Fast Robust Model Selection in Large Datasets

DUPUIS, Debbie, VICTORIA-FESER, Maria-Pia

Abstract

Large datasets upon which classical statistical analysis cannot be performed because of the curse of dimensionality are more and more common in many research fields. In particular, in the linear regression context, it is often the case that a huge number of potential covariates are available to explain a response variable, and the first step of a reasonable statistical analysis is to reduce the number of covariates using appropriate statistical criteria. Alternative fast methods that alleviate the problem of computational time with classical procedures have been recently proposed in the literature. However, these fast methods are based on classical statistical theory and are non robust to extreme observations. And, simply replacing the classical statistical criteria by robust ones is not possible because the complexity of the robust estimators and the testing procedures lead to infeasible computations. In this paper, we propose alternative robust estimators, selection criteria and testing procedures for the linear regression model that are fast to compute and hence can be used in a fast model selection procedure. The [...]
Fast Robust Model Selection in Large Datasets*

Debbie J. Dupuis and Maria-Pia Victoria-Feser
HEC Montréal and HEC Genève

Revised March 25, 2010

Abstract

Large datasets upon which classical statistical analysis cannot be performed because of the curse of dimensionality are more and more common in many research fields. In particular, in the linear regression context, it is often the case that a huge number of potential covariates are available to explain a response variable, and the first step of a reasonable statistical analysis is to reduce the number of covariates using appropriate statistical criteria. Alternative fast methods that alleviate the problem of computational time with classical procedures have been recently proposed in the literature. However, these fast methods are based on classical statistical theory and are non robust to extreme observations. And, simply replacing the classical statistical criteria by robust ones is not possible because the complexity of the robust estimators and the testing procedures lead to infeasible computations. In this paper, we propose alternative robust estimators, selection criteria and testing procedures for the linear regression model that are fast to compute and hence can be used in a fast model selection procedure. The robust estimator is a one-step weighted $M$-estimator that can be biased if the covariates are not orthogonal. We show that the bias is relatively small and can be made smaller by iterating the $M$-estimator one or more steps further. In the variable selection process, we propose a simplified robust criterion based on a robust $t$-statistic for significance. We propose a complete algorithm for fast robust model selection, including considerations for huge sample sizes, and show the performance of our method in a simulation study. We also analyze two datasets and show that the results obtained by our method outperform those from robust LARS and random forests. Supplemental materials are available online, see http link.

Keywords: Linear regression, multicollinearity, $M$-estimator, robust $t$-test, partial correlation, LARS, random forests.

*The first author acknowledges the support of the Natural Sciences and Engineering Research Council of Canada. The second author acknowledges the support of the Swiss National Science Foundation (grant no 100012-109532). The authors thank Dr. Susanne Ragg from the Center for Computational Diagnostics at Indiana University for providing access to the protein data as well as Dr. Olga Vitek of the Department of Statistics and the Department of Computer Science at Purdue University for her advice on the analysis of these data. The authors also thank the Editor, the Associate Editor and two referees for comments that greatly improved the presentation of the results.
1 Introduction

Datasets with millions of observations and a huge number of variables are now quite common, especially in finance, management, computer sciences, health sciences, etc. Performing statistical inference with such datasets requires the development of new techniques otherwise the calculations of the necessary statistics are impossible. One group of techniques is concerned with the improvement of the efficiency of algorithms so as to circumvent computer memory problems. Another group of techniques seeks to develop new statistical techniques that can achieve the same goal, i.e. make the calculations feasible. In this paper, we are concerned with the latter.

When the number $p$ of variables is (too) large, good statistical practice requires that this number be reduced by means of proper statistical criteria in order to better understand the phenomenon under investigation. However, standard statistical techniques which often take into account all the variables simultaneously cannot be used when $p$ is too large, either because of the curse of dimensionality or because they involve computations of statistical criteria on all combinations of subsets of variables. For the linear regression problem, simple forward selection procedures that are computationally feasible exist and can be readily used. They involve the estimation of an initial model, to which one or a group of potential explanatory variables is added according to a given criterion, usually based on a test statistic, until the criterion is no longer satisfied. Recently, criteria that take into account the false discovery rate (FDR) have flourished; see e.g. Benjamini and Gavrilov (2009), Gavrilov, Benjamini, and Sakar (2009), Wu, Boos, and Stefanski (2007), Abramovich, Benjamini, Donoho, and Johnstone (2006), Benjamini, Krieger, and Yekutieli (2006), Foster and Stine (2004), etc. These methods are all based on the classical least squares (LS) estimator for the linear regression model, but the criterion can be adapted to other estimators. Basically one needs a computationally fast method to select the best explanatory vari-
ables at each stage of the forward selection procedure and a test statistic for testing
the significance of the regression coefficient of the selected explanatory variable(s).
The $p$-value is then compared to an adaptive (in the number of explanatory variables
already in the model) threshold.

In this paper, we are concerned with the use of such procedures but based on
robust estimators (and hence test statistics) of the linear regression model. As shown
in Ronchetti and Staudte (1994), spurious model deviations such as outliers can lead
to a completely different, and suboptimal, selected model when a non robust criterion
(like Mallows’ $C_p$) is used. This happens because under (slight) data contamination
the estimated model parameters and consequently the model choice criterion can be
seriously biased. The consequence is that decisions (tests) are taken at the wrong
level which has an effect on the computation of the empirical FDR as illustrated
in simulations in §4. The problem is not new and several robust model selections
procedures have been proposed. Global criteria such as the AIC (Akaike 1973), the
BIC (Schwarz 1978) and Mallows’ $C_p$ (Mallows 1973) have been extended to the ro-
bust case by Ronchetti (1982) and Müller and Welsh (2005) for the AIC, Machado
(1993) for the BIC and Ronchetti and Staudte (1994) for Mallows’ $C_p$, respectively.
Ronchetti, Field, and Blanchard (1997) develop a robust criterion based on cross-
validation (CV). All these criteria are very computer intensive when all the possible
models are evaluated which means that only forward selection procedures can be used
in large datasets. Moreover, and even worse, the available robust estimators are just
impossible to compute at the full model if $p$ is too large, and they become useless
in a forward selection procedure as the number of explanatory variables increases
in the selected model. For example, the robust AIC used in Maronna, Martin, and
Yohai (2006) in conjunction with a stepwise selection is too long, or even impossible
to compute, with our data sets. In the first, we seek to predict the abundance of
a protein on the basis of $p=159$ measured covariate values for $n=231$ observations.
In the second we seek to explain the average price asked for a housing unit given $p=1770$ covariate values for $n=13,970$ observations. All available robust estimators with nice properties of efficiency and breakdown cannot be used without some modifications/simplifications.

We propose here a simplified robust estimator and a test statistic that are computationally feasible in high dimensions for estimating and selecting the different models in the forward selection procedure. The basic idea is to use weighted $M$-estimators, weighted partial correlations and robust test statistics based on the weighted $M$-estimator, with weights chosen appropriately for robustness considerations. Our fast robust forward model selection procedure is summarized as follows:

1. Given a selected model at step $k$, select the explanatory variable (among the remaining) with the largest absolute weighted partial correlation to enter the model (see §3.1).
2. Using a robust $z$-type test statistic for significance and its associated $p$-value, end the selection or search for a new explanatory variable to add. The $p$-value is evaluated making use of a FDR correction (see §3.2).
3. Estimate the augmented model using a weighted $M$-estimator (see §2.1) and return to Step 1.

We choose our weights to serve two purposes: (1) robustness considerations; (2) computational speed. This is achieved at the cost of a (small) bias when there is no data contamination. The bias of the weighted $M$-estimator, which is bounded when model deviations such as outliers are present in the data and is negligible when there is no data contamination and the covariates are not too highly collinear, is discussed in §2.2. However, since relatively high collinearity can happen in real datasets, especially when first order interactions are added as potential explanatory variables as in done in our examples, we propose in §3.3 a more computationally intensive adjustment,
but still feasible in high dimensions, that reduces the bias and hence provides a better
model selection procedure in these cases. Computational problems arising because
of the large sample sizes are also dealt with in a simple manner. In §4, simulations
show that the weighted $M$-estimator is fast to compute on large datasets, and with
its associated weighted partial correlation and robust test statistic for significance, we
have a variable selection procedure that is not only feasible, but also provides very
good results in terms of the probability of choosing the correct model.

Finally, when analyzing the large real datasets in §5, we compare our approach
with alternative methods for model selection in linear regression. More specifically, we
consider the least angle regression (LARS) of Efron, Hastie, Johnstone, and Tibshirani
(2004), the robust LARS (RLARS) of Khan, Van Aelst, and Zamar (2007b) as well
as the more general random forests of Breiman (2001).

2 Fast robust estimator for linear regression

2.1 Simple and fast weighted $M$-estimator

An $M$-estimator (Huber 1964, 1967) of the parameters $\beta = [\beta_0, \beta_1, \ldots, \beta_p]^T$ of the
regression model with response variable $y$ and explanatory variables $x = [1 \ x_1 \ldots x_p]^T$
such that $y|x \sim N(x^T \beta; \sigma^2)$, is defined implicitly by

$$\sum_{i=1}^{n} \psi(r_i; c)x_i = 0 \quad (1)$$

for a given sample of $n$ observations, with $\psi$ a chosen function which is at least
bounded, possibly redescending for a high breakdown point robust estimator, $c$ a
tuning constant that controls the relationship between efficiency and robustness and
$r_i = (y_i - x_i^T \beta)/\sigma$ the standardized residuals. For $\psi$, we choose Tukey’s (redescending)
bieight function $\psi(r_i; c) = w(r_i; c)r_i$ based on the weights.
\[
\begin{align*}
    w(r_i; c) &= \begin{cases} 
        \left(\frac{r_i}{c}\right)^2 - 1 & \text{if } |r_i| \leq c, \\
        0 & \text{if } |r_i| > c.
    \end{cases} 
\end{align*}
\] (2)

For better efficiency properties, one can consider a MM-estimator given by the solution in \( \beta \) of (1) in which \( \sigma \) is replaced by an initial scale estimator \( \hat{\sigma} \), e.g. a high breakdown S-estimator (Rousseeuw and Yohai 1984). Throughout the paper, we choose \( c = 4.685 \) which corresponds to an efficiency level of 95% for the robust estimator compared to the LS estimator at the normal model.

The solution to (1) is implicit and hence too computationally intensive in high dimensions. We consider instead a weighted M-estimator of the form

\[
\sum_{i=1}^{n} w_i(r_i^0; c)(y_i - x_i^T \beta)x_i = 0
\] (3)

where the weights \( w_i(r_i^0; c) := w_i^0 \) are computed at residuals \( r_i^0 \) based on an initial robust estimator for \( \beta \) and the scale. We base the latter on robust estimators of all one-covariate sub-models. Namely, let \( \hat{\beta}_1, \ldots, \hat{\beta}_p \) be the robust slope estimators for the \( p \) marginal models \( y = \beta_{01} + x_1\beta_1 + \varepsilon_1, \ldots, y = \beta_{0p} + x_p\beta_p + \varepsilon_p \). Since computational time is not an issue for the univariate sub-models, we use MM-estimators for the \( \hat{\beta}_j \) (together with the \( \hat{\beta}_{0j} \)). Note that in general \( \hat{\beta}_j \) is not a consistent estimator of \( \beta_j \) where \( \beta_j \) corresponds to the model \( y = x^T \beta + \varepsilon \). Let \( w_{ij}, \forall j = 1, \ldots, p \) be the weights (2) computed at the residuals \( r_{ij} = (y_i - \hat{\beta}_{0j} - x_{ij}\hat{\beta}_j)/\hat{\sigma}_j \) with \( \hat{\sigma}_j = 1.483 \text{med}|\tilde{r}_{ij} - \text{med}(\tilde{r}_{ij})| \), the median absolute deviation (MAD) of the residuals \( \tilde{r}_{ij} = (y_i - \hat{\beta}_{0j} - x_{ij}\hat{\beta}_j) \). Now, considering the weighted regressors matrices

\[
X_{w0} = \begin{bmatrix} 1 & \sqrt{w_{i1}x_{i1}} & \cdots & \sqrt{w_{ip}x_{ip}} \end{bmatrix}_{i=1,\ldots,n}
\] (4)

\[
X_{w0}^{w2} = \begin{bmatrix} 1 & w_{i1}x_{i1} & \cdots & w_{ip}x_{ip} \end{bmatrix}_{i=1,\ldots,n}
\] (5)
we define, for \( p < n \), the initial estimator for the model \( y = x^T \beta + \varepsilon \) as

\[
\hat{\beta}^0 = \left[ (X^w_0)^T X^w_0 \right]^{-1} (X^w_0 y^w) \tag{6}
\]

with \( y \) the \( n \)-vector of responses and with corresponding residual scale \( \hat{\sigma}^0 = \text{MAD}(\hat{r}_i) \) with \( \hat{r}_i = y_i - x_i^T \hat{\beta}^0 \). One might be tempted to consider the univariate MM-estimators and their corresponding scale estimates as initial estimators, however although computational time is increased, we favor \( \hat{\beta}^0 \) since: (1) it is otherwise unclear what to consider as intercept and scale estimates; (2) as seen in §2.2, \( \hat{\beta}^0 \) has a smaller bias than an estimator built from univariate \( \hat{\beta}_j \); (3) weights based on residuals standardized by the MAD performed better in the simulation studies.

Using Tukey’s weights \( w^0_i \) in equation (2) with residuals \( r^0_i = (y_i - x_i^T \hat{\beta}^0) / \hat{\sigma}^0 \), our weighted M-estimator can be written as \( \hat{\beta}^1 = \left[ (X^w)^T X^w \right]^{-1} (X^w y^w) \) with

\[
X^w = \text{diag}(\sqrt{w^0_i}) \cdot X \tag{7}
\]
\[
y^w = \text{diag}(\sqrt{w^0_i}) \cdot y. \tag{8}
\]

2.2 Properties of \( \hat{\beta}^1 \)

In this section, we consider in turn the robustness and the consistency properties of \( \hat{\beta}^1 \). For the robustness properties, it is well known that an estimator of the regression slopes is robust if large residuals (and large leverage values in the covariates) are downweighted. The weighting scheme of \( \hat{\beta}^1 \) is based on residuals from an initial robust fit \( \hat{\beta}^0 \) which in turn is based on robust fits of the univariate models \( \hat{\beta}_j \). In a worst case scenario, it is possible that extreme residuals for the full model are not visible in the univariate regressions and might not be correctly downweighted by the robust estimators of the univariate regressions. However, \( \hat{\beta}^1 \) downweights extreme residuals with respect to the full model, so that it is expected that “missed” extreme
residuals for the univariate regressions will be somewhat downweighted at the full model by $\hat{\beta}^1$. Moreover, $\hat{\beta}^0$ is a coordinate-wise robust estimator and Alqallaf, Van Aelst, Yohai, and Zamar (2009) show through the computation of (a generalized version of) the influence function of Hampel (1968, 1974) and different contamination schemes in the (normal) multivariate setting, that coordinate-wise robust estimators can be less sensitive to extreme observations when they occur independently at the univariate level.

The consistency of the regression model estimator is also important, since biased slope estimates can lead to wrongly selected covariates in the forward selection procedure. Our $\hat{\beta}^1$ depends on the $\hat{\beta}_j$ which are biased estimators of the slopes if the covariates are not orthogonal. In http link, we show that at the (true) regression model, under the assumptions that the explanatory variables are centered, i.e. $E[x_j] = 0, \forall j$, and moments higher than 2 for the covariates are nil, the asymptotic bias of $\hat{\beta}^1$ is

$$-\delta \left[ \frac{1}{\gamma_{jj}} \sum_{l \neq j} \gamma_{jl} \hat{\beta}_l \right]_{j=1,\ldots,p}$$

with $\gamma_{jk} = \frac{1}{n} \sum x_{ij} x_{ik}$ and $\delta = 0.0057$. The bias on $\hat{\beta}^0$ is nil. Note that (9) is computed at the uncontaminated model; the effect of model contamination on the possible bias of the estimator is discussed through its robustness properties above.

The relatively small bias depends as expected on the correlation structure between the covariates and is relative to the (true) slopes. In http link the relationship between $\delta$ and $c$ through the efficiency level of the robust estimator is presented. When the efficiency level increases (i.e. $c$ increases), the bias tends to 0, which corresponds to the bias of LS.

It should be noted that the hypothesis that moments higher than 2 for the covariates are nil is there only because otherwise the computation of the bias is simply not tractable. If higher non null moments are present in the data, this might cer-
tainly affect the bias of $\hat{\beta}^1$, but it is not clear how. Actually, in all cases, the bias can be made smaller to even nil if the weighted $M$-estimator serves as an initial estimator of another weighted $M$-estimator in an iterative fashion. Namely, (3) is computed at the updated weights based on the residuals $r_i^{(2)} = (y_i - \mathbf{x}_i^T \hat{\beta}^{(1)}) / \hat{\sigma}^{(1)}$, $\ldots$, $r_i^{(k)} = (y_i - \mathbf{x}_i^T \hat{\beta}^{(k-1)}) / \hat{\sigma}^{(k-1)}$. The (asymptotic) bias of the weighted $M$-estimator iterated $k$ times would be $0.1719254^k \cdot 0.03337403 \left[ \gamma_{jj}^{-1} \sum_{l \neq j} \gamma_{jl} \beta_l \right]$, which converges to 0 as $k$ increases, see [http link].

3 Simple robust variable selection

We propose here to exploit the marginal robust weights computed for constructing $\hat{\beta}^0$ in (6) for estimating robust partial correlations to select the variables to enter in a forward selection procedure as is done in e.g. Foster and Stine (2004). We show that these robust partial correlations arise as a natural information criterion used in significance testing. We also propose a simple method for dealing with large sample sizes. The stopping criterion is based on robust $z$-test statistics for significance testing and takes into account the false discovery rate in the spirit of multiple testing as in Benjamini and Hochberg (1995). When the correlation between the explanatory variables is strong however, the marginal weights are too biased, and in that case a one-step weighted $M$-estimator and corresponding $z$-values for significance can be used. Although the resulting procedure is longer, it is still feasible in large dimensions.

3.1 Selecting the variables

In the standard robust theory for regression models, significance can be tested using a robust Wald-type test statistic (Heritier and Ronchetti 1994) or $z$-statistic given
generally by

\[ z_\psi \text{-ratio} = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \]  

(10)

with \( SE(\hat{\beta}_j)^2 = \text{var}(\hat{\beta})_{jj} \) and for an \( M \)-estimator defined through (1), \( \text{var}(\hat{\beta}) = n^{-1}M(\psi, \Phi)^{-1}Q(\psi, \Phi)M(\psi, \Phi)^{-1} \) with \( M(\psi, \Phi) = -n^{-1}\sum_{i=1}^{n}\int(\partial/\partial\beta^T)\psi(r; c)x_i d\Phi(r) \) and \( Q(\psi, \Phi) = n^{-1}\sum_{i=1}^{n}\int\psi^2(r; c)x_i x_i^T d\Phi(r) \). The variance \( \text{var}(\hat{\beta})_{jj} \) can be estimated using

\[ \hat{SE}(\hat{\beta}_j)^2 = \frac{\hat{\sigma}^2}{n} \left[ \left( \frac{1}{\sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i x_i^T} \right)^{-1} \right] e_c^{-1} \]  

(11)

(see e.g. Maronna et al. 2006) where the weights \( w_i \) are the robust weights corresponding to the \( M \)-estimator on the model with \( x \) and \( e_c \) its efficiency. With Tukey’s bi-weight function, the weights are (2) and we have \( e_c = \left[ \int_{-c}^{c} (5r^4/c^4 - 6r^2/c^2 + 1) d\Phi(r) \right]^2 / \int_{-c}^{c} r^2(r^2/c^2 - 1)^4 d\Phi(r) \) (see e.g. Heritier, Cantoni, Copt, and Victoria-Feser 2009, p. 57). Using (10) instead of the standard \( t \)-test guarantees that the level \( \alpha \) at which the significance test is performed remains stable under model contamination. Indeed, \( \alpha \) is specified through \( P(|T| > t_\alpha) = \alpha \) using the (asymptotic) distribution of the test statistic \( T \) and corresponding quantile \( t_\alpha \) under the hypothesis that the data upon which the test statistic is computed are generated exactly by the postulated model. If this is not the case, then the (asymptotic) distribution is not correct and consequently neither is the level (for a more formal treatment and examples, see e.g. Heritier et al. 2009, §2.4). The price to pay for a robust test statistic is however a (relatively small) loss of power when there is no model contamination. This loss is quantified in the simulation study of §4 when we compare the performance of classical and robust selection procedures under no model contamination.

In order to be able to compute a robust test statistic for \( H_0 : \beta_j = 0 \) in a forward selection procedure in high dimensions, we need to adapt the \( z_\psi \)-ratio so that it can be computed rapidly with the information available at the current forward selection.
At a given stage of the forward selection procedure, we have a model including $S$ explanatory variables (plus intercept) $\mathbf{x}_S$ to which we can add one of the remaining explanatory variables, say $z_j$. We also have $\hat{\beta}_j^1$, the weighted $M$-estimator on the sub-model with $\mathbf{x}_S$, with corresponding weights $w_{iS}$ and scale estimate $\hat{\sigma}^2$, and the $MM$-estimator $\hat{\beta}_j$ of the simple linear regression model with $z_j$ (and intercept) with corresponding weights $w_{ij}$. Let $\mathbf{X}_S^w = \text{diag}(\sqrt{w_{iS}^0}) \mathbf{X}_S$ where $\mathbf{X}_S$ is the $n \times (S+1)$ design matrix for the sub-model with $\mathbf{x}_S$, $\mathbf{z}_j^w = \text{diag}(\sqrt{w_{ij}}) \mathbf{z}_j$ and $\mathbf{y}_j^w = \text{diag}(\sqrt{w_{ij}}) \mathbf{y}$.

In http link, we show that the robust test statistic for testing the significance of $\beta_j$ in the model with $[\mathbf{x}_S^T, z_j]^T$, can be reasonably approximated by

$$z_{\psi}\text{-ratio}^2 \approx \frac{1}{\hat{\sigma}^2} \sum_{i=1}^n w_{ij} c_i y_j^{wT} z_j^{w} \phi^w_{\mathbf{YZ}|\mathbf{X}}$$

where

$$\phi^w_{\mathbf{YZ}|\mathbf{X}} = (z_j^{wT} z_j^{w} - z_j^{wT} H_S z_j^{w})^{-1} z_j^{wT} (\mathbf{I} - H_S) y_j^w$$

and $H_S = \mathbf{X}_S^w [\mathbf{X}_S^{wT} \mathbf{X}_S^w]^{-1} \mathbf{X}_S^{wT}$. Equation (13) is the partial correlation between $y_j^w$ and $z_j^w$ given $\mathbf{X}^w$. Note that $z_j^{wT} (\mathbf{I} - H_S) y_j^w$ is $n - 1$ times the estimated partial covariance between $y_j^w$ and $z_j^w$ given $\mathbf{X}^w$ and $z_j^{wT} z_j^w - z_j^{wT} H_S z_j^w$ is the partial variance of $z_j^w$ given $\mathbf{X}^w$. Hence, the predictor $z_j$ with the (in absolute value) largest (weighted) partial correlation offers the greatest reduction in the (weighted) residual sum of squares and is selected in a robust forward stepwise regression, analogously to the approach followed by Foster and Stine (2004) in the classical setting.

When both the number of regressors and the sample size are very large, it may be advantageous, from a computational perspective, to take a subsampling approach. That is, we will choose candidate regressors to include in the model based on a computation of partial correlations on random subsamples of the data. We propose the following. Select a subsample of size $n_n$ among the $n$ observations and compute
weighted partial correlations following (13). For the variable with the largest absolute weighted partial correlation, compute its weighted partial correlation using all the data. Repeat the process $n_s$ times and choose to add the variable with the largest absolute (full sample) weighted partial correlation.

### 3.2 Stopping criterion

We propose a stopping criterion based on the robust significance test of the variable selected according to its weighted partial correlation. Denote by $z_j$ the variable that is added to the subset $x_S$ and by $\hat{\beta}_j^1$ the corresponding weighted $M$-estimator computed using $x_{S+1} = [x_S^T \ z_j]^T$, with corresponding weights $w_i^0$. We use $z_\psi$-ratio $= \hat{\beta}_j^1 / SE(\hat{\beta}_j^1)$ as the test statistic for testing the significance of the regression parameter $\beta_j$ of the regressor $z_j$ newly entered into the model. The standard deviation $SE(\hat{\beta}_j^1)$ is computed using (11) with $x_i = x_{S+1,i}$ and weights $w_i^0$, and $\hat{\sigma}$ as the MAD of the residuals.

We proceed in a step-down manner (see Foster and Stine (2004) for full motivation and theory) so that the $(S+1)$th suggested regressor is added if

$$|z_\psi\text{-ratio}| > \Phi^{-1}\left(1 - \frac{S + 1 \ \alpha}{p \ \alpha} \right).$$

Use of $(S+1)/p$ is shown to control FDR in Benjamini and Hochberg (1995). Because the weighted $M$-estimator has a (small) bias, the standard normal distribution $\Phi$ might not be appropriate. We do not expect this bias to have important consequences in practice, especially with nearly orthogonal to moderately correlated covariates. If bias is an issue on the stopping criterion, then the weighted $M$-estimator can be iterated fully (a few steps suffice), at the cost of more computational time (see §5.1).
3.3 Heavily Correlated Covariates

The robust weighted $M$-estimator and the selection procedure based on weighted partial correlations using marginal weights allow fast robust estimation and model selection in high dimensions. The cost for this dramatic gain in computational time is the bias of the marginal $MM$-estimators and weights, then to a lesser extent of the weighted $M$-estimator, and consequently of the weighted partial correlations used as a criterion for selecting the potentially best covariates to enter the model. The marginal $MM$-estimators for say the regression model with $[1 \ x_j]$ have a bias that is proportional to $\sum_{l \neq j} \gamma_{jl} \beta_l$ (see (S.12) in http link) and which is larger than the bias of the weighted $M$-estimator including several explanatory variables (see (9)). The bias is proportional to the correlation structure of the covariates, and the higher the latter, the larger the bias and consequently the selection procedure may be affected. In other words, when collinearity is high among the covariates, the bias on the estimated marginal regression coefficients and weights can be large and the information provided by the weighted partial correlations can lead to wrong selection or even to too short a selection (the stopping criterion is met after too few steps). An alternative approach that is more time consuming but still feasible in high dimensions is to compute the weighted $M$-estimator for all potential models (those augmented by one variable at a time) and then use the associated weights for the computation of the weighted partial correlations. More precisely, one replaces in (12) and (13) the (marginal) weights $w_{ij}$ by the weighted $M$-estimator weights $w_{i,S+1}^0$ on the augmented models. This amounts to selecting the covariate with the largest (in absolute value) $z_{\psi}$-ratio. The stopping criterion is then applied to this $z_{\psi}$-ratio.

To deal with large sample size, we use a sampling technique that computes the weighted $M$-estimators for $n_s$ subsamples of size $n_s$ to find candidate regressors, from which the best candidate is selected based on the all-data weighted $M$-estimator.
4 Simulation Study

We carry out a simulation study to assess the effectiveness of the model selection approaches outlined above and to showcase the strengths of our proposed robust weighted $M$-estimator. First, we create a linear model

\[ y = L_1 + L_2 + \ldots + L_k + \sigma \varepsilon \]  

(14)

where $L_1, L_2, \ldots, L_k$ are multivariate normal (MVN) with $E(L_i) = 0$, $\text{Var}(L_i) = 1$, and $\text{corr}(L_i, L_j) = \rho$, $i \neq j, i,j = 1,\ldots,k$, and $\varepsilon$ an independent standard normal variable. We choose $\rho$ to produce a range of theoretical $R^2$ values ($R^2 = (\text{Var}(y) - \sigma^2)/\text{Var}(y)$ for (14)) and $\sigma$ to give $z$ values for our target regressors of about 6 under normality as in Ronchetti et al. (1997). A set of $p$ candidate regressors is created as follows. The first $k$ covariates are

\[ X_i = L_i \quad i = 1, \ldots, k. \]

Let $e_{k+1}, \ldots, e_p$ be independent standard normal variables and use the first $2k$ to give the $2k$ covariates

\[
\begin{align*}
X_{k+1} &= L_1 + \lambda e_{k+1}, & X_{k+2} &= L_1 + \lambda e_{k+2}, \\
X_{k+3} &= L_2 + \lambda e_{k+3}, & X_{k+4} &= L_2 + \lambda e_{k+4}, \\
\vdots & & \vdots \\
X_{3k-1} &= L_k + \lambda e_{3k-1}, & X_{3k} &= L_k + \lambda e_{3k};
\end{align*}
\]

and the final $p - 3k$ to give the $p - 3k$ covariates

\[ X_i = e_i, \quad i = 3k + 1, \ldots, p. \]

The covariates $X_1, \ldots, X_k$ are our target covariates. Variables $X_{k+1}, \ldots, X_{3k}$ are noise covariates that are correlated with our target covariates, and variables $X_{3k+1}, \ldots, X_p$ are independent noise covariates.
We consider samples without and with contamination. Samples with no contamination are generated using \( \varepsilon \sim N(0,1) \). To allow for 5% outliers, we generate using \( \varepsilon \sim 95\%N(0,1) + 5\%N(30,1) \). These contaminated cases also have high leverage \( X \)-values: \( L_1, \ldots, L_k \sim \text{MVN} \) as before, except \( \text{Var}(L_i) = 5, i = 1, \ldots, k \). This represents the most difficult contamination scheme: large residuals at high leverage points.

We choose \( \lambda = 3.18 \) so that \( \text{corr}(X_1, X_{k+1}) = \text{corr}(X_1, X_{k+2}) = \text{corr}(X_2, X_{k+3}) = \ldots = \text{corr}(X_k, X_{3k}) = 0.3 \). First, we consider a small number of regressors and a small sample size so as to compare our proposals with another robust selection method. We consider a sample of size \( n = 200, p = 10 \) candidate regressors with \( k = 3 \) target covariates. We consider our forward selection methods with \( \alpha = 0.05 \), robust AIC as implemented in \texttt{step.lmRob} of the \texttt{robust} package for R, and LS. We consider both low and high signal-to-noise ratios. Results are in Table 1.

Table 1: Model selection results. Simulated data, as described in §4, have \( n = 200 \) observations with \( p = 10 \) potential regressors, including \( k = 3 \) target regressors with \( \text{corr}(X_1, X_2) = \text{corr}(X_1, X_3) = \text{corr}(X_2, X_3) = \rho = 0 \) (\( R^2 = 0.35 \)) and \( \rho = 0.76 \) (\( R^2 = 0.80 \)), and \( \text{corr}(X_1, X_{k+1}) = \text{corr}(X_1, X_{k+2}) = \text{corr}(X_2, X_{k+3}) = \ldots = \text{corr}(X_k, X_{3k}) = 0.3 \) in both cases. Methods r1, r2, and LS are based on forward selection with FDR stopping with candidates for entry selected following r1: maximizing robust partial correlation; r2: maximizing robust \( z_\psi \)-ratio; LS: maximizing partial correlation. r3 is backward selection based on robust AIC and robust linear regression with high breakdown point and high efficiency regression as implemented in \texttt{step.lmRob} of the \texttt{robust} package for R. Empirical FDR appears in the last row. Mean execution times for the r1, r2, r3, and LS methods are \( \sim 0.60, 0.9, 3.0 \) and 0.1 s, respectively. Results are based on 200 simulations for each configuration.

\[
\begin{array}{cccccccccccc}
 & \multicolumn{2}{c}{R^2 = 0.35} & & \multicolumn{2}{c}{R^2 = 0.80} \\
 & \text{no outliers} & \text{5\% outliers} & & \text{no outliers} & \text{5\% outliers} \\
\hline
\text{Correct} & \begin{array}{cccc}
176 & 174 & 60 & 183 \\
169 & 167 & 54 & 0 \\
182 & 174 & 60 & 183 \\
179 & 166 & 54 & 4 \\
\end{array} & \begin{array}{cccc}
126 & 124 & 60 & 183 \\
167 & 167 & 54 & 0 \\
182 & 174 & 60 & 183 \\
179 & 166 & 54 & 4 \\
\end{array} & & \begin{array}{cccc}
176 & 174 & 60 & 183 \\
169 & 167 & 54 & 0 \\
182 & 174 & 60 & 183 \\
179 & 166 & 54 & 4 \\
\end{array} & & \begin{array}{cccc}
126 & 124 & 60 & 183 \\
167 & 167 & 54 & 0 \\
182 & 174 & 60 & 183 \\
179 & 166 & 54 & 4 \\
\end{array} \\
\text{Extra} & \begin{array}{cccc}
24 & 26 & 140 & 17 \\
31 & 34 & 146 & 0 \\
18 & 26 & 140 & 17 \\
20 & 33 & 146 & 0 \\
\end{array} & & \begin{array}{cccc}
0 & 1 & 0 & 16 \\
0 & 1 & 0 & 16 \\
0 & 1 & 0 & 16 \\
0 & 1 & 0 & 16 \\
\end{array} & & \begin{array}{cccc}
0 & 1 & 0 & 16 \\
0 & 1 & 0 & 16 \\
0 & 1 & 0 & 16 \\
0 & 1 & 0 & 16 \\
\end{array} & & \begin{array}{cccc}
0 & 1 & 0 & 16 \\
0 & 1 & 0 & 16 \\
0 & 1 & 0 & 16 \\
0 & 1 & 0 & 16 \\
\end{array} \\
\text{Missing 1} & \begin{array}{cccc}
0 & 0 & 0 & 0 \\
0 & 1 & 0 & 16 \\
0 & 1 & 0 & 16 \\
0 & 1 & 0 & 16 \\
\end{array} & & \begin{array}{cccc}
0 & 0 & 0 & 0 \\
0 & 1 & 0 & 16 \\
0 & 1 & 0 & 16 \\
0 & 1 & 0 & 16 \\
\end{array} & & \begin{array}{cccc}
0 & 0 & 0 & 0 \\
0 & 1 & 0 & 16 \\
0 & 1 & 0 & 16 \\
0 & 1 & 0 & 16 \\
\end{array} & & \begin{array}{cccc}
0 & 0 & 0 & 0 \\
0 & 1 & 0 & 16 \\
0 & 1 & 0 & 16 \\
0 & 1 & 0 & 16 \\
\end{array} \\
\text{Missing 2} & \begin{array}{cccc}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
\end{array} & & \begin{array}{cccc}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
\end{array} & & \begin{array}{cccc}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
\end{array} & & \begin{array}{cccc}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
\end{array} \\
\text{Other} & \begin{array}{cccc}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
\end{array} & & \begin{array}{cccc}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
\end{array} & & \begin{array}{cccc}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
\end{array} & & \begin{array}{cccc}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
\end{array} \\
\hline
\%FDR & 3.4 & 3.6 & 26 & 2.2 & 4.1 & 4.4 & 26 & 2.2 & 2.6 & 3.7 & 26 & 2.2 & 2.5 & 4.4 & 26 & 14
\end{array}
\]
Entries in the top panel give the actual number of runs falling into each category. The category “Correct” means that the correct model was chosen and is the key measure of the performance. “Extra” means that a model was chosen for which the true model is a proper subset. “Missing 1” means that the model chosen differed from the true model only in that it was missing one of the target covariates; “Missing 2” is defined analogously. At a 95% level of confidence, Monte Carlo variation (MCV) on entries with value around 190 are ±8.5, and no entry has MCV greater than ±14. We also report the empirical FDR: the proportion of noise covariates among selected covariates, see Benjamini and Hochberg (1995).

Both our robust selection methods perform very well in terms of proportion of correctly selected models: (1) performing as well as LS when there are no outliers, i.e. there is very little loss of power, and (2) outperforming LS when there are outliers. The backward selection using robust AIC is not very good in all cases; while it often yields a model which includes the correct model, the final model also contains many noise covariates. The classical LS approach fails miserably in the presence of outliers and rarely chooses the correct model. The empirical FDR brings another insight. The latter is stable across all data configurations for our robust method based on maximizing the robust $z_\psi$-ratio ($r^2$) with a value of around 4%. For our robust method based on robust partial correlations, there is a (small) drop in the FDR when the target covariates are heavily correlated ($R^2 = 0.80$). This is likely due to the impact of multicollinearity on the bias of the $M$-estimator and/or on the use of marginal weights in (13). It is for this reason that we also proposed our $z_\psi$-ratio approach. For LS, the FDR is reasonable without data contamination (but no better than for both our robust methods), but reaches large values with data contamination. The backward selection using robust AIC has a too large FDR in all settings.

Now, we consider a sample of size $n = 1000$, $p = 100$ candidate regressors with $k = 5$ target covariates. Robust AIC is not a viable approach here as selection
Table 2: Forward selection results. As in Table 1, except $n = 1000$, $p = 100$, $k = 5$, $\rho = 0.10$ ($R^2 \approx .20$) and $\rho = 0.85$ ($R^2 \approx .80$). Subsampling approach uses $n_s$ samples of size $n_n$ and is applied to r1, r2 and LS methods in middle and bottom panels.

### $R^2 \approx 0.20$

<table>
<thead>
<tr>
<th></th>
<th>no outliers</th>
<th>5% outliers</th>
<th>no outliers</th>
<th>5% outliers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>158 156 160</td>
<td>142 132 0</td>
<td>181 156 160</td>
<td>171 135 0</td>
</tr>
<tr>
<td>Extra</td>
<td>42 44 39</td>
<td>44 50 0</td>
<td>17 42 39</td>
<td>20 54 0</td>
</tr>
<tr>
<td>Missing 1</td>
<td>0 0 1</td>
<td>13 11 1</td>
<td>2 2 1</td>
<td>9 10 1</td>
</tr>
<tr>
<td>Missing 2</td>
<td>0 0 0</td>
<td>0 0 3</td>
<td>0 0 0</td>
<td>0 0 1</td>
</tr>
<tr>
<td>Missing 3</td>
<td>0 0 0</td>
<td>0 0 12</td>
<td>0 0 0</td>
<td>0 0 12</td>
</tr>
<tr>
<td>Other</td>
<td>0 0 0</td>
<td>1 7 184</td>
<td>0 0 0</td>
<td>0 1 186</td>
</tr>
<tr>
<td>%FDR</td>
<td>4.1 4.4 3.7</td>
<td>4.2 6.0 28</td>
<td>1.5 4.1 3.7</td>
<td>1.7 5.4 4.5</td>
</tr>
<tr>
<td>time</td>
<td>28 54 8.7</td>
<td>28 53 2.6</td>
<td>28 56 8.7</td>
<td>28 55 2.8</td>
</tr>
</tbody>
</table>

### $R^2 \approx 0.80$

<table>
<thead>
<tr>
<th></th>
<th>no outliers</th>
<th>5% outliers</th>
<th>no outliers</th>
<th>5% outliers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>144 101 130</td>
<td>144 90 0</td>
<td>183 117 128</td>
<td>182 110 0</td>
</tr>
<tr>
<td>Extra</td>
<td>18 18 14</td>
<td>16 20 0</td>
<td>15 14 12</td>
<td>9 19 0</td>
</tr>
<tr>
<td>Missing 1</td>
<td>34 67 48</td>
<td>36 64 2</td>
<td>2 66 58</td>
<td>9 68 1</td>
</tr>
<tr>
<td>Missing 2</td>
<td>2 10 6</td>
<td>2 17 3</td>
<td>0 2 1</td>
<td>0 2 1</td>
</tr>
<tr>
<td>Missing 3</td>
<td>0 1 1</td>
<td>0 3 10</td>
<td>0 0 0</td>
<td>0 0 10</td>
</tr>
<tr>
<td>Other</td>
<td>2 3 1</td>
<td>2 6 185</td>
<td>0 1 1</td>
<td>0 1 188</td>
</tr>
<tr>
<td>%FDR</td>
<td>1.7 2.0 1.3</td>
<td>1.7 2.5 27</td>
<td>1.3 1.3 1.2</td>
<td>0.8 1.7 2.3</td>
</tr>
<tr>
<td>time</td>
<td>22 59 3.2</td>
<td>23 60 1.4</td>
<td>23 59 3.2</td>
<td>22 60 1.3</td>
</tr>
</tbody>
</table>

### subsampling with $n_n = 100$, $n_s = 10$

<table>
<thead>
<tr>
<th></th>
<th>no outliers</th>
<th>5% outliers</th>
<th>no outliers</th>
<th>5% outliers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>140 92 138</td>
<td>140 90 0</td>
<td>183 117 128</td>
<td>182 110 0</td>
</tr>
<tr>
<td>Extra</td>
<td>16 15 14</td>
<td>15 20 0</td>
<td>15 14 12</td>
<td>9 19 0</td>
</tr>
<tr>
<td>Missing 1</td>
<td>34 67 48</td>
<td>36 64 2</td>
<td>2 66 58</td>
<td>9 68 1</td>
</tr>
<tr>
<td>Missing 2</td>
<td>2 10 6</td>
<td>2 17 3</td>
<td>0 2 1</td>
<td>0 2 1</td>
</tr>
<tr>
<td>Missing 3</td>
<td>0 1 1</td>
<td>0 3 10</td>
<td>0 0 0</td>
<td>0 0 10</td>
</tr>
<tr>
<td>Other</td>
<td>2 3 1</td>
<td>2 6 185</td>
<td>0 1 1</td>
<td>0 1 188</td>
</tr>
<tr>
<td>%FDR</td>
<td>1.7 2.0 1.3</td>
<td>1.7 2.5 27</td>
<td>1.3 1.3 1.2</td>
<td>0.8 1.7 2.3</td>
</tr>
<tr>
<td>time</td>
<td>22 59 3.2</td>
<td>23 60 1.4</td>
<td>23 59 3.2</td>
<td>22 60 1.3</td>
</tr>
</tbody>
</table>

### subsampling with $n_n = 100$, $n_s = 50$

<table>
<thead>
<tr>
<th></th>
<th>no outliers</th>
<th>5% outliers</th>
<th>no outliers</th>
<th>5% outliers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>160 156 161</td>
<td>147 130 0</td>
<td>181 156 161</td>
<td>173 137 0</td>
</tr>
<tr>
<td>Extra</td>
<td>40 36 39</td>
<td>39 47 0</td>
<td>17 38 36</td>
<td>18 46 0</td>
</tr>
<tr>
<td>Missing 1</td>
<td>0 3 2</td>
<td>13 16 2</td>
<td>2 6 3</td>
<td>9 16 1</td>
</tr>
<tr>
<td>Missing 2</td>
<td>0 0 0</td>
<td>0 0 3</td>
<td>0 0 0</td>
<td>0 0 1</td>
</tr>
<tr>
<td>Missing 3</td>
<td>0 0 0</td>
<td>0 0 13</td>
<td>0 0 0</td>
<td>0 0 12</td>
</tr>
<tr>
<td>Other</td>
<td>0 1 1</td>
<td>1 7 182</td>
<td>0 0 0</td>
<td>0 1 186</td>
</tr>
<tr>
<td>%FDR</td>
<td>3.9 4.1 3.6</td>
<td>3.8 5.7 28</td>
<td>1.5 3.6 3.5</td>
<td>1.5 4.6 3.6</td>
</tr>
<tr>
<td>time</td>
<td>30 248 12</td>
<td>30 244 3.2</td>
<td>31 255 12</td>
<td>30 250 3.3</td>
</tr>
</tbody>
</table>
takes over 20 hours on one sample. Results for the other methods are in Table 2. Computations are carried out in R and the largest contributor to the larger time for the robust analysis is the time required to compute all $p$ marginal robust models using \texttt{lmRob} (to obtain the required starting values). The work that follows is only slightly more computationally expensive when we use robust partial correlations to choose the candidate regressors for entry with our robust estimator (with the obvious gains shown in Table 2). In terms of correctly selected models and empirical FDR, the conclusions on the performance comparison between the three methods is the same as with the smaller scale simulation. Both robust methods are very stable across settings, with a relative performance loss for the our robust method based on robust partial correlations in the highly correlated case ($R^2 \approx 0.80$). LS performs miserably with data contamination, never choosing the correct model or a larger model of which the correct model is a subset. It does manage however to maintain a respectable FDR in the $R^2 \approx 0.80$ case, often choosing a one-variable (target) model. On average, the robust estimators assign weights below 0.20 to only 0.6% of the observations when there are no outliers, and 6.4% of the observations when there are 5% outliers.

Moreover, the results clearly show that subsampling can work just as well as considering the full sample, suffices to take an adequate number of subsamples, and that the forward selection is sufficient. What we cannot appreciate given the times in Table 2 is that subsampling is really faster when $p$ and $n$ are really large, and sometimes the only option, e.g. we cannot proceed without subsampling (due to lack of resident memory on most machines) in the census data example analyzed in §5.

As mentioned in §1, an alternative approach for model selection is given by the popular least angle regression (LARS) of Efron, Hastie, Johnstone, and Tibshirani (2004), an extremely efficient algorithm for computing the entire LASSO (Tibshirani 1996) path. LASSO shrinks regression coefficients by imposing a penalty on their size. Making the penalty sufficiently large necessarily causes some of the coefficients
to be exactly zero, i.e. providing a form of model selection. Khan et al. (2007b) propose a robust alternative, namely RLARS, which replaces the non robust building blocks of LARS (mean, variance and correlation) by their robust counterparts. Both methods can deal with large \( p \) but provide only an order of entry into the model for the explanatory variables. Model selection based on minimizing \( K \)-fold CV mean squared prediction error for augmented models suggested by LARS is however readily available in the \texttt{lars} package for R.

Running RLARS (using the \texttt{brlars} R code made available on the authors’ website) on one simulated sample takes about 9.5 minutes, and this yields only the suggested order of entry of the variables (not where to stop). Our forward selection approach in conjunction with our robust estimator selects a model in less than 30 seconds when using weighted partial correlation to propose candidates (and less than 60 seconds when using \( z_\psi \)-ratios), see Table 2. Note that throughout the paper, our time comparisons are not intrinsic comparisons, but rather comparisons of available implementations of the methods only.

In \textit{http link}, the results of a simulation study with 10\% data contamination show that both robust approaches maintain some level of robustness but present signs of breakdown. Our very difficult data generating setting is not favourable to any model selection procedure and it is not reasonable to expect contaminated data in amounts exceeding 5-7\% in most real datasets. We are confident that our robust approaches can be used in many applications and this is illustrated with the real datasets in §5.

Finally, it should be noted that there exists alternative robust methods such as the robust boosting of Lutz et al. (2008) and the robust forward and stepwise selection approach based on robust correlation estimates, an approach similar to RLARS, of Khan, Van Aelst, and Zamar (2007a), but these methods are not compared to our methods since code is not available.
5 Examples

In this section, we analyze two real data sets. In each case, the explanatory variables have been centered and scaled to have mean equal to zero and standard deviation equal to one. For the census data, we also scale the dependent variable, the average price asked for a housing unit, which has a very large range and otherwise yields very large values for the estimated coefficients. Models are compared using the median absolute error (MAE), as measured by 10-fold CV. First, we find the proposed model for a given method using all the data. Then, we split the data into 10 roughly equal-sized parts. For the $k^{th}$ part, we fit the proposed model to the other nine parts of the data and calculate the MAE of the fitted model when predicting the $k^{th}$ part of the data. We do this for $k = 1, \ldots, 10$ and report the mean of the 10 MAE estimates.

In some cases (where computations remain reasonable) we compute a full median absolute prediction error (MAPE), as measured by 10-fold CV. That is, we split the data into 10 roughly equal-sized parts. For the $k^{th}$ part, we find the proposed model using the other nine parts of the data and calculate the MAPE of this chosen model when predicting the $k^{th}$ part of the data. We do this for $k = 1, \ldots, 10$ and report the mean of the 10 estimates of the MAPE. For all methods, the data were split in the same way. We also report the standard error (SE) of the MAE and MAPE estimates.

In both our examples, there is no a priori reason to believe that a linear regression model is the best approach. As usual, it remains the easiest route in the hope of obtaining some interpretability of the data. If simply prediction, and not interpretation, is sought then we can turn to random forests (Breiman 2001). We compare our predictions with those of random forests for completeness.

5.1 Protein data

We analyze data from a clinical investigation of patients with cardiovascular disease conducted by the Medizinische Klinik und Deutsches Herzzentrum der Technischen
Universität in Munich and the Center for Computational Diagnostics at Indiana University Purdue University. See Clough et al. (2009) and Patil et al. (2007) for details. Among the measures taken on the participants is the abundance of a series of proteins and an information of interest is the link between the abundance of each protein and a series of clinical variables. We consider here the abundance of the protein Fibrinogen beta chain that we seek to explain by means of 19 potential variables and their 171 interactions. Some regressors are binary however and over the \( n = 231 \) observations collected, some potential regressors yield a column of zeros, are exactly collinear with one column (or a linear combination of several columns), and must be removed. After removal of problematic interactions, we are left with 140 interactions and a total of \( p = 159 \) regressors. We carry out classical and robust forward selection as described in §3 using \( \alpha = 0.10 \). As \( n \) is only 231 we do not need to resort to subsampling. Also, as the chosen models remain small, we also consider further iterating our one-step weighted \( M \)-estimator to examine the impact of the bias. Results appear in Table 3.

First, we consider our two one-step estimates and the classical estimates. All three approaches pick up the importance of \( X_1X_3 \), but the robust approaches choose \( X_2X_11 \) over \( X_3X_11 \), the second variable in the classically chosen model. The correlation between \( X_2 \) and \( X_3 \) is only 0.05, however the correlation between \( X_2X_11 \) and \( X_3X_11 \) is 0.86, so these models differ only slightly, as is further confirmed by respective MAE values. Estimates for our two approaches using three-step estimates (results are the same for any greater number of steps) include more variables. Most notably, the \( z_\psi \)-ratio approach adds \( X_6X_11 \) and \( X_2X_19 \) to the \( X_1X_3 \) variable. This reduces the estimate of standard error (SE) and the MAE value. Note that while \( X_6X_11 \) and \( X_3X_11 \) are heavily correlated (0.93), the correlation between \( X_3X_11 \) and \( X_2X_19 \) is -0.05, and between \( X_1X_3 \) and \( X_2X_19 \) is -0.13, so \( X_2X_19 \) is adding new information to the model (useful information as the MAE is reduced to 0.52). The LS estimate of \( \sigma \) is inflated due to the presence of four outliers, patients 116, 206, 159 and 21,
which had the largest abundances of protein (three standard deviations larger than the mean abundance), see Figure 1.

Table 3: Estimates (SE) for models selected by forward selection with FDR stopping for the protein data. Candidates for entry selected to maximize robust partial correlation, robust \( z_\psi \)-ratio, and partial correlation, respectively. For robust estimators, one-step and three-(or more) steps weighted \( M \)-estimators are considered. \( X_1 \) is age, \( X_2 \) is gender, \( X_3 \) is nicotine, \( X_6 \) is cholesterol, \( X_{11} \) is an indicator of unstable angina, \( X_{19} \) is the glomerular filtration rate, an overall index of kidney function. The \( p \)-values are: \( ^* < 10^{-3}, ^{**} < 10^{-4}, ^{***} < 10^{-5} \). In robust \( z_\psi \)-ratio analysis with 3+ steps, 4 of 231 observations (patients 206\((w = 0)\), 21\((w = 0.10)\), 159\((w = 0.13)\) and 116\((w = 0.15)\)) had robust weights \( w \) less than 0.20. MAE and MAPE are measured following 10-fold CV as described in the text. Last row shows CPU time in seconds. MAPE for random forests is 0.64 (0.05).

<table>
<thead>
<tr>
<th></th>
<th>robust partial correlation</th>
<th>( z_\psi )-ratio</th>
<th>partial</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X_1X_3 )</td>
<td>0.22 (0.06)(^*)</td>
<td>0.26 (0.05)(***)</td>
<td>0.22 (0.06)(^*)</td>
</tr>
<tr>
<td>( X_2X_{11} )</td>
<td>0.25 (0.06)(**)</td>
<td>0.25 (0.06)(**)</td>
<td></td>
</tr>
<tr>
<td>( X_3X_{11} )</td>
<td></td>
<td></td>
<td>0.21 (0.06)(^*)</td>
</tr>
<tr>
<td>( X_6X_{11} )</td>
<td>0.23 (0.05)(**)</td>
<td>0.23 (0.05)(**)</td>
<td></td>
</tr>
<tr>
<td>( X_2X_{19} )</td>
<td></td>
<td>-0.17 (0.05)(^*)</td>
<td></td>
</tr>
<tr>
<td>( \hat{\sigma} )</td>
<td>0.83</td>
<td>0.80</td>
<td>0.83</td>
</tr>
<tr>
<td>MAE</td>
<td>0.61 (0.05)</td>
<td>0.58 (0.05)</td>
<td>0.61 (0.05)</td>
</tr>
<tr>
<td>MAPE</td>
<td>0.60 (0.06)</td>
<td>0.61 (0.05)</td>
<td>0.61 (0.04)</td>
</tr>
<tr>
<td>time(s)</td>
<td>6.4</td>
<td>6.4</td>
<td>10.6</td>
</tr>
</tbody>
</table>

The robust model chosen using robust partial correlation also outperforms the LS model, but as was our experience in the simulation study, it is more parsimonious than its \( z_\psi \)-ratio counterpart (which provides an additional predictor and better predictions). Note that with \( p = 159 \) regressors an all-possible-subsets approach, whether robust or non-robust, is not possible. We tried model selection following RLARS. The latter, as implemented in the \texttt{brlars} function provided by the authors, does not run to completion. RLARS proceeds by sampling with replacement and given the 14 binary covariates and their numerous interactions, subsamples lead to singularity in the
computation of the Huber correlations and the function terminates in error. We can carry out model selection by minimizing 10-fold CV mean squared prediction error for LARS using the function `cv.lars` in the R package `lars`. The selected model contains no regressors. It is hypothesized that the outliers and/or multicollinearity in the data baffle LARS, however only 5% of the \(159 \times 158/2 = 12561\) correlations are greater than 0.66 (and only 2% are greater than 0.90).

Figure 1: Three-step robust \(z_\psi\)-ratio and LS standardized residual analysis of models in Table 3 for the protein data.

Also note that with only 19 first-order variables, it is quite straightforward to obtain predictions using random forests as implemented in the `randomForest` function of the `randomForest` package for R. MAPE using the classical, both robust approaches, and random forests also appear in Table 3. While random forests outperform the classical approach, they do not do as well as either robust approach.
5.2 Census data

We analyze data constructed from the 1990 US Census. The original data were assembled by Data for Evaluating Learning in Valid Experiments (DELVE) at the University of Toronto and may be downloaded at http://www.cs.toronto.edu/~delve/data/census-house/desc.html. There are 22,784 observations and 139 variables in the original data. The goal is to explain the average price asked for a housing unit. Some of the other 138 variables are collinear, e.g. one variable is % female in sector and another is % male in sector, and some deletion is required to carry out the analysis. Also, there are many of the covariates which are all zeros, or almost all zeros, and a few that are categories essentially of the average price. After removing the latter and cleaning the covariates, we are left with \( n = 13,970 \) observations and 59 first-order covariates. Data are available at http link.

We include all possible interactions and carry out model selection. With \( n = 13,970 \) and \( p = 1770 \), the subsampling approach outlined in §3 and tested in §4 has to be used on most computers. As we had access to a high performance machine (2 CPU type AMD Opteron\textsuperscript{TM} Processor 250 running at 2390.073 Mhz under the Linux/x86\_64 – unknown – linux\_co4.2 – gnu operating system with 16 GB of RAM and 33 GB of swap memory) we can report the findings with and without subsampling. We report subsampling results for \((n_n,n_s) = (200,10)\) but results are qualitatively similar for many other schemes. The MAE of models selected following the different approaches are shown in Figure 2. Table 4 gives the CPU running times for each approach. Subsampling performs well and is clearly sufficient. Models chosen via the robust \( z_\psi \)-ratio approach outperform all the classical models.

Our robust estimates for a model consisting of the first 10 variables entered using a forward selection with robust \( z_\psi \)-ratio approach are listed in Table 5. A description of the regressors appears in Table 6. It is perhaps not surprising that a LS approach
Table 4: Number of variables entered in model [execution times in minutes] for forward selection with FDR stopping for census data. Computations are on high performance machine described in the text. Results for intermediate 10- and 20-variable models are included for purposes of comparison. Subsampling with \((n_n, n_s) = (200, 10)\). \(f\): final model selected; \(a\): selection not finished at time of submission.

<table>
<thead>
<tr>
<th>Robust part. corr.</th>
<th>Robust (z_\psi)-ratio</th>
<th>Classical part. corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>subsamp.</td>
<td>full sample</td>
</tr>
<tr>
<td>(71f) [1716]</td>
<td>102(^a) [30022]</td>
<td>68(^f) [110]</td>
</tr>
</tbody>
</table>

performs poorly here as 986 observations (7.1%) get downweighted below 0.01 by our robust estimator. We note that 10% of the observations are assigned weights less than 0.20. This amount of hard downweighting is more than that in our simulated data with 5% outliers in §4, but less than that in our simulated data with 10% outliers, see http link.

Table 5: Estimates and tests for robust 10-variable model selected by forward selection with FDR stopping using robust \(z_\psi\)-ratios based on one-step weighted \(M\)-estimators and subsampling \((n_n = 200, n_s = 10)\) to predict average asking price for a housing unit in a sector for census data.

<table>
<thead>
<tr>
<th>(z_\psi)-ratio</th>
<th>Estimate (SE)</th>
<th>(p)-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>2.74 (0.03)</td>
<td>0</td>
</tr>
<tr>
<td>H9.4*H5.4</td>
<td>0.04 (0.03)</td>
<td>0.20</td>
</tr>
<tr>
<td>H9.9*P6.4</td>
<td>-0.6 (0.3)</td>
<td>0.02</td>
</tr>
<tr>
<td>H15.1*P11.3</td>
<td>0.260 (0.004)</td>
<td>0</td>
</tr>
<tr>
<td>H9.4*P6.4</td>
<td>5.9 (0.5)</td>
<td>&lt; 10(^{-30})</td>
</tr>
<tr>
<td>H9.9*H9.4</td>
<td>20.5 (0.8)</td>
<td>&lt; 10(^{-146})</td>
</tr>
<tr>
<td>P18.1*P16.1</td>
<td>-0.138 (0.004)</td>
<td>&lt; 10(^{-302})</td>
</tr>
<tr>
<td>H15.2*P14.2</td>
<td>-0.146 (0.004)</td>
<td>0</td>
</tr>
<tr>
<td>H7.1*H5.2</td>
<td>0.052 (0.003)</td>
<td>&lt; 10(^{-53})</td>
</tr>
<tr>
<td>H9.9*P18.2</td>
<td>-0.31 (0.02)</td>
<td>&lt; 10(^{-73})</td>
</tr>
<tr>
<td>H5.4*P25.1</td>
<td>0.070 (0.004)</td>
<td>&lt; 10(^{-86})</td>
</tr>
<tr>
<td>(\hat{\sigma})</td>
<td>0.311</td>
<td></td>
</tr>
</tbody>
</table>
Figure 2: MAE (±2SE) of different models for census data as evaluated by 10-fold CV. Points labels: base indicates the number of variables in the model, exponent identifies the selection method following the notation in the legend.

When using the subsampling approach, only two regressors appear both in our robust 10-variable model and the final 68-variable model chosen by LS: H15.1*P11.3 and P9.9*P18.2. The 68-variable LS model does contain three first-order regressors: H15.1, the average number of rooms in a owner-occupied housing unit (HU); P18.3, the percentage of households (HH) with no persons under 18 which are family HH; and H14, the average number of rooms in a HU. The LS estimate of $\sigma$ is 0.523.

The cv.lars procedure requires 11 hours to run on the same machine and once again selects no regressors even though LS forward selection leads to a 68-variable model which reduces MAE by 59% with respect to a mean-only model. Note that a model consisting of the first 10 variables selected by the forward selection already reduce the MAE by 49%. It is again hypothesized that the outliers and/or the near multicollinearity in the data baffle LARS.

With these data we do not seek a ‘true’ model, but rather to best explain the average asking price in a sector with very few regressors. As there is near multicollinearity, many models should perform equivalently. What is most impressive is that we find a 10-variable robust model that outperforms the selected 68-variable LS
model with respect to MAE, and we do so in less time than the classical analysis, 85 minutes versus 110 minutes, respectively.

Table 6: Description of regressors selected in Table 5 to predict average asking price for a HU in a sector for census data.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>P6.4</td>
<td>%-tage Asian or Pacific Inslander (asian)</td>
</tr>
<tr>
<td>P11.3</td>
<td>%-tage [25:64] years old</td>
</tr>
<tr>
<td>P14.2</td>
<td>%-tage males married, not separated</td>
</tr>
<tr>
<td>P16.1</td>
<td>%-tage of HH-lds with 1 person</td>
</tr>
<tr>
<td>P18.1</td>
<td>%-tage of HH-lds with 1+ persons under 18 which are family HH-lds</td>
</tr>
<tr>
<td>P18.2</td>
<td>%-tage of HH-lds with 1+ persons under 18 which are non-family HH-lds</td>
</tr>
<tr>
<td>P25.1</td>
<td>%-tage of HH-lds with 1+ persons 65 and over</td>
</tr>
<tr>
<td>H5.2</td>
<td>%-tage of vacant HU for sale only</td>
</tr>
<tr>
<td>H5.4</td>
<td>%-tage of vacant HU for seasonal, recreational or occasional use</td>
</tr>
<tr>
<td>H7.1</td>
<td>%-tage of vacant HU with usual home elsewhere</td>
</tr>
<tr>
<td>H9.4</td>
<td>%-tage of ownOcc HU with asian HH-lder</td>
</tr>
<tr>
<td>H9.9</td>
<td>%-tage of rentOcc HU with asian HH-lder</td>
</tr>
<tr>
<td>H15.1</td>
<td>Average number of rooms in a ownOcc HU</td>
</tr>
<tr>
<td>H15.2</td>
<td>Average number of rooms in a rentOcc HU</td>
</tr>
</tbody>
</table>

Finally, we put these data to random forests and get a MAPE of 0.197 (0.003). This is clearly better than the prediction power of the full classical model found with subsampling, 0.225 (0.002), but only the same as that of the 30-variable (i.e. we follow forward selection but only keep the first 30 variables even if the entry criterion would allow more to enter) robust model found with the $z_\psi$-ratio and subsampling: 0.196 (0.002). The robust 10-variable and 20-variable models had MAPE of 0.213 (0.003) and 0.202 (0.003), respectively.

6 Concluding remarks

Although outliers are prevalent in very large data sets, existing robust methods for model fitting and selection are not computationally feasible in high dimensions. We present a simplified robust estimator that is fast and effective, allowing for much needed robust inference in high dimensions.
Note that all our computations were done in R and the code for our proposed approach is not optimized for speed, but rather for ease of use. Computing times reported should thus not be examined in the absolute, but rather on a comparative basis for the methods considered. Increased computational efficiency is a topic for future research.

Also note that we chose to scale predictors using (non-robust) mean and standard deviation. More robust location and scale estimators can be used, but we found that use of the median and the MAD led to a poorer performance of the robust partial correlation approach while causing no change in the performance of the robust $z_\psi$-ratio approach.

Finally, note that we used an $MM$-estimator for the marginal regression models (with one explanatory variable) to get the starting weights, but one could also consider different and possibly simpler robust estimators. A redescending $\psi$-function is recommended in high dimensions, but for the marginal models, a non redescending $\psi$-function, such as Huber’s (Huber 1964) $\psi$-function could in principle be used. In that case, the bias of the one-step $M$-estimator would be different.

**Supplemental Materials**

**Technical Details:** More detailed presentation of technical results and additional simulation results.

**References**


