and the out-of-bag estimator.

A robust version of the apparent error rate of a robust prediction rule $\eta^R(x_i, Z)$ is directly obtained by calculating the α -trimmed mean of the prediction errors in the original dataset. That is, the robust apparent error rate is given by

$$\overline{\text{err}}_R(Z_n) = m_\alpha \left(Q[y_i, \eta(x_i, \hat{\beta}^R(Z_n))] \right). \quad (18)$$

To obtain robust out-of-bag predictions in a computationally efficient way, we need approximations for the MM-estimates of the regression coefficients in the bootstrap samples. To this end, Salibian-Barrera and Zamar (2002) developed a fast and robust bootstrap procedure. This procedure calculates initial approximations $\hat{\beta}_0^*$ and $\hat{\sigma}_0^*$ to the MM-estimates of a bootstrap sample Z_n^* by applying (9)-(10) on the bootstrap sample, but with the weights w_i^* and v_i^* of each observation in the bootstrap sample equal to the weights of that observation in the original sample Z_n . As before, this does not require calculating any new robust estimates and thus can be performed quickly. To correct for the re-use of the initial weights in each bootstrap sample, Salibian-Barrera and Zamar (2002) derive a linear correction which yields an updated approximation $\hat{\beta}_1^*$ that largely improves the initial approximation $\hat{\beta}_0^*$ (see Salibian-Barrera and Zamar, 2002, for details).

The fast and robust counterpart of the out-of-bag estimate $\widehat{\text{Err}}^{(\text{OOB})}$ in (4) is given by

$$\widehat{\text{Err}}_{\text{FR}}^{(\text{OOB})} = \frac{1}{B} \sum_{b=1}^B m_\alpha \left(Q[y_{b_{\text{oo}b}}, \eta(x_{b_{\text{oo}b}}, \hat{\beta}_1^*(Z_n^{*,b}))] \right), \quad (19)$$

where $b_{\text{oo}b}$ again are the observations not contained in the b th bootstrap sample $Z_n^{*,b}$. Finally, the fast robust .632 bootstrap estimate of the error rate is given by

$$\widehat{\text{Err}}_{\text{FR}}^{(.632)} = 0.368 \overline{\text{err}}_{\text{R}}(Z_n) + 0.632 \widehat{\text{Err}}_{\text{FR}}^{(\text{OOB})}. \quad (20)$$

4. A simulation study

To investigate the behavior of our robust estimators of prediction error, we consider a simulation setting similar as in Khan et al. (2007a). We first create a linear model

$$y = L_1 + 2L_2 + 3L_3 + 4L_4 + 5L_5 + \varepsilon, \quad (21)$$

with 5 latent variables, where L_1, \dots, L_5 are independent standard normal variables. ε is a normal variable with mean 0 and standard deviation $\sigma = \sqrt{55}/2$, which is chosen so that the signal to noise ratio is equal to 2. A set

of $d = 30$ candidate predictors is created as follows. Let

$$X_{ij} = L_i + e_{ij}, \quad i = 1, \dots, 5; j = 1, 2, 3,$$

and

$$X_k = u_k, \quad k = 1, \dots, 15.$$

where all e_{ij} and u_k are independent standard normal variables. The first 15 candidate predictors are active predictors related to the latent variables in (21). Hence, related to each latent variable, there are three active predictor variables which are moderately correlated ($\text{cor}=0.5$). The remaining candidate predictors are noise variables.

We compare the following 6 distinct models:

Model 1: All candidate predictors,

Model 2: All active predictors plus 5 first noise variables,

Model 3: All active predictors (without noise variables),

Model 4: All active predictors related to the three most important latent variables L_3, L_4, L_5 ,

Model 5: Only one active predictor related to each of the 5 latent variables,

Model 6: Only one active predictor related to each of the three most important latent variables.

We considered data without outliers (**Case 1**) and we also considered contaminated data with 10% of bad leverage points. In the contaminated data sets, the clean data are generated as explained above. On the other hand, the 10% of bad leverage points are obtained by generating the errors ε with mean -250 and variance 1, while at the same time all or part of the corresponding predictor values are contaminated. The contaminated candidate predictors are generated with mean 10 and variance 1. We considered the following contamination settings:

Case 2: All predictors are contaminated,

Case 3: Only active predictors are contaminated,

Estimator	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
LS	34.44	31.75	30.62	33.38	44.84	42.90
MM	35.47	32.14	30.92	33.52	44.89	42.96

Table 1: True prediction error for each of the 6 models in the simulation.

Case 4: Only noise predictors are contaminated.

We measured prediction error using squared error losses $Q_{L_2}[y_i, \eta(x_i, \hat{\beta}(Z'))]$. We considered the fast robust one-step (1-FR CV_n) and two-step (2-FR CV_n) leave-one-out 10% trimmed CV procedures as well as the fast robust one-step (1-FR CV_5) and two-step (2-FR CV_5) 5-fold 10% trimmed CV procedures. For comparison, we also included the classical leave-one-out (CV_n) and 5-fold (CV_5) cross-validation based on least squares regression. Finally, we considered the classical (B.632) and fast robustified .632 (FRB.632) bootstrap estimates of the prediction error. For 5-fold CV we averaged over $R = 50$ random runs and for bootstrap we considered $B = 250$ bootstrap samples, such that we have the same number of resamples in both cases. The classical CV and bootstrap procedures use least squares regression in the prediction rule, while the fast robust estimates of prediction error use MM-estimators. The simulations and examples in this paper were performed in R (R Development Core Team, 2009) where we used the MM-estimator as implemented in the R package 'robustbase' (Robustbase Development Team, 2008).

To evaluate the performance of the different estimators, we first determine the expected true prediction error $E_{Z_n}\{E_H(Q[Y_0, \eta(X_0, \hat{\beta}(Z_n))])\}$ for each of the 6 models under consideration. Notice that the expected true prediction error averages over the training samples (Z_n) as well as over future data distributed according to distribution H . In case of squared error loss, it follows from standard calculus that for a given data set Z_n the middle expectation $E_H(Q[Y_0, \eta(X_0, \hat{\beta}(Z_n))]) = E_H[(Y_0 - X_0^t \hat{\beta}(Z_n))^2]$ reduces to $(-\hat{\beta}(Z_n)^t, 1)\Sigma_{X_0, Y_0}(-\hat{\beta}(Z_n)^t, 1)^t$, where Σ_{X_0, Y_0} is the joint covariance matrix of the predictors and response. This joint covariance matrix can easily be calculated for each of the six models described above. To approximate the outer expectation (w.r.t. the training sample Z_n), we took the average over $M = 100$ randomly generated training data sets. The resulting expected true prediction errors are shown in Table 1, both for the prediction rule based on least squares (LS) regression and the prediction rule using the MM-estimator.

For each setting we generated $M = 200$ samples of size $n = 150$ and calculated the estimated prediction error according to the different methods, for each of the six models. In Table 2 we report the average, and between brackets the standard deviation, of the estimated prediction errors of each method for each of the models under the different contamination settings. The top panel of Table 2 shows the results for uncontaminated data (case 1). When comparing these results with the values in Table 1, we see that classical leave-one-out and 5-fold CV both yield reliable estimates of the prediction error. The .632 bootstrap estimator is also reasonably accurate, but tends to slightly underestimate the prediction error in this setting. The fast and robust CV estimates and the FRB.632 estimates are clearly smaller than the expected true prediction errors for all the future data. However, this could be expected because the trimming implies that these estimates express how well the prediction rule can predict the closest 90% of the future data, and not how well the prediction rule can predict all future data. The three bottom panels of Table 2 show the results for the three contamination settings with bad leverage points. The results clearly show the large influence of the outliers on the classical CV and bootstrap estimators of prediction error, both in terms of accuracy and precision. On the other hand, the robust estimators of prediction error are far less influenced by the outliers and still produce reliable estimates for the prediction error of the clean future data. The fast robust CV measures of prediction error seem to perform better than the FRB.632 estimator of prediction error in terms of accuracy towards the targets displayed in Table 1, as well as in terms of precision in case of model 1. Comparing robust leave-one-out and 5-fold CV estimates, we see from Table 2 that 5-fold CV tends to produce slightly larger estimates of the prediction error. The comparison between the one-step and two-step fast and robust CV estimators of prediction error does not reveal a clear winner in this setting. However, in general we expect that the two-step version will be somewhat more reliable because it produces more variety in the regression estimates for the subsamples.

5. Examples

As explained before, robust estimates of prediction error can be used to compare a moderate number of competing models based on their prediction performance (e.g. all subset selection with up to 10 candidate variables).

We consider two examples to illustrate the performance of the fast robust

Method	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
CV_n	34.5 (3.3)	31.8 (2.9)	30.7 (2.8)	33.3 (2.8)	46.8 (3.6)	41.3 (3.3)
1-FR CV_n	21.5 (3.5)	19.9 (3.0)	19.2 (2.9)	20.9 (2.8)	28.0 (3.6)	27.0 (3.4)
2-FR CV_n	21.9 (3.5)	20.1 (3.0)	19.4 (2.9)	20.9 (2.9)	28.1 (3.6)	27.0 (3.4)
CV_5	36.9 (3.4)	33.2 (3.0)	31.6 (2.8)	33.9 (2.8)	47.2 (3.6)	41.7 (3.4)
1-FR CV_5	22.9 (3.6)	20.7 (3.1)	19.8 (2.9)	21.2 (2.9)	28.2 (3.7)	27.2 (3.4)
2-FR CV_5	23.4 (3.7)	21.0 (3.1)	20.0 (2.9)	21.3 (2.9)	28.3 (3.7)	27.3 (3.4)
B.632	31.5 (2.8)	29.9 (2.5)	29.3 (2.4)	32.7 (2.5)	47.2 (3.4)	41.1 (3.2)
FRB .632	28.4 (6.4)	23.1 (3.4)	21.5 (2.1)	23.2 (2.1)	32.6 (2.8)	28.7 (2.5)
CV_n	124.3 (11.0)	201.2 (17.6)	379.6 (32.2)	518.6 (44.9)	663.7 (51.5)	423.3 (38.3)
1-FR CV_n	35.8 (5.1)	32.8 (4.3)	31.4 (4.0)	33.8 (4.3)	45.2 (5.9)	43.5 (5.7)
2-FR CV_n	36.1 (5.1)	32.9 (4.3)	31.5 (4.1)	33.9 (4.4)	45.3 (5.9)	43.5 (5.7)
CV_5	133.1 (11.7)	209.8 (18.0)	391.5 (33.0)	528.6 (45.2)	668.7 (51.6)	428.2 (38.5)
1-FR CV_5	38.6 (5.5)	34.3 (4.6)	32.5 (4.2)	34.5 (4.4)	45.6 (6.0)	44.0 (5.7)
2-FR CV_5	38.9 (5.6)	34.5 (4.6)	32.6 (4.2)	34.6 (4.4)	45.6 (6.0)	44.0 (5.8)
B.632	163.5 (9.8)	268.4 (15.5)	524.0 (29.9)	730.9 (41.5)	951.9 (48.6)	606.5 (36.7)
FRB .632	48.3 (17.5)	32.9 (3.2)	32.0 (4.3)	33.7 (4.1)	46.1 (4.3)	40.6 (4.1)
CV_n	426.6 (36.9)	394.8 (31.6)	379.8 (29.9)	522.2 (44.3)	664.1 (50.3)	425.6 (36.7)
1-FR CV_n	35.7 (5.1)	32.5 (4.4)	31.1 (4.2)	33.7 (4.6)	44.9 (5.5)	42.9 (5.3)
2-FR CV_n	36.0 (5.1)	32.6 (4.4)	31.2 (4.2)	33.8 (4.6)	44.9 (5.5)	42.9 (5.3)
CV_5	457.2 (39.2)	412.2 (32.2)	391.8 (30.6)	532.5 (44.7)	669.3 (50.7)	430.4 (37.0)
1-FR CV_5	38.5 (5.5)	34.0 (4.6)	32.1 (4.3)	34.4 (4.6)	45.2 (5.6)	43.4 (5.4)
2-FR CV_5	38.9 (5.5)	34.2 (4.7)	32.3 (4.3)	34.5 (4.7)	45.3 (5.6)	43.4 (5.4)
B.632	565.3 (32.1)	530.5 (28.2)	525.0 (27.6)	736.8 (41.4)	955.3 (46.3)	607.8 (34.2)
FRB .632	50.0 (37.0)	32.8 (3.90)	30.7 (3.9)	32.9 (3.7)	45.9 (4.2)	40.2 (4.1)
CV_n	84.3 (7.1)	173.7 (14.6)	2083.8 (150.6)	1647.8 (115.6)	1328.0 (69.5)	1343.3 (92.6)
1-FR CV_n	36.1 (5.0)	32.8 (4.3)	31.4 (4.0)	34.0 (4.5)	45.1 (5.5)	43.2 (5.4)
2-FR CV_n	36.4 (5.0)	33.0 (4.3)	31.5 (4.0)	34.1 (4.5)	45.2 (5.5)	43.3 (5.4)
CV_5	90.2 (7.3)	180.9 (14.8)	2396.3 (148.0)	1823.2 (113.3)	1395.5 (69.8)	1437.2 (90.9)
1-FR CV_5	39.0 (5.4)	34.3 (4.5)	32.4 (4.1)	34.7 (4.6)	45.5 (5.6)	43.7 (5.4)
2-FR CV_5	39.4 (5.5)	34.6 (4.6)	32.6 (4.2)	34.8 (4.6)	45.5 (5.6)	43.8 (5.4)
B.632	110.9 (5.8)	230.9 (12.5)	6014.0 (51.7)	5768.4 (41.8)	5896.4 (42.2)	5429.9 (46.0)
FRB .632	44.83 (9.5)	32.9 (3.50)	30.8 (3.2)	32.8 (3.2)	46.0 (4.1)	40.7 (4.0)

Table 2: Average (standard deviation) of the estimated mean squared prediction error for each of the models as obtained by the different estimators of prediction error. Top panel is for clean data (case 1), the next panels correspond to contamination cases 2, 3, and 4 respectively.

estimators of prediction error in practice. To robustly measure prediction error in these examples, we use the fast robust two-step 5-fold cross-validation, because it performed well in the simulation study shown in the previous section and does not yield overly optimistic estimates of prediction error. Moreover, 5-fold CV is expected to be less conservative when models of different dimensions are compared (see e.g. Shao, 1993). The choice of trimming level was based on the fraction of outliers detected by the MM-estimator in the full model. We compare the results of the robust cross-validation with the results obtained by classical 5-fold cross-validation based on least squares regression. To guarantee the stability of the obtained results we used the average of $R = 1,000$ random 5-fold cross-validation runs. Note that prediction error was again measured through squared losses.

The first example uses the pulpfiber data set considered in (Rousseeuw, et al., 2004) which is available in the R package 'robustbase' (Robustbase Development Team, 2008). The data set contains measurements of four candidate predictor variables (pulpfiber properties) and four response variables (paper properties) for $n = 62$ observations. We only use the first response variable which is paper breaking length. We fitted the full model as well as all possible submodels. For each of these models we estimated the fast robust (10% trimming) and classical 5-fold CV prediction errors. The resulting prediction errors are shown in Table 3 for all submodels with at least two predictors. All models with only one predictor yielded estimated prediction errors that were much larger than those of the optimal model. As can be seen from Table 3, the optimal model selected by classical 5-fold CV is the full model with all four predictors. On the other hand, fast and robust 5-fold CV selects a model with the following three predictors: long fiber fraction, fine fiber fraction, and zero span tensile. Table 3 also shows the results for classical cross-validation with 10% trimming which provides robustness at the validation level but not at the estimation level. In this example, 10% trimmed CV selects the same model as the fast robust 10% trimmed CV. It thus seems that the outliers are not very influential at the estimation level but mainly affect the predictions. This will be different in the following example.

To compare the two optimal models with respect to prediction accuracy, we calculated the square root mean squared prediction errors for 5000 random 5-fold cross-validation runs. Figure 1 shows the boxplots based on the estimated prediction errors with 10% trimming (panel a) and without trimming (panel b). From panel a we can clearly see that the optimal robust

	Trimming	1,2,3,4	1,2,3	1,2,4	1,3,4	2,3,4	1,2	1,3	1,4	2,3	2,4	3,4
RCV	10%	0.88	2.72	1.17	0.93	0.84	2.77	3.55	1.19	2.61	1.13	0.93
CV	10%	1.26	2.88	1.36	1.26	1.13	2.79	3.58	1.46	2.74	1.31	1.28
CV	0%	2.68	4.87	2.80	2.87	2.70	4.27	5.21	3.12	4.68	2.85	2.78

Table 3: Estimated 5-fold CV prediction errors for all submodels with at least two predictors for the pulpfiber data. The prediction error of the optimal model in each case is shown in boldface.

model predicts better the majority (90%) of the data. On the other hand, the classical model tries to predict all the data. Since the data contains influential outliers, the classical model yields large prediction errors for all the observations, including the majority of good points as can be seen from panel a. Panel b shows the importance of the trimming in the robust estimation of prediction error. Since the optimal robust model is not affected much by the outliers, it yields very large predictions for the outlying observations which results in the large overall mean squared prediction errors displayed in panel b, reflecting the fact that good prediction of all future observations (including future outliers) is not possible. Note the difference in the scale on the vertical axis between panel a and panel b. The behavior of the optimal robust model as shown in both panels of Figure 1 is consistent with the goal of robust prediction, namely to predict well the majority of good observations while ignoring the outliers in the training and validation data. The use of robust MM-estimation guarantees robustness when fitting the models at the training level while the trimming of the prediction error ensures robustness at the validation level.

The second example uses the well-known Hawkins-Bradru-Kass data set (Hawkins et al., 1984) which is also available in the R package 'robustbase' (Robustbase Development Team, 2008). The data set consists of $n = 75$ observations in 4 dimensions. As before we fitted the full model and all possible submodels. For each of these models we estimated the fast robust (15% trimming) and classical 5-fold CV prediction errors. The results in Table 4 show that both classical and fast robust CV select a model with only 1 predictor. However, while robust CV selects predictor 2, classical CV (with or without trimming) selects predictor 3. To compare the two optimal models with respect to prediction accuracy, we again constructed boxplots of square root mean squared prediction errors for 5000 random 5-fold cross-validation runs. From Figure 2 we can clearly see that the optimal robust

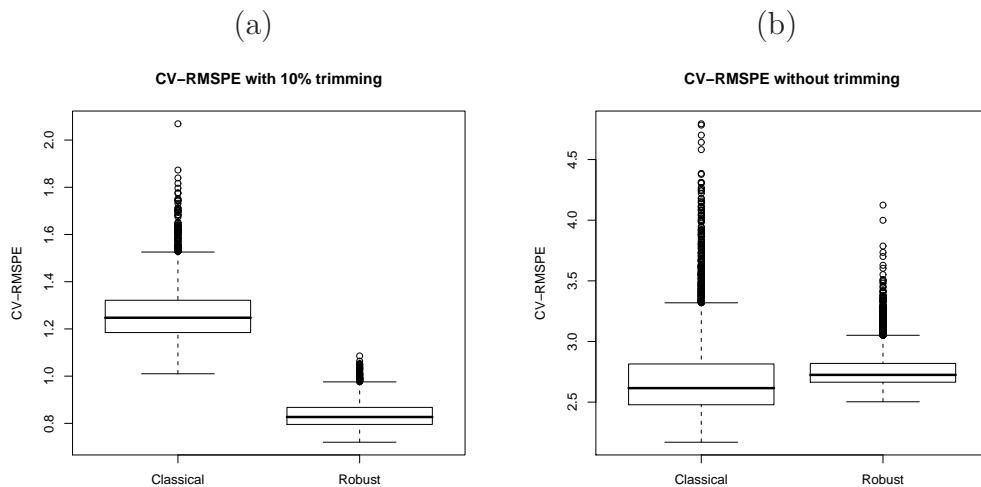


Figure 1: Boxplots of 5000 estimated square roots of mean prediction errors in the optimal models for the pulpfiber data. Panel a shows 10% trimmed mean prediction errors while panel b shows results without trimming. Note the difference in the scale on the vertical axis between panel a and panel b.

	Trimming	1,2,3	1,2	1,3	2,3	1	2	3
RCV	15%	0.311	0.313	0.308	0.312	0.302	0.301	0.305
CV	15%	0.929	1.093	0.679	0.891	1.140	0.809	0.655
CV	0%	6.826	7.183	6.172	6.926	6.268	7.418	6.141

Table 4: Estimated 5-fold CV prediction errors for all submodels for the Hawkins-Bradu-Kass data. The prediction error of the optimal model in each case is shown in boldface.

model again achieves its goal of predicting well the majority (85%) of the data.

6. Conclusions

We proposed several fast and robust estimators for the mean squared prediction error of linear models and showed that these robust estimators perform well with clean data and substantially better than their non-robust counterparts with contaminated data. In particular, the fast and robust CV procedures performed well in our simulation study and examples. Based on these results, we can recommend them for reliable estimation of the prediction error.

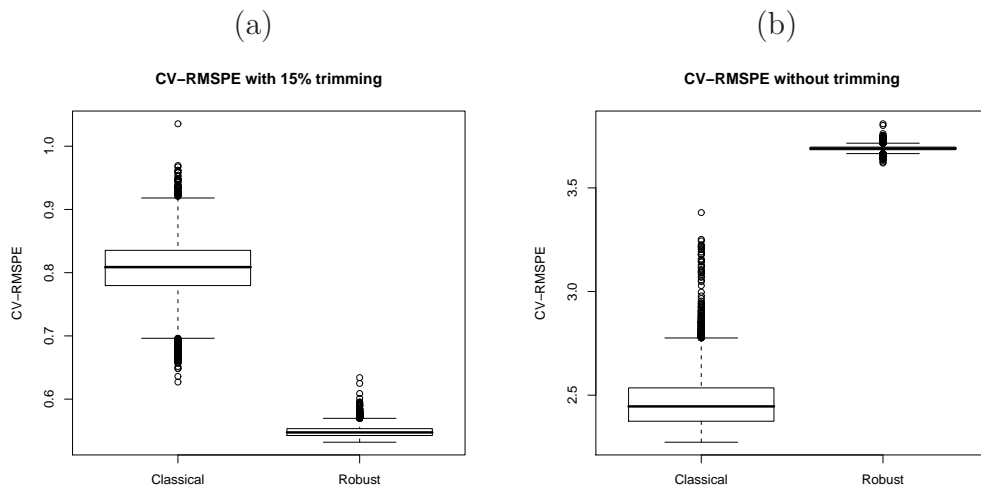


Figure 2: Boxplots of 5000 estimated square roots of mean prediction errors in the optimal models for the Hawkins-Bradru-Kass data. Panel a shows 15% trimmed mean prediction errors while panel b shows results without trimming. Note the difference in the scale on the vertical axis between panel a and panel b.

To make robust procedures useful in practice, it does not only have to be guaranteed that they can resist the effect of outliers, but it must also be able to obtain the result in a reasonable amount of time. Therefore, a considerable amount of effort has been devoted to the development of computationally efficient algorithms for robust methods (see e.g. Salibian-Barrera and Yohai, 2006; Salibian-Barrera, Willems and Zamar, 2008; Loisel and Takane, 2009, for recent contributions). In this paper we proposed robust estimators for the prediction error based on a fast approximation for the robust estimates in each resample, as first proposed by Salibian-Barrera and Zamar (2002) and later successfully extended and applied by e.g. Willems and Van Aelst (2005); Van Aelst and Willems (2005); Salibian-Barrera, Van Aelst, and Willems (2006); Salibian-Barrera and Van Aelst (2008); Roelant, Van Aelst and Croux (2009).

We considered linear prediction rules $\eta(x, \hat{\beta}(Z_n)) = x' \hat{\beta}$ and prediction error based on squared loss $Q_{L_2}[y_0, \eta(x_0, \hat{\beta}(Z_n))]$ in our simulations and examples. However, the approach introduced in this paper can easily be extended to other loss functions, e.g. mean absolute prediction error instead of mean squared prediction error and to prediction rules based on other models, e.g.

nonlinear regression. Moreover, we used MM-estimators in this paper, however, the method is very general and can easily be adapted for other robust estimators based on a smooth loss function. See Hubert, Rousseeuw and Van Aelst (2008) for an overview of such robust regression estimators.

We have shown that trimming is very important to obtain robustness at the validation level when estimating prediction errors. An additional advantage of trimming is its clear interpretation because it reflects what fraction of the new outcomes we are aiming at predicting well. On the other hand, the fraction of trimming is a tuning parameter that needs to be specified by the user. Alternatively, more flexible robust summary measures of the prediction errors could be considered.

7. Acknowledgement

The research of Stefan Van Aelst was 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). The research of Ruben H. Zamar was supported by NSERC.

References

- Efron, B., 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Amer. Statist. Assoc.*, 78, 316-331.
- Hubert, M. and Engelen, S., 2007. Fast cross-validation of high-breakdown resampling algorithms for PCA. *Comp. Statist. Data Anal.*, 51, 5013-5024.
- 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.
- Hastie, T., Tibshirani, R. and Friedman, J. 2009. *The elements of statistical learning* (2nd edition). Springer.
- Hawkins, D.M., Bradu, D. and Kass, G.V., 1984. Location of several outliers in multiple regression data using elemental sets. *Technometrics* 26, 197-208.

- 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.
- Hubert, M., Rousseeuw, P.J. and Van Aelst, S., 2008. High-breakdown robust multivariate methods. *Statist. Science*, 23, 92-119.
- Khan, J.A. 2006. Robust Linear Model Selection for High-Dimensional Datasets. Unpublished doctoral thesis, University of British Columbia, Dept. of Statistics.
- 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.
- Kim, J.-H. 2009. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Comp. Statist. Data Anal.*, 53, 3735-3745.
- Loisel S. and Takane M., 2009. Fast indirect robust generalized method of moments. *Comp. Statist. Data Anal.*, 53, 3571-3579.
- Lutz, R.W., Kalisch, M. and Bühlmann, P., 2008. Robustified L2 boosting. *Comp. Statist. Data Anal.*, 52, 3331-3341.
- Maronna, R.A., Martin, R.D. and Yohai, V.J., 2006. Robust statistics: Theory and methods. John Wiley and Sons.
- McCann, L. and Welsch R.E., 2007. Robust variable selection using least angle regression and elemental set sampling. *Comp. Statist. Data Anal.*, 52, 249-257.
- Miller, A., 2002. Subset Selection in Regression. Chapman & Hall/CRC.
- Morgenthaler, S., Welsch, R. E. and Zenide, A., 2003. Algorithms for robust model selection in linear regression. In: M. Hubert, G. Pison, A. Struyf, and S. Van Aelst (Eds.), *Theory and Applications of Recent Robust Methods*, Birkhäuser-Verlag, Basel (Switzerland), 195-206.

- R Development Core Team (2009). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.
- Robustbase Development Team, 2008. robustbase: Basic Robust Statistics. R package version 0.4-3. URL <http://cran.R-project.org/package=robustbase>.
- Roelant, E., Van Aelst, S. and Croux, C. (2009), Multivariate generalized S-estimators. *J. Multivar. Anal.*, 100, 876–887.
- Rousseeuw, P. J., Van Aelst, S., Van Driessen, K. and Agulló, J. 2004. Robust multivariate regression. *Technometrics*, 46, 293-305.
- 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., Willems, G. and Zamar, R.H., 2008. The fast-tau estimator for regression. *J. Comput. Graph. Statist.*, 17, 659-682.
- Salibian-Barrera, M. and Van Aelst, S. 2008. Robust model selection using fast and robust bootstrap. *Computat. Statist. Data Anal.*, 52, 5121-5135.
- Salibian-Barrera, M., Van Aelst, S. and Willems, G. (2006). PCA based on multivariate MM-estimators with fast and robust bootstrap. *J. Amer. Statist. Assoc.*, 101, 1198–1211.
- Salibian-Barrera, M. and Yohai, V.J. 2006. A fast algorithm for S-regression estimates. *J. Comput. Graph. Statist.*, 15, 414-427.
- Salibian-Barrera, M. and Zamar, R.H. 2002. Bootstrapping robust estimates of regression. *Ann. Statist.*, 30, 556–582.
- Serneels, S., Filzmoser, P., Croux, C. and Van Espen, P. J. 2005. Robust continuum regression. *Chemometrics and Intelligent Laboratory Systems*, 76, 197–204.
- Shao, J. 1993. Linear model selection by cross-validation. *J. Amer. Statist. Assoc.*, 88, 486–494.

- Van Aelst, S. and Willems, G. (2005). Multivariate regression S-estimators for robust estimation and inference. *Statistica Sinica*, 15, 981–1001.
- Willems, G. and Van Aelst, S. (2005). Fast and robust bootstrap for LTS. *Computat. Statist. Data Anal.*, 48, 703–715.
- Yohai, V.J., 1987. High breakdown point and high efficiency robust estimates for regression. *Ann. Statist.*, 15, 642–656.