Robust groupwise least angle regression

Andreas Alfons[a,b,*], Christophe Croux[b], Sarah Gelper[c]

[a]Erasmus School of Economics, Erasmus Universiteit Rotterdam, PO Box 1738, 3000DR Rotterdam, The Netherlands
[b]ORSTAT Research Center, Faculty of Economics and Business, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium
[c]Rotterdam School of Management, Erasmus Universiteit Rotterdam, PO Box 1738, 3000DR Rotterdam, The Netherlands

Abstract

Many regression problems exhibit a natural grouping among predictor variables. Examples are groups of dummy variables representing categorical variables, or present and lagged values of time series data. Since model selection in such cases typically aims for selecting groups of variables rather than individual covariates, an extension of the popular least angle regression (LARS) procedure to groupwise variable selection is considered. Data sets occurring in applied statistics frequently contain outliers that do not follow the model or the majority of the data. Therefore a modification of the groupwise LARS algorithm is introduced that reduces the influence of outlying data points. Simulation studies and a real data example demonstrate the excellent performance of groupwise LARS and, when outliers are present, its robustification.

Keywords: categorical variables, model selection, outliers, time series

1. Introduction

In many applications of linear regression, there exists a natural grouping among the predictor variables. One common example is regression with categorical variables, where each categorical variable is represented by a group of dummy variables. Another example is regression with time series data, where typically not only the original series are considered in the model, but also several lags of each series. Furthermore, time series models frequently contain an autoregressive part, i.e., lags of the response are included as covariates. Such models are commonly referred to as autoregressive models with exogenous inputs, or ARX models for short. Note that in both situations, groups of covariates emerge from the measured variables.

With increasing availability of data sets containing a large number of variables, model selection continues to be a topic of high importance in regression analysis. Linear models that include a large set of variables tend towards having large variance, often resulting in poor prediction performance. Selecting only the important variables can therefore improve prediction accuracy. Furthermore, traditional regression methods cannot be applied if the number of variables is larger than the number of observations due to the rank deficiency of the design matrix.

Whenever the regression problem involves groups of covariates, variable selection methods should select these groups rather than individual covariates. This ensures that all information of a selected measured variable enters the model, which is in general not the case when selecting individual covariates (Yuan and Lin, 2006). In addition, retaining the groupwise structure in submodels allows for better interpretation of the results.

Concerning notation, let n denote the number of observations and p the total number of covariates from m predictor groups. Moreover, let $\boldsymbol{y} = (y_1, \dots, y_n)'$ be the response variable, and $\boldsymbol{X}_j$ an $(n \times p_j)$ matrix corresponding to the j-th predictor group, $j = 1, \dots, m$, with $\sum_{j=1}^{m} p_j = p$. The regression problem with grouped predictor variables can then be written as

$$\boldsymbol{y} = \sum_{j=1}^{m} \boldsymbol{X}_j \boldsymbol{\beta}_j + \boldsymbol{\varepsilon}, \tag{1}$$

[*] Corresponding author

Email addresses: alfons@ese.eur.nl (Andreas Alfons), christophe.croux@kuleuven.be (Christophe Croux), sgelper@rsm.nl (Sarah Gelper)

Preprint submitted to Computational Statistics & Data Analysis *February 9, 2015*

where $\boldsymbol{\beta}_j$ is a coefficient vector of size p_j, $j = 1, \dots, m$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)'$ are the random error terms. Our aim is to find a subset $J \subseteq \{1, \dots, m\}$ of the important predictor groups such that only those are included in the regression model, which is equivalent to setting the coefficient vectors $\boldsymbol{\beta}_j$ with $j \notin J$ in (1) to zero vectors.

In the traditional variable selection setting with all $p_j = 1$, a considerable amount of research has been done. Popular methods are the least absolute shrinkage and selection operator (lasso; Tibshirani, 1996), least angle regression (LARS; Efron et al., 2004), and the nonnegative garrote (Breiman, 1995). All three methods have been adjusted by Yuan and Lin (2006) to handle grouped variables. Zhao et al. (2009) introduced a family of composite absolute penalty functions for grouped and hierarchical variable selection via penalized regression. Furthermore, a groupwise version of the lasso for logistic regression was developed by Meier et al. (2008). Breheny and Huang (2009) follow a different philosophy and introduced a penalized regression framework for bi-level variable selection with grouped variables, i.e., their method first selects the important groups of variables and then the important variables within those groups. Nevertheless, none of these contributions consider the problem of outlying data points. While many robust methods for model selection in the traditional setting are available (e.g. Ronchetti et al., 1997; Khan et al., 2007; McCann and Welsch, 2007; Salibian-Barrera and Van Aelst, 2008; Khan et al., 2010; Alfons et al., 2011, 2013), almost no work has been done on robust groupwise variable selection. Chen et al. (2010) apply a more robust version of the groupwise lasso based on a convex combination of L_1 and L_2 loss functions. However, their procedure is only robust against heavily-tailed errors, but not against leverage points, i.e., outliers in the predictor space.

This paper focuses on the LARS procedure, which produces a sequence of variables in the order of their predictive content. Khan et al. (2007) point out that only correlations are required for variable sequencing with LARS and propose robustified versions of LARS, referred to as RLARS. While those authors express LARS in terms of correlations, we propose to use an extension of LARS to grouped variables that is formulated in terms of R^2 measures from short regressions. Here the term short regressions refers to regressions that use only one of the predictor groups. As the groupwise LARS approach is sensitive to outliers, we propose a robustification of the procedure such that the influence of outliers is reduced. We focus on sequencing the groups of variables, i.e., obtaining a sequence of groups in the order of their importance that can be further investigated for model selection and estimation.

The rest of the paper is organized as follows. Groupwise LARS is discussed in Section 2. Its robustification is then introduced in Section 3. Simulation studies are performed in Section 4, and Section 5 contains a real data example. Finally, Section 6 concludes. Proofs and technical details on the algorithms can be found in the appendix.

2. Groupwise least angle regression

First, we review the idea of least angle regression (LARS) in the traditional setting with non-grouped variables. It proceeds in the following stepwise fashion (for details on the LARS algorithm, see Efron et al., 2004):

First step. Find the predictor with the highest correlation to the response and add it to the set of active predictors.

$(k + 1)$*-th step.* Move along the equiangular direction among all active predictors until a new predictor has equal correlation with the current residual, and add that predictor to the active set. The key to the algorithm is that this step size can easily be computed.

We generalize LARS by reformulating it in terms of R^2 measures from short regressions. A short regression has only the variables belonging to one single group as covariates, hence it has a limited number of regressors. This is in contrast to the full regression, where all covariates are included. If p is large with respect to n, short regressions can be carried out, while the full regression may not. Our approach is similar to the groupwise LARS algorithm of Yuan and Lin (2006), but our algorithm allows for more groups to be sequenced. The key steps of the groupwise LARS algorithm are discussed in the following. A complete schematic overview of the algorithm including technical details is given in the appendix. Let z_0 be the standardized response and $\underline{\boldsymbol{X}}_j$, $j = 1, \dots, m$, the groups of standardized covariates such that all variables have zero mean and unit variance. Furthermore, let $R^2(\boldsymbol{z} \sim \boldsymbol{X})$ denote the R^2 measure of least squares regression of the vector $\boldsymbol{z}$ on the variables given by the columns of the matrix $\boldsymbol{X}$, let A denote the active set, i.e., the index set of the sequenced predictor groups, and let the complement A^c denote the inactive set, i.e., the index set of the not yet sequenced predictor groups.

First step. Find the predictor group with the largest R^2 measure

$$R^2(z_0 \sim \underline{\boldsymbol{X}}_j), \qquad j = 1, \ldots, m, \tag{2}$$

and add its index to the active set A.

$(k+1)$-th step. At the beginning of the $(k+1)$-th step, $k \geq 1$, the active set A contains the indices of the k first ranked predictor groups $\underline{\boldsymbol{X}}_{(1)}, \ldots, \underline{\boldsymbol{X}}_{(k)}$. Let the current response be denoted by z_{k-1}, and let $\tilde{\boldsymbol{x}}_{(l)}$ be the standardized vector of fitted values of regressing z_{l-1} on $\underline{\boldsymbol{X}}_{(l)}$, $l = 1, \ldots, k$. The idea of the $(k+1)$-th step is to move along the equiangular direction among the predictor groups until a new group with equal R^2 enters the model. The equiangular vector $\boldsymbol{u}_k$ is defined as the standardized vector having equal correlation with all of the vectors $\tilde{\boldsymbol{x}}_{(1)}, \ldots, \tilde{\boldsymbol{x}}_{(k)}$ and is easy to compute (see, e.g., Khan et al., 2007). Let this correlation be denoted by

$$a_k = \operatorname{cor}(\boldsymbol{u}_k, \tilde{\boldsymbol{x}}_{(1)}) = \ldots = \operatorname{cor}(\boldsymbol{u}_k, \tilde{\boldsymbol{x}}_{(k)}). \tag{3}$$

Then the response is updated by moving along the direction of the equiangular vector:

$$z_k = \frac{z_{k-1} - \gamma_k \boldsymbol{u}_k}{\sigma_k} \qquad \text{with} \qquad \sigma_k^2 = \operatorname{var}(z_{k-1} - \gamma_k \boldsymbol{u}_k). \tag{4}$$

Note that we scale the response in each step to unit variance to simplify the calculations. The shrinking factor γ_k is thereby chosen as the smallest positive solution such that it holds for a predictor group $\underline{\boldsymbol{X}}_j$ with $j \in A^c$ that

$$R^2(z_{k-1} - \gamma_k \boldsymbol{u}_k \sim \tilde{\boldsymbol{x}}_{(k)}) = R^2(z_{k-1} - \gamma_k \boldsymbol{u}_k \sim \underline{\boldsymbol{X}}_j). \tag{5}$$

Condition (5) is a generalization of the equi-correlation condition of the LARS procedure developed by Efron et al. (2004). In standard LARS, individual variables are added one by one. In this case, the R^2 in (5) reduces to a squared correlation between the response and the corresponding covariate, and the resulting equation is trivial to solve. In groupwise LARS, a whole group of variables is added in each step, which makes solving (5) more complex. In the appendix, we prove the following lemma.

Lemma 1. *In the $(k+1)$-th step of the groupwise LARS algorithm, for every $k \geq 1$, the following statements hold.*

(a) The current response z_{k-1} has equal and positive correlation with all $\tilde{\boldsymbol{x}}_{(1)}, \ldots, \tilde{\boldsymbol{x}}_{(k)}$:

$$r_k = cor(z_{k-1}, \tilde{\boldsymbol{x}}_{(1)}) = \ldots = cor(z_{k-1}, \tilde{\boldsymbol{x}}_{(k)}) \geq 0. \tag{6}$$

(b) For every $j \in A^c$, it holds that $R^2(z_{k-1} \sim \underline{\boldsymbol{X}}_j) \leq r_k^2$.

(c) For every $j \in A^c$, the solution γ_k to (5) verifies $0 \leq \gamma_k \leq r_k/a_k$.

From the above lemma, equation (6), it follows that $\tilde{\boldsymbol{x}}_{(k)}$ on the left hand side of (5) is an arbitrary choice and can be replaced by any other $\tilde{\boldsymbol{x}}_{(l)}$, $l = 1, \ldots, k$. For a convenient way of solving the generalized equi-correlation condition (5), let $\hat{\boldsymbol{z}}_{k-1}^j$ and $\hat{\boldsymbol{u}}_k^j$ denote the fitted values of least squares regression of z_{k-1} and $\boldsymbol{u}_k$, respectively, on $\underline{\boldsymbol{X}}_j$ for $j \in A^c$. In the appendix, we show that solving equation (5) is equivalent to solving the following quadratic equation in γ_k:

$$(r_k^2 - r_{k,j}^2) + 2(r_{k,j}a_{k,j} - r_k a_k)\gamma_k + (a_k^2 - \tau_{k,j}^2)\gamma_k^2 = 0, \tag{7}$$

where

$$r_{k,j} = \operatorname{cor}(z_{k-1}, \hat{\boldsymbol{z}}_{k-1}^j), \tag{8}$$

$$a_{k,j} = \operatorname{cor}(\boldsymbol{u}_k, \hat{\boldsymbol{z}}_{k-1}^j), \tag{9}$$

$$\tau_{k,j} = \operatorname{cor}(\boldsymbol{u}_k, \hat{\boldsymbol{u}}_k^j). \tag{10}$$

We know from Lemma 1 that the equivalent conditions (5) and (7) always have a positive solution for all $j \in A^c$. The index corresponding to the smallest positive solution over all $j \in A^c$ is added to the active set. With step size γ_k now determined, the response can be updated as in (4) to conclude the $(k+1)$-th step.

A potential problem with the procedure outlined above is that large predictor groups consume more degrees of freedom. They therefore tend to have larger R^2 and consequently tend to be selected by the generalized equiangular condition (5). Inspired by Yuan and Lin (2006), groupwise LARS can easily be adjusted for unequal group sizes p_j. The first selected predictor group is obtained by maximizing

$$R^2(z_0 \sim \underline{\boldsymbol{X}}_j)/p_j, \qquad j = 1, \dots, m. \tag{11}$$

and the equiangular condition is adjusted to

$$R^2(z_{k-1} - \gamma_k \boldsymbol{u}_k \sim \tilde{\boldsymbol{x}}_{(k)})/p_{(k)} = R^2(z_{k-1} - \gamma_k \boldsymbol{u}_k \sim \underline{\boldsymbol{X}}_j)/p_j. \tag{12}$$

Other scalings are possible as well, but the proposed scaling is chosen for reasons of interpretation. With this adjustment, the procedure does not consider the total explained variance that a predictor group adds to the model, but the explained variance *per covariate* in the group. Hence it avoids favoring groups with a large number of variables. For the quadratic equation (7), the quantities r_k, a_k, $r_{k,j}$, $a_{k,j}$ and $\tau_{k,j}$ change to

$$r_k = \mathrm{cor}(z_{k-1}, \tilde{\boldsymbol{x}}_{(1)})/\sqrt{p_{(1)}} = \dots = \mathrm{cor}(z_{k-1}, \tilde{\boldsymbol{x}}_{(k)})/\sqrt{p_{(k)}} \geq 0, \tag{13}$$

$$a_k = \mathrm{cor}(\boldsymbol{u}_k, \tilde{\boldsymbol{x}}_{(1)})/\sqrt{p_{(1)}} = \dots = \mathrm{cor}(\boldsymbol{u}_k, \tilde{\boldsymbol{x}}_{(k)})/\sqrt{p_{(k)}}, \tag{14}$$

$$r_{k,j} = \mathrm{cor}(z_{k-1}, \hat{\boldsymbol{z}}_{k-1}^j)/\sqrt{p_j}, \tag{15}$$

$$a_{k,j} = \mathrm{cor}(\boldsymbol{u}_k, \hat{\boldsymbol{z}}_{k-1}^j)/\sqrt{p_j}, \tag{16}$$

$$\tau_{k,j} = \mathrm{cor}(\boldsymbol{u}_k, \hat{\boldsymbol{u}}_k^j)/\sqrt{p_j}. \tag{17}$$

With these adjustments, Lemma 1 and equation (7) still hold. The proofs follow exactly the same lines as the ones given in the appendix for the unadjusted case. It is important to note that (14) implies a definition of the equiangular vector $\boldsymbol{u}_k$ such that the correlations are adjusted for the respective group sizes. The computation of $\boldsymbol{u}_k$ with this adjustment is described in the appendix.

Our procedure differs from the method proposed by Yuan and Lin (2006) in the construction of the equiangular vector. We only use the standardized fitted values from short regressions once a predictor group enters the active set. Yuan and Lin (2006), on the other hand, compute the equiangular vector between all covariates of the active groups. The latter approach has the advantage that coefficient estimates can be accurately computed along the coefficient path, leading to the least squares solution after the final step (if $n > p$). The advantage of our approach is that it can sequence as many groups as there are observations, while the approach of Yuan and Lin (2006) is limited to as many individual covariates as there are observations. Since our approach is focused on *sequencing* the candidate predictor groups, model selection is achieved in a two-step procedure. First, a sequence of the predictor groups is obtained. Second, submodels along the sequence are investigated. By adding the grouped variables one by one to a series of least squares regression models, the final model can be determined via an optimality criterion such as the BIC (Schwarz, 1978), or by estimating prediction performance via cross-validation. Such a two-step approach is also proposed by Khan et al. (2007) in their robustification of the traditional LARS algorithm.

3. Robust groupwise least angle regression

Robust groupwise LARS combines ideas of the groupwise LARS algorithm from the previous section and the robust LARS procedure by Khan et al. (2007) for the traditional variable selection setting. Khan et al. (2007) propose two variants of robust LARS: *plug-in* robust LARS and *data cleaning* robust LARS. In the plug-in approach, the classical building blocks of the algorithm (means, standard deviations and correlations) are replaced by robust counterparts. The data cleaning approach, on the other hand, uses multivariate winsorization to clean the data first, after which classical LARS is applied to the cleaned data. Plug-in robust LARS has the advantage that it can be applied to high-dimensional data and that computation time scales with the number of sequenced variables. Data cleaning LARS requires more observations than variables and the data cleaning step dominates computation time. Nevertheless, the numerical experiments of Khan et al. (2007) suggest that the data cleaning approach in some situations performs better than the plug-in approach.

Unfortunately, such a plug-in approach leads to severe computational problems in the case of grouped variables. Groupwise LARS relies heavily on the properties of the classical estimators to ensure that the quadratic equation (7) always has a positive solution. When least squares regression and correlations are replaced by robust counterparts, this cannot be guaranteed anymore. In the traditional case of selecting individual variables, only a linear equation needs to be considered. Even though the theoretical properties leading to that equation do not hold anymore if the classical estimators are replaced by robust counterparts, there is still always a solution. Hence plug-in robust LARS in the traditional setting does not suffer from the same problem as in the groupwise setting.

As a consequence, we focus on a data cleaning approach for robust groupwise LARS, by applying the algorithm from Section 2 to a cleaned data set. While Khan et al. (2007) use multivariate winsorization to clean the data, this is not practical for grouped variables. For instance, winsorization is problematic in the case of categorical data as it is based on the assumption of a normal distribution.

To motivate our data cleaning method, keep in mind that the algorithm is formulated in terms of short regressions of the current response z_{k-1} and the equiangular vector $\boldsymbol{u}_k$ on predictor groups $\underline{\boldsymbol{X}}_j$. In addition, the equiangular vector $\boldsymbol{u}_k$ is a linear combination of the vectors $\tilde{\boldsymbol{x}}_{(1)}, \dots, \tilde{\boldsymbol{x}}_{(k)}$, which are standardized fitted values from short regressions in the steps carried out so far. Suppose that the data are cleaned such that the fitted values $\hat{z}_0^j$ do not deviate too much from the majority of the data for all $j = 1, \dots, m$. Then due to the definition of z_k the elements of the vectors $\hat{z}_k^l$ for $k > 0$ behave like the majority of the data as well, and the influence of the outliers on the result remains *by construction* limited in the further steps of the algorithm.

The aim of our data cleaning approach is to find a set of weights such that multiplying each observation with the corresponding weight results in a data set that behaves like the majority of the data. We can leverage the fact that certain robust regression methods can be interpreted as weighted least squares methods with data-dependent weights. In the case of robust groupwise LARS, let $z = (z_1, \dots, z_n)'$ denote the robustly standardized response (e.g., using median and median absolute deviation), and let $\underline{\boldsymbol{X}}_j$ denote the groups of robustly standardized predictor variables for $j = 1, \dots, m$. In addition, let $\underline{\mathbf{x}}_{1j}, \dots, \underline{\mathbf{x}}_{nj}$ denote the p_j-dimensional observations of $\underline{\boldsymbol{X}}_j$ (i.e., the rows). For instance, MM-regression (Yohai, 1987; Maronna et al., 2006) of z on $\underline{\boldsymbol{X}}_j$ then solves the problem

$$\hat{\boldsymbol{\beta}}_j = \arg\min_{\boldsymbol{\beta}_j} \sum_{i=1}^{n} \rho\left(\frac{r_{ij}(\boldsymbol{\beta}_j)}{\hat{\sigma}_j}\right), \tag{18}$$

where $r_{ij}(\boldsymbol{\beta}_j) = z_i - \underline{\mathbf{x}}'_{ij}\boldsymbol{\beta}_j$ denotes the residuals, $\rho(.)$ is a bounded loss function, and $\hat{\sigma}_j$ is a robust scale estimate of the residuals from a robust but inefficient S-estimator. If ρ is differentiable, its derivative $\psi = \rho'$ can be used to define weights

$$\omega_{ij} = \frac{\psi\left(r_{ij}(\boldsymbol{\beta}_j)/\hat{\sigma}_j\right)}{r_{ij}(\boldsymbol{\beta}_j)/\hat{\sigma}_j}, \qquad i = 1, \dots, n, \tag{19}$$

such that weighted least squares regression of z on $\underline{\boldsymbol{X}}_j$ with weights $\boldsymbol{\omega}_j = (\omega_{1j}, \dots, \omega_{nj})'$ yields the same results. Equivalently, least squares regression with response $z_j^* = (\sqrt{\omega_{1j}}z_1, \dots, \sqrt{\omega_{nj}}z_n)'$ and predictors $\underline{\boldsymbol{X}}_j^* = (\sqrt{\omega_{1j}}\underline{\mathbf{x}}_1, \dots, \sqrt{\omega_{nj}}\underline{\mathbf{x}}_n)'$ can be performed.

Thus data cleaning weights $\sqrt{\omega_{1j}}, \dots, \sqrt{\omega_{nj}}$ can be obtained by such robust regressions for each predictor group $\underline{\boldsymbol{X}}_j$. However, this would result in a different cleaned response z_j^* for each cleaned predictor group $\underline{\boldsymbol{X}}_j^*$. As a remedy, a set of weights $\boldsymbol{\omega} = (\omega_1, \dots, \omega_n)'$ for the whole data set can be defined by

$$\omega_i = \min_{j \in \{1,\dots,m\}} \sqrt{\omega_{ij}}, \qquad i = 1, \dots, n. \tag{20}$$

The algorithm for groupwise LARS from Section 2 can then be applied to the cleaned data $\boldsymbol{y}^* = (\omega_1 z_1, \dots, \omega_n z_n)'$ and $\boldsymbol{X}_j^* = (\omega_1 \underline{\mathbf{x}}_{1j}, \dots, \omega_n \underline{\mathbf{x}}_{nj})'$, $j = 1, \dots, m$. The data are cleaned such that the fitted values from the initial short regressions behave like the majority of the data, avoiding a strong influence of outliers by construction. Furthermore, since the weights are based on robust short regressions for each predictor group, the procedure can also be applied to high-dimensional data.

Another strategy for cleaning the data is to obtain weights based on Mahalanobis distances of the multivariate set of standardized residuals from the robust short regressions. Let $\hat{\boldsymbol{V}}$ denote a robust correlation matrix of the matrix of standardized residuals $\boldsymbol{S} = \left(s_{ij}\right)_{1 \leq i \leq n, 1 \leq j \leq m}$ with $s_{ij} = r_{ij}(\boldsymbol{\beta}_j)/\hat{\sigma}_j$. Then a set of weights $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)'$ can be obtained by

$$\omega_i = \min\left(\sqrt{\frac{c}{\mathbf{s}_i' \hat{\boldsymbol{V}}^{-1} \mathbf{s}_i}}, 1\right), \qquad i = 1, \ldots, n, \tag{21}$$

where $\mathbf{s}_i = (s_{i1}, \ldots, s_{im})'$. This weighting scheme generalizes the data cleaning robust LARS proposed by Khan et al. (2007) in the traditional variable selection setting. Concerning the correlation matrix $\hat{\boldsymbol{V}}$ and the tuning constant c, the proposals of Khan et al. (2007) are reasonable choices. That is, the elements of $\hat{\boldsymbol{V}}$ are computed pairwise via bivariate winsorization, after which positive-definiteness is restored if necessary, and c is set to the 95% quantile of the χ_m^2 distribution with m degrees of freedom. However, a drawback of this method is that the number of observations must be larger than the number of predictor groups m, otherwise the robust correlation matrix $\hat{\boldsymbol{V}}$ cannot be inverted.

This restriction can be circumvented by using the Euclidean distance instead of the Mahalanobis distance, i.e., by disregarding the correlations between the standardized residuals. Literature on other applications of high-dimensional data analysis shows that imposing a diagonal covariance or correlation matrix can lead to good results (e.g. Dudoit et al., 2001; Tibshirani et al., 2003; Witten et al., 2009). In this case, the data cleaning weights $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)'$ are defined as

$$\omega_i = \min\left(\sqrt{\frac{c}{\mathbf{s}_i' \mathbf{s}_i}}, 1\right), \qquad i = 1, \ldots, n. \tag{22}$$

The tuning constant c can be chosen in the same way as the 95% quantile of the χ_m^2 distribution.

In the remainder of the paper, the abbreviation *GrpLARS* is sometimes used for groupwise LARS. The three proposed methods for robust groupwise LARS are referred to as *RGrpLARS-min* (for taking the minimum weight from robust short regressions), *RGrpLARS-ED* (for weights based on Euclidean distances) and *RGrpLARS-MD* (for weights based on Mahalanobis distances).

4. Simulation studies

In this section, we assess the performance of our groupwise LARS procedures by means of simulation. The final models are obtained by fitting least squares and MM-regression along the respective sequence of predictor groups, and choosing the respective optimal model via BIC. Let A_k denote the set of the first k sequenced groups. Then the BIC for the MM-regression model using the subset of predictor groups given by A_k can be written as

$$\text{BIC}(A_k) = \log \hat{\sigma}_{A_k} + \text{df}(A_k) \frac{\log(n)}{n}, \tag{23}$$

where $\hat{\sigma}_{A_k}$ denotes the corresponding residual scale estimate from the initial S-estimator, and $\text{df}(A_k)$ are the degrees of freedom of the model (i.e., the number of estimated coefficients). Furthermore, the loss function ρ for MM-regression is chosen to be Tukey's bisquare function tuned towards 95% efficiency.

We evaluate the performance of robust and classical groupwise LARS by comparing them to plug-in RLARS and data cleaning RLARS (Khan et al., 2007), as well as classical LARS (Efron et al., 2004). As for robust groupwise LARS, the final RLARS models are obtained by fitting MM-regression models along the respective sequences and computing the BIC based on the corresponding residual scale estimate, except that variables are of course added individually to the model. Classical LARS, on the other hand, computes the coefficient path leading to the least squares solution of the full model. Submodels along each step of the coefficient path are evaluated via BIC. Note that the degrees of freedom of a model in the LARS path are well approximated by the number of non zero coefficients (see Efron et al., 2004).

The simulations are performed in R (R Development Core Team, 2014) with package **simFrame** (Alfons et al., 2010; Alfons, 2014b), which is a general framework for simulation studies in statistics. Moreover, robust and classical groupwise LARS, as well as robust LARS, are implemented in the R package **robustHD** (Alfons, 2014a), while classical LARS is available in package **lars** (Hastie and Efron, 2013).

4.1. Data configurations

Two data configurations are investigated in the simulation study. The first configuration corresponds to a model with numerical and categorical predictors, the second to a time series model. For each data configuration, we perform $R = 1000$ simulation runs.

4.1.1. Numerical and categorical data

First we generate latent variables $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_{20}$ and $\boldsymbol{w}$ independently from a standard normal distribution $N(0, 1)$. Then we define predictors $\boldsymbol{x}_j = (\boldsymbol{f}_j + \boldsymbol{w})/\sqrt{2}$, $j = 1, \ldots, 20$. Finally, the last 10 variables $\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{20}$ are trichotomized as -1, 1 or 0 if they are smaller than $\Phi^{-1}(\frac{1}{3})$, larger than $\Phi^{-1}(\frac{2}{3})$ or in between. There are 10 groups containing the linear and quadratic terms of the numerical predictors, as well as 10 groups of two dummy variables representing the categorical predictors, resulting in a total of 40 individual candidate predictor variables. We generate the response $\boldsymbol{y}$ from the model

$$\boldsymbol{y} = \boldsymbol{x}_3 + \boldsymbol{x}_3^2 + \frac{2}{3}\boldsymbol{x}_6 - \boldsymbol{x}_6^2 + 2I(\boldsymbol{x}_{11} = -1) + I(\boldsymbol{x}_{11} = 1) + \sigma\boldsymbol{\varepsilon}, \tag{24}$$

where $I(.)$ is the indicator function, $\boldsymbol{\varepsilon} \sim N(0, 1)$ and $\sigma = 2$. The number of relevant groups of variables is thus equal to three. The number of observations is set to $n = 100$. This data configuration is similar to model IV from Yuan and Lin (2006).

We consider the following contamination scenarios:

1. *No contamination.*
2. *Vertical outliers*: 10% of the error terms $\boldsymbol{\varepsilon}$ are generated from a normal distribution $N(20, 1)$ instead of $N(0, 1)$.
3. *Independent normal leverage points*: 10% of the observations are generated as follows. Outliers in the numerical predictors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{10}$ are drawn from independent normal distributions $N(\mu, \sigma^2)$ with $\mu = 3$ and $\sigma = 0.01$. Denoting such a leverage point by $\tilde{\mathbf{x}}_i = (\tilde{x}_{i,1}, \ldots, \tilde{x}_{i,10}, x_{i,11}, \ldots, x_{i,20})'$, we generate the value of the response as

$$\tilde{y}_i = -0.5\left[\textstyle\sum_{j=1}^{10}(\tilde{x}_{ij} + \tilde{x}_{ij}^2) + \sum_{j=11}^{20}(I(x_{ij} = -1) + I(x_{ij} = 1))\right]. \tag{25}$$

   Thus the model for the leverage points is very different from the true regression model (24).
4. *Multivariate normal leverage points*: 10% of the observations are generated as follows. Outliers in the numerical predictors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{10}$ are drawn from a multivariate normal distribution $N(\boldsymbol{\mu}, \Sigma)$ with $\mu = (1, \ldots, 1)'$ and an ill-conditioned covariance matrix Σ. Starting with a positive-definite matrix $\boldsymbol{\Sigma}_0 = (0.5^{|i-j|})_{1 \leq i,j \leq p}$, we first compute the Eigendecomposition $\boldsymbol{\Sigma}_0 = \boldsymbol{\Gamma}\boldsymbol{\Lambda}_0\boldsymbol{\Gamma}'$, and then construct the ill-conditioned covariance matrix as $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}'$ with $\Lambda = diag(100, 0.01, \ldots, 0.01)$. Finally, we generate the values of the response as in (25).

We considered alternative contamination scenarios as well. As such, we generated large outliers in predictors not contributing to the regression model. This creates good leverage points, and the predictor will not enter the active set in the early steps, as it should be. This type of contamination is pretty harmless, even for the non-robust methods. Furthermore, we generated outliers in such a way that they were not detected in the short regressions, but still appeared as vertical outliers in the full model. The RGrpLARS method can still handle this type of outliers to a certain extent, and performs better than the other methods. The reason is that this type of outliers does not distort the sequencing of the predictors too much, and the MM-estimator applied on the selected groups is restoring the robustness.

4.1.2. Time series

In the time series case, we generate the data from an ARX(2,2) model

$$y_{t+1} = \alpha + \beta_{0,0}y_t + \beta_{0,1}y_{t-1} + \sum_{j=1}^{m}\beta_{j,0}x_{t,j} + \sum_{j=1}^{m}\beta_{j,1}x_{t-1,j} + \varepsilon_{t+1} \tag{26}$$

with i.i.d. standard normal error terms ε_{t+1}. The length of time series is set to $T = 100$ and the number of exogenous predictors $x_{t,j}$ is set to $m = 150$, resulting in three times as many individual covariates than observations. Moreover, the regression coefficients are given by $\beta_{0,0} = 0.4$, $\beta_{0,1} = 0.2$, $\beta_{1,0} = \beta_{3,0} = 1$, $\beta_{1,1} = \beta_{3,1} = 0.5$, $\beta_{2,0} = \beta_{4,0} = -1$, $\beta_{2,1} = \beta_{4,1} = -0.5$, as well as $\beta_{j,0} = \beta_{j,1} = 0$ for $j = 5, \ldots, m$. In order to generate serial correlations and cross-correlations among the exogenous candidate predictors $x_{t,j}$, they are generated from latent dynamic factors. To be more specific, we generate the latent dynamic factor $\boldsymbol{f}_t$ with $l = 2$ components by a VAR(1) model with common scalar autoregressive parameter θ as

$$\boldsymbol{f}_t = \theta \boldsymbol{f}_{t-1} + \boldsymbol{u}_t, \tag{27}$$

where $\boldsymbol{u}_t$ is i.i.d. l-dimensional standard normal. Furthermore, the choice of $\theta = 0.8$ results in strong serial correlations. Then we obtain the exogenous candidate predictors $x_{t,j}$ from

$$x_{t,j} = \boldsymbol{f}_t' \boldsymbol{\lambda}_{j,0} + \boldsymbol{f}_{t-1}' \boldsymbol{\lambda}_{j,1} + \varepsilon_{t,j}, \qquad j = 1, \ldots, m. \tag{28}$$

To generate cross-correlations between the predictor series, the elements of the l-dimensional factor loadings $\boldsymbol{\lambda}_{j,0}$ and $\boldsymbol{\lambda}_{j,1}$ are drawn from a uniform distribution $U(0, 0.5)$.

Concerning contamination, we consider both additive outliers and innovation outliers:

1. *No contamination.*

2. *Innovation outliers*: 5% of the error terms ε_{t+1} are generated from a normal distribution $N(0, \sigma^2)$ with $\sigma = 30$ instead of $N(0, 1)$.

3. *Additive outliers*: 5% of the responses y_{t+1} are replaced by values drawn from a normal distribution $N(0, \sigma^2)$ with $\sigma = 30$.

Note that an innovation outlier in the response is only problematic at the time point of its occurrence. In the consecutive two observations, it is used in the respective lagged predictor to construct the values of the response, therefore having a good leverage effect. An additive outlier in the response, on the other hand, also appears as a bad leverage point in the following two observations due to the autoregressive part of the series.

4.2. Performance measures

First, we evaluate the performance of the methods with respect to sequencing by *recall curves*, which plot the number of target variables among the first k sequenced variables, with k ranging from 1 to the total number of sequenced variables. A target variable has a non-zero coefficient in the regression function. The steeper the curve approaches the total number of target variables, the better. For comparison, we include recall curves for the oracle estimator, which sequences all target variables before any other variables.

Second, we investigate the prediction performance of the final models since the aim of variable selection is to improve prediction accuracy. For the numerical and categorical data, the estimators are evaluated by the *root mean squared prediction error* (RMSPE). In each simulation run, we generate n additional observations as test data for this purpose. Then the RMSPE is given by

$$\mathrm{RMSPE}(\hat{\boldsymbol{\beta}}) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(y_i^r - \mathbf{x}_i^{r\prime} \hat{\boldsymbol{\beta}} \right)^2}, \tag{29}$$

where y_i^r and $\mathbf{x}_i^r$, $i = 1, \ldots, n$, denote the observations of the response and predictors in the test data from the r-th simulation run. Finally, we report the average RMSPE over all R simulation runs. For the time series data, we evaluate the out-of-sample predictions via the *root mean squared forecast error* (RMSFE) at horizon h

$$\mathrm{RMSFE}_h = \sqrt{\frac{1}{R} \sum_{r=1}^{R} \left(y_{T+h}^r - \hat{y}_{T+h}^r \right)^2}, \tag{30}$$

where R denotes the number of simulation runs, y_{T+h}^r is the realized out-of-sample observation at horizon h in the r-th simulation run, and $\hat{y}_{T+h}^r$ is the corresponding h-step-ahead forecast. For both data configurations, we compute

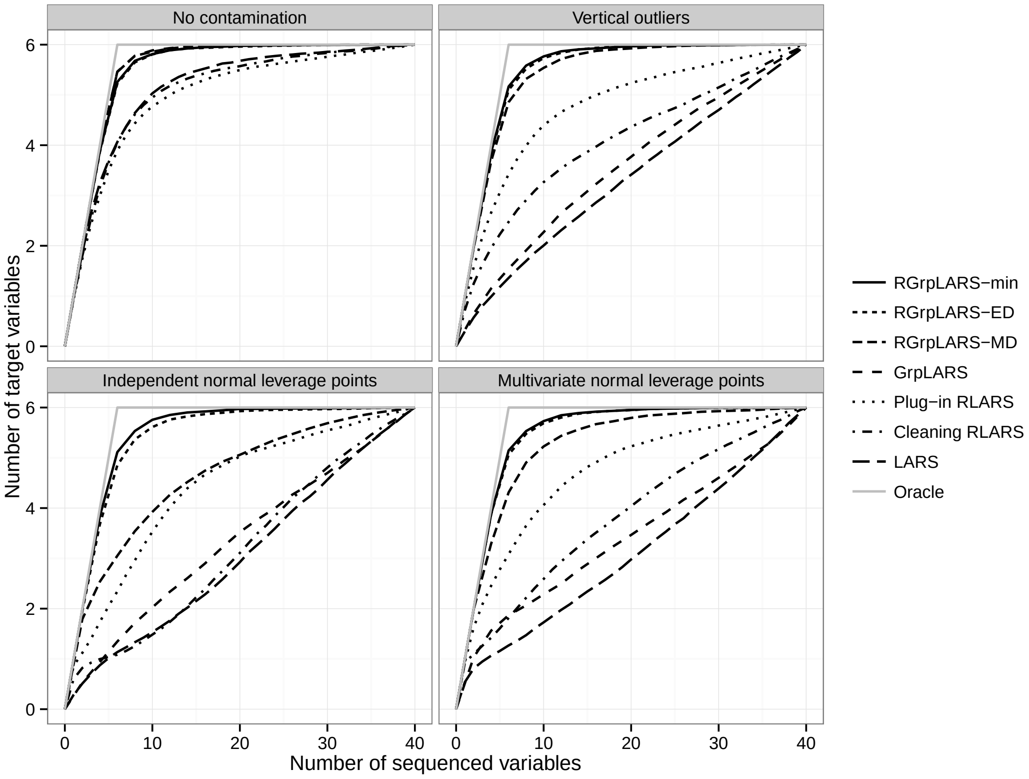

Figure 1: Recall curves for the numerical and categorical data: no contamination (*top left*), vertical outliers (*top right*), independent normal leverage points (*bottom left*) and multivariate normal leverage points (*bottom right*). The contamination level $\varepsilon = 0.1$ is used in the latter three plots.

the respective prediction performance measure for the oracle estimator, which uses the true coefficient values $\boldsymbol{\beta}$, as a benchmark for the evaluated methods.

Concerning variable selection, we evaluate the obtained models by the *false positive rate* (FPR) and the *false negative rate* (FNR). A false positive is a coefficient that is zero in the true model, but is estimated as nonzero. Analogously, a false negative is a coefficient that is nonzero in the true model, but is estimated as zero. In mathematical terms, the FPR and FNR are defined as

$$\mathrm{FPR}(\hat{\boldsymbol{\beta}}) = \frac{|\{j \in \{1, \dots, p\} : \hat{\beta}_j \neq 0 \wedge \beta_j = 0\}|}{|\{j \in \{1, \dots, p\} : \beta_j = 0\}|}, \tag{31}$$

$$\mathrm{FNR}(\hat{\boldsymbol{\beta}}) = \frac{|\{j \in \{1, \dots, p\} : \hat{\beta}_j = 0 \wedge \beta_j \neq 0\}|}{|\{j \in \{1, \dots, p\} : \beta_j \neq 0\}|}. \tag{32}$$

Both FPR and FNR are averaged over all simulation runs, and should be as small as possible for reliable model selection.

4.3. Simulation results

In this section, the simulation results for the different data configurations and contamination settings are discussed in detail.

4.3.1. Numerical and categorical data

Figure 1 (*top left*) shows the recall curves for the case without contamination. All groupwise variable selection methods perform almost as well as the oracle with respect to sequencing the important predictor variables. Differences

Table 1: Results for the numerical and categorical data from 1000 simulation runs. Methods are evaluated by the false positive rate (FPR), the false negative rate (FNR) and the root mean squared prediction error (RMSPE).

	No contamination			**Vertical outliers**			**Independent normal leverage points**			**Multivariate normal leverage points**		
Method	FPR	FNR	RMSPE	FPR	FNR	RMSPE	FPR	FNR	RMSPE	FPR	FNR	RMSPE
RGrpLARS-min	0.08	0.13	2.25	0.04	0.18	2.25	0.04	0.18	2.24	0.05	0.16	2.27
RGrpLARS-ED	0.09	0.13	2.26	0.04	0.18	2.25	0.04	0.20	2.25	0.05	0.16	2.27
RGrpLARS-MD	0.08	0.11	2.24	0.04	0.20	2.26	0.07	0.44	2.63	0.06	0.22	2.37
GrpLARS	0.02	0.08	2.14	0.02	0.95	5.50	0.21	0.67	7.62	0.19	0.64	5.76
Plug-in RLARS	0.13	0.27	2.36	0.08	0.40	2.46	0.12	0.50	2.59	0.11	0.40	2.54
Cleaning RLARS	0.12	0.26	2.34	0.08	0.59	2.65	0.06	0.85	3.06	0.13	0.61	2.83
LARS	0.11	0.25	2.31	0.01	0.98	5.04	0.33	0.66	6.57	0.37	0.61	5.14
Oracle			2.00			2.00			1.99			2.00

among the curves for the methods that select individual variables are small as well, but those methods clearly perform worse than the groupwise methods. When vertical outliers are introduced (Figure 1, *top right*), the three approaches for RGrpLARS do not lose much of their sequencing power. They outperform RLARS, where the plug-in approach in turn performs much better than the cleaning approach. As expected, GrpLARS and LARS are highly influenced by the outliers and give the worst results. In the presence of leverage points (Figure 1, *bottom row*), RGrpLARS-min and RGrpLARS-ED remain close to the oracle, but RGrpLARS-MD no longer performs so well in the case of independent normal leverage points. Also the difference between plug-in RLARS and cleaning RLARS is considerably larger than in the case with vertical outliers.

The results on FPR, FNR and RMSPE are shown in Table 1. Without contamination, the robust methods in general exhibit slightly higher FPR and RMSPE than their classical counterparts. Clearly, GrpLARS performs best with respect to all three measures. The different weighting methods for RGrpLARS yield very similar results and perform well. Differences between LARS and RLARS are small, but in particular the FNR is larger than for the groupwise methods. With vertical outliers, RGrpLARS now outperforms the other methods. The different weighting methods thereby again give very similar results. RLARS also yields good prediction performance, but suffers from a large FNR. LARS and GrpLARS even exhibit an FNR close to 1, resulting in poor prediction performance. In the two settings with leverage points, the performance of RGrpLARS-min and RGrpLARS-ED is virtually unchanged and remains the best. For independent normal leverage points, FNR and RMSPE of RGrpLARS-MD increase considerably. Plug-in RLARS in this case performs similar to RGrpLARS-MD, but somewhat worse for multivariate normal leverage points. In both cases, plug-in RLARS gives better results than cleaning RLARS. LARS and GrpLARS yield lower FNR but higher FPR than in the case with vertical outliers, in total further increasing their RMSPE.

To summarize, RGrpLARS outperforms its competitors, with RGrpLARS-min and RGrpLARS-ED being preferable as they remain stable also in the case of leverage points.

4.3.2. Time series

For the time series data, keep in mind that the number of predictor groups exceeds the number of observations, hence RGrpLARS-MD and cleaning RLARS cannot be applied.

Recall curves for the data without contamination are displayed in Figure 2 (*left*). While the groupwise methods perform very well, LARS and RLARS fail to sequence all target variables. Note also that the curves of LARS and RLARS are shorter than the ones of their groupwise counterparts. The reason is that the number of steps is limited by the number of observations in the high-dimensional case, and the groupwise methods in each step add a whole group of variables to the sequence. In the setting with innovation outliers (Figure 2, *center*), RGrpLARS stays rather close to the oracle. Surprisingly, RLARS barely outperforms GrpLARS and LARS. The situation is similar for additive outliers (Figure 2, *right*), except that the sequencing power is now somewhat lower for RGrpLARS-ED than for RGrpLARS-min. Furthermore, RLARS performs better than in the case of innovation outliers.

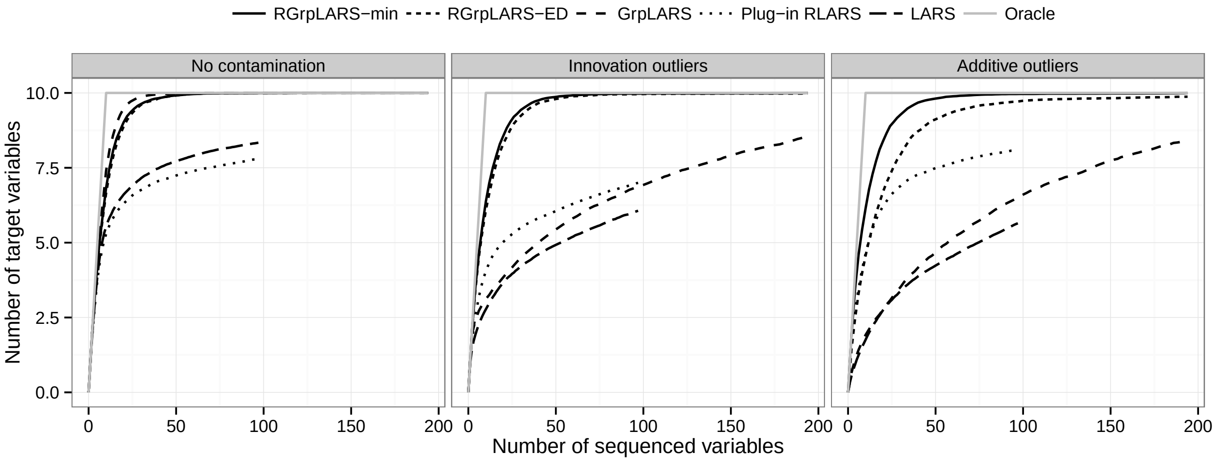

Figure 2: Recall curves for the time series data: no contamination (*left*), innovation outliers (*center*) and additive outliers (*right*). Contamination level $\varepsilon = 0.05$ is used in the latter two plots.

Table 2: Results for the time series data from 1000 simulation runs. Methods are evaluated by the average false positive rate (FPR), the average false negative rate (FNR) and the root mean squared forecast error (RMSFE).

	No contamination			**Innovation outliers**			**Additive outliers**		
Method	FPR	FNR	RMSFE	FPR	FNR	RMSFE	FPR	FNR	RMSFE
RGrpLARS-min	0.04	0.03	1.38	0.04	0.05	1.38	0.04	0.09	1.51
RGrpLARS-ED	0.05	0.03	1.39	0.05	0.06	1.45	0.05	0.24	1.97
GrpLARS	0.02	0.01	1.23	0.01	0.73	3.92	0.01	0.88	4.48
Plug-in RLARS	0.05	0.37	1.55	0.03	0.51	1.64	0.03	0.41	1.57
LARS	0.03	0.38	1.49	0.01	0.82	3.47	0.00	0.94	4.06
Oracle			1.03			1.04			0.98

Table 2 contains the results on FPR, FNR and RMSFE. When there is no contamination, GrpLARS yields the best forecast performance, with RGrpLARS being in second place. In addition, FPR and FNR are close to 0 for the groupwise methods. LARS and RLARS, on the other hand, have somewhat higher RMSFE than their groupwise counterparts due to a considerably higher FNR. When innovation outliers are introduced, the results for RGrpLARS barely change. RGrpLARS now outperforms the other methods, followed by RLARS. LARS and GrpLARS exhibit a very high FNR, yielding poor forecast performance. With additive outliers, RGrpLARS is still the best with respect to all three performance measures, although the RMSFE slightly increased. Also RLARS remains stable and performs well. For LARS and GrpLARS, there is a further increase in FNR (which is now close to 1) and RMSFE.

Again, RGrpLARS clearly outperforms the other methods. RGrpLARS-min thereby shows somewhat higher sequencing power than RGrpLARS-ED in the presence of additive outliers.

5. Real data example

As real data example, we use information on cars that we scraped from the website of the popular BBC television show *Top Gear* (`http://www.topgear.com/uk/`). The data set is included in the R package **robustHD** (Alfons, 2014a) and contains $n = 242$ complete observations on $m = 29$ numerical and categorical variables. There are 4 categorical variables with two possible outcomes, and 12 categorical variables having three levels. A description of

Table 3: Description of the *Top Gear* car data.

Variable	**Description**	**Possible outcomes**		
MPG	Fuel consumption (in miles per gallon)			
Fuel	Type of fuel	Diesel	Petrol	
Price	List price (in UK pounds)			
Cylinders	Number of cylinders in the engine			
Displacement	Displacement of the engine (in cc)			
DriveWheel	Type of drive wheel	4WD	Front	Rear
BHP	Power of the engine (in bhp)			
Torque	Torque of the engine (in lb/ft)			
Acceleration	Time from 0 to 62 mph (in seconds)			
TopSpeed	Top speed (in mph)			
Weight	Curb weight (in kg)			
Length	Length (in mm)			
Width	Width (in mm)			
Height	Height (in mm)			
AdaptiveHeadlights	Whether the car has adaptive headlights	no	optional	standard
AdjustableSteering	Whether the car has adjustable steering	no		standard
AlarmSystem	Whether the car has an alarm system	no/optional		standard
Automatic	Whether the car has an automatic transmission	no	optional	standard
Bluetooth	Whether the car has bluetooth	no	optional	standard
ClimateControl	Whether the car has climate control	no	optional	standard
CruiseControl	Whether the car has cruise control	no	optional	standard
ElectricSeats	Whether the car has electric seats	no	optional	standard
Leather	Whether the car has a leather interior	no	optional	standard
ParkingSensors	Whether the car has parking sensors	no	optional	standard
PowerSteering	Whether the car has power steering	no		standard
SatNav	Whether the car has a satellite navigation system	no	optional	standard
ESP	Whether the car has ESP	no	optional	standard
Verdict	Review score			
Origin	Origin of the car maker	Asia	Europe	USA

the variables is provided in Table 3. We use variable *MPG* (fuel consumption) as the response and the remaining variables as predictors. The resulting design matrix consists of 12 numerical variables, 4 individual dummy variables, and 12 groups of two dummy variables each, giving a total of 40 individual covariates. Note that we log-transform the variable *Price* (list price) to remove skewness.

We apply the same methods as in the simulations and use the same strategies to select the final models. In Table 4, the selected predictors for each method are listed in the order that they are sequenced. The three approaches for RGrpLARS yield very similar sequences, with RGrpLARS-Min and RGrpLARS-ED even producing the same final model. RGrpLARS-MD only differs in that it does not include variable *Width*. All selected variables seem reasonable in the context of fuel consumption, making interpretation easy. The GrpLARS model, on the other hand, is rather small and omits some variables that seem important in the context (e.g., *BHP*, *Weight*, *Acceleration*). Thus GrpLARS may be influenced by outliers in the data. The model selected by Plug-in RLARS appears to make more sense, but keeps only one dummy variable for the predictor group *DriveWheel*. Furthermore, cleaning RLARS and LARS both include many variables that seem difficult to interpret and are therefore not further discussed.

Concerning outlier detection, Figure 3 (*left*) shows a regression diagnostic plot (Rousseeuw and van Zomeren, 1990) of the final model obtained via RGrpLARS-min. It reveals three clear outliers: BMW i3 (observation 40), Chevrolet Volt (observation 53) and Vauxhall Ampera (observation 216). All three are electric cars with an additional

Table 4: *Top Gear* car data: predictors in the final models for the response variable *MPG* (fuel consumption), in the order that they are sequenced. For the groupwise methods, we report the selected predictor groups by the names of the underlying original variables. For the methods selecting individual covariates, we use the indicator function notation *I*(.) to report dummy variables.

Methods	**Predictors in the final model**
RGrpLARS-min	*BHP*, *Displacement*, *Acceleration*, *Fuel*, *Weight*, *DriveWheel*, *Width*, *Height*, *TopSpeed*
RGrpLARS-ED	*BHP*, *Displacement*, *Acceleration*, *Weight*, *Fuel*, *DriveWheel*, *Width*, *Height*, *TopSpeed*
RGrpLARS-MD	*BHP*, *Displacement*, *Fuel*, *Weight*, *DriveWheel*, *Acceleration*, *TopSpeed*, *Height*
GrpLARS	*Displacement*, *TopSpeed*, *Verdict*, *Automatic*, *Height*
Plug-in RLARS	*BHP*, *I*(*DriveWheel* = *Front*), *Acceleration*, *Displacement*, *I*(*Fuel* = *Petrol*), *Weight*, *Width*, *TopSpeed*, *I*(*AdustableSteering* = *standard*), *Height*
Cleaning RLARS	*BHP*, *Displacement*, *I*(*DriveWheel* = *Front*), *I*(*Fuel* = *Petrol*), *Weight*, *Height*, *TopSpeed*, *I*(*Bluetooth* = *standard*), *I*(*CruiseControl* = *optional*), *I*(*Automatic* = *optional*), *I*(*Origin* = *Europe*), *I*(*Automatic* = *standard*), *Verdict*, *Width*
LARS	*Displacement*, *TopSpeed*, *I*(*Automatic* = *standard*), *Verdict*, *I*(*ParkingSensors* = *optional*), *Height*, *I*(*Leather* = *optional*), *I*(*Bluetooth* = *standard*)

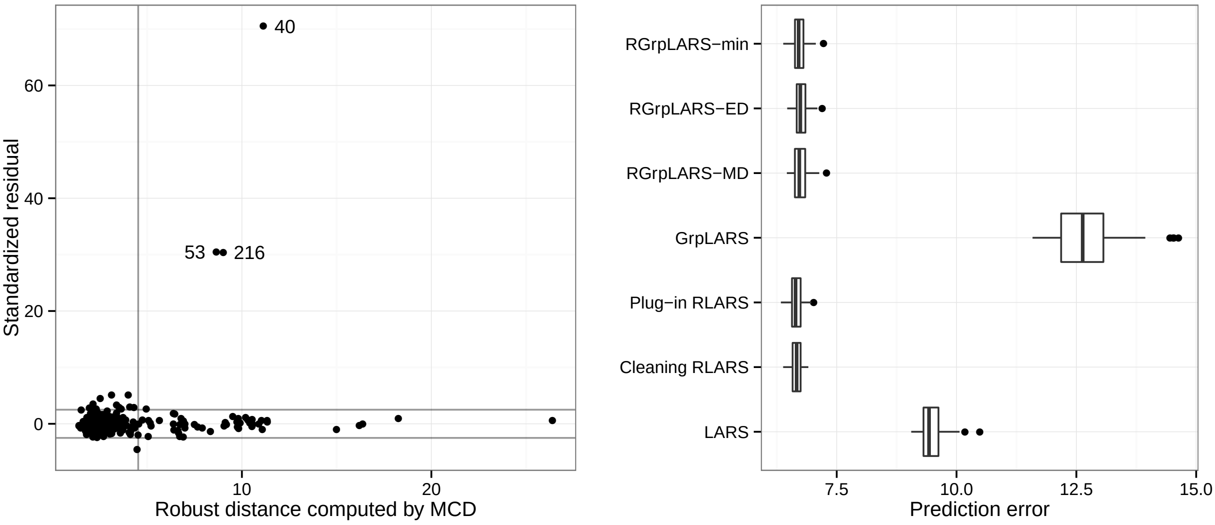

Figure 3: *Top Gear* car data: regression diagnostic plot for the final RGrpLARS-min model (*left*), and root trimmed mean squared prediction error with 5% trimming estimated via repeated five-fold cross-validation with 100 repetitions (*right*).

petrol-powered range extender engine. Those three car models are also flagged as outliers by the other robust methods.

In Figure 3 (*right*) and Table 5, we evaluate the prediction performance of the different methods via repeated five fold cross-validation with 100 replications. We use the root trimmed mean squared prediction error (RTMSPE) as prediction loss, that is, the root mean squared prediction error where a certain percentage of the largest squared prediction errors are dropped. The RTMSPE is an intuitively appealing robust measure of prediction loss, as it allows to identify the models that can be expected to best predict a given percentage of future data. Even though only three clear outliers were detected among the $n = 242$ observations (about 1.2%), we use the RTMSPE with 5% trimming to be on the conservative side. Hence we identify the models that best predict 95% of future data. Clearly, all robust methods perform comparably well, whereas LARS and groupwise LARS result in poor prediction performance. Interestingly, groupwise LARS is much more affected by the outliers than LARS. Other trimming proportions (e.g., 2% or 10%) give a similar picture.

Table 5: *Top Gear* car data: root trimmed mean squared prediction error with 5% trimming estimated via repeated five-fold cross-validation with 100 repetitions. The results from the 100 repetitions are summarized by median, interquartile range (IQR), mean and standard deviation (SD).

	Median	IQR	Mean	SD
RGrpLARS-min	6.71	0.18	6.72	0.15
RGrpLARS-ED	6.74	0.18	6.76	0.15
RGrpLARS-MD	6.72	0.22	6.74	0.17
GrpLARS	12.63	0.88	12.72	0.69
Plug-in RLARS	6.64	0.18	6.65	0.14
Cleaning RLARS	6.66	0.17	6.66	0.12
LARS	9.43	0.31	9.48	0.26

Table 6: *Top Gear* car data without the three clear outliers: predictors in the final models for the response variable *MPG* (fuel consumption), in the order that they are sequenced. For the groupwise methods, we report the selected predictor groups by the names of the underlying original variables. For the methods selecting individual covariates, we use the indicator function notation *I*(.) to report dummy variables.

Methods	**Predictors in the final model**
RGrpLARS-min	*BHP*, *Displacement*, *Acceleration*, *Fuel*, *Weight*, *DriveWheel*, *Width*, *Height*, *TopSpeed*
RGrpLARS-ED	*BHP*, *Displacement*, *Acceleration*, *Weight*, *Fuel*, *DriveWheel*, *Width*, *Height*, *TopSpeed*
RGrpLARS-MD	*BHP*, *Displacement*, *Fuel*, *Acceleration*, *Weight*, *DriveWheel*, *Width*, *Height*, *TopSpeed*
GrpLARS	*BHP*, *Displacement*, *Acceleration*, *Fuel*, *Weight*, *DriveWheel*, *Width*, *Height*, *AdjustableSteering*, *TopSpeed*
Plug-in RLARS	*BHP*, *I*(*DriveWheel* = *Front*), *Acceleration*, *Displacement*, *Cylinders*, *I*(*Fuel* = *Petrol*), *Width*, *Weight*, *I*(*AdustableSteering* = *standard*), *I*(*Bluetooth* = *standard*), *I*(*AdaptiveHeadlights* = *standard*), *TopSpeed*, *Height*
Cleaning RLARS	*BHP*, *Displacement*, *I*(*DriveWheel* = *Front*), *I*(*Fuel* = *Petrol*), *Weight*, *Width*, *Height*, *I*(*Bluetooth* = *standard*), *Acceleration*, *TopSpeed*
LARS	*BHP*, *Displacement*, *I*(*DriveWheel* = *Front*), *Acceleration*, *I*(*Fuel* = *Petrol*), *Weight*, *Width*, *I*(*Bluetooth* = *standard*), *Height*, *TopSpeed*

For comparison, we apply the methods again to the data without the three clear outliers. The results are shown in Table 6. RGrpLARS-min and RGrpLARS-ED yield the same model as before. RGrpLARS-MD now results in the same model as the other two RGrpLARS approaches, although the order of the predictors is different. GrpLARS mostly agrees with RGrpLARS, but also includes the variable *AdjustableSteering*. The models obtained by LARS and RLARS are still difficult to interpret, in particular the the plug-in RLARS model. Cleaning RLARS now yields the same model as LARS except for the order of the covariates.

6. Conclusions

When regression problems exhibit a natural grouping among predictor variables, model selection typically aims at selecting whole predictor groups rather than individual covariates for better interpretation of the resulting models. This paper discusses an algorithm for groupwise least angle regression together with a robust extension based on different data cleaning strategies. The flexibility and the excellent performance of the proposed method are demonstrated by simulation studies for numerical and categorical predictor groups, as well as for an autoregressive time series model with exogenous inputs. A real data example regarding fuel consumption of cars provides further indication that our procedure works well in practice.

We propose three approaches for cleaning the data: (a) taking the minimum weight from robust short regressions of the response on each predictor group, (b) computing weights based on Euclidean distances of the multivariate set of standardized residuals from such robust short regressions, or (c) computing weights based on Mahalanobis distances of those standardized residuals. Besides being applicable when the number of predictor groups is larger than the number of observations, the first two approaches led to better results in our numerical experiments and are therefore recommended. While we only showed the results for moderate contamination levels in this paper, further experiments indicate that robust groupwise LARS can withstand much higher contamination levels in many situations. Nevertheless, it should be noted that the performance decreases as the contamination level approaches the breakdown point of the robust regression estimator used.

The proposed procedures are implemented in the R package **robustHD**, which is freely available on CRAN (the Comprehensive R Archive Network, `http://www.CRAN.R-project.org`). Since the sequencing step is fully implemented in C++, groupwise LARS and its robust modifications are very fast to compute.

Appendix A. Proofs

Proof of Lemma 1. The three statements of Lemma 1 are proven by induction. We know that

$$r_1 = \operatorname{cor}(z_0, \tilde{\boldsymbol{x}}_{(1)}) = \operatorname{cor}(z_0, \hat{z}_0^{(1)}) \geq 0,$$

with $\hat{z}_0^{(1)}$ the fitted value of the least squares regression of z_0 on $\underline{\boldsymbol{X}}_1$. Therefore (a) holds for $k = 1$. Since the first index in the active set A by construction yields the largest $R^2(z_0 \sim \underline{\boldsymbol{X}}_j)$ and $R^2(z_0 \sim \underline{\boldsymbol{X}}_{(1)}) = R^2(z_0 \sim \tilde{\boldsymbol{x}}_{(1)}) = r_1^2$, (b) holds for $k = 1$. Condition (c) for $k = 1$ reduces to $0 \leq \gamma_1 \leq r_1$, as the equiangular vector $\boldsymbol{u}_1 = \tilde{\boldsymbol{x}}_{(1)}$ implies $a_1 = 1$. For $\gamma = 0$, we have $R^2(z_0 \sim \tilde{\boldsymbol{x}}_{(1)}) = R^2(z_0 \sim \underline{\boldsymbol{X}}_{(1)}) \geq R^2(z_0 \sim \underline{\boldsymbol{X}}_j)$ by construction of the first index in the active set A. For $\gamma = r_1$, on the other hand, we have $R^2(z_0 - r_1\tilde{\boldsymbol{x}}_{(1)} \sim \tilde{\boldsymbol{x}}_{(1)}) = R^2(z_0 - \hat{z}_0^{(1)} \sim \underline{\boldsymbol{X}}_{(1)}) = 0 \leq R^2(z_0 - r_1\tilde{\boldsymbol{x}}_{(1)} \sim \underline{\boldsymbol{X}}_j)$, since $z_0 - \hat{z}_0^{(1)}$ contains the residuals from least squares regression of z_0 on $\underline{\boldsymbol{X}}_{(1)}$. Hence there must exist a γ_1 between 0 and r_1 for which the generalized equi-correlation condition (5) holds, from which it follows that (c) holds for $k = 1$.

Suppose now that the three statements of Lemma 1 hold for $k - 1$. We prove that then they also hold for k. Concerning (a), we have for every $l = 1, \ldots, k - 1$

$$\begin{aligned}\operatorname{cor}(\boldsymbol{z}_{k-1}, \tilde{\boldsymbol{x}}_{(l)}) = \operatorname{cov}(\boldsymbol{z}_{k-1}, \tilde{\boldsymbol{x}}_{(l)}) &= \operatorname{cov}\left(\frac{\boldsymbol{z}_{k-2} - \gamma_{k-1}\boldsymbol{u}_{k-1}}{\sigma_{k-1}}, \tilde{\boldsymbol{x}}_{(l)}\right) = \frac{\operatorname{cor}(\boldsymbol{z}_{k-2}, \tilde{\boldsymbol{x}}_{(l)}) - \gamma_{k-1}\operatorname{cor}(\boldsymbol{u}_{k-1}, \tilde{\boldsymbol{x}}_{(l)})}{\sigma_{k-1}} \\ &= \frac{r_{k-1} - \gamma_{k-1}a_{k-1}}{\sigma_{k-1}} \geq 0, \end{aligned} \tag{A.1}$$

which does not depend on l. Here we used (a) for $k - 1$ and (3), as well as the fact that $\boldsymbol{z}_{k-1}$ and $\tilde{\boldsymbol{x}}_{(l)}$ are standardized. The last inequality holds since (c) holds for $k - 1$. Furthermore, using (5) for $k - 1$ yields

$$R^2(\boldsymbol{z}_{k-1} \sim \tilde{\boldsymbol{x}}_{(k-1)}) = R^2(\boldsymbol{z}_{k-1} \sim \underline{\boldsymbol{X}}_{(k)}),$$

or, equivalently,

$$\operatorname{cor}^2(\boldsymbol{z}_{k-1}, \tilde{\boldsymbol{x}}_{(k-1)}) = \operatorname{cor}^2(\boldsymbol{z}_{k-1}, \tilde{\boldsymbol{x}}_{(k)}).$$

Since $\tilde{\boldsymbol{x}}_{(k)}$ are the standardized fitted values from least squares regression of $\boldsymbol{z}_{k-1}$ on $\underline{\boldsymbol{X}}_{(k)}$, we also have $\operatorname{cor}(\boldsymbol{z}_{k-1}, \tilde{\boldsymbol{x}}_{(k-1)}) = \operatorname{cor}(\boldsymbol{z}_{k-1}, \tilde{\boldsymbol{x}}_{(k)}) \geq 0$. We conclude that (a) holds for k.

To prove (b) for k, suppose that there exists a $j \in A^c$ such that

$$r_k^2 = R^2(\boldsymbol{z}_{k-1} \sim \tilde{\boldsymbol{x}}_{(1)}) = \ldots = R^2(\boldsymbol{z}_{k-1} \sim \tilde{\boldsymbol{x}}_{(k)}) < R^2(\boldsymbol{z}_{k-1} \sim \underline{\boldsymbol{X}}_j). \tag{A.2}$$

We use the shortened notation $f_0(\gamma) = R^2(\boldsymbol{z}_{k-2} - \gamma\boldsymbol{u}_{k-1} \sim \tilde{\boldsymbol{x}}_{(k-1)})$ for the left hand side of (5) for $k - 1$, and $f_j(\gamma) = R^2(\boldsymbol{z}_{k-2} - \gamma\boldsymbol{u}_{k-1} \sim \underline{\boldsymbol{X}}_j)$ for the right hand side of (5) for $k - 1$. By definition, γ_{k-1} is smaller than any positive solution of $f_0(\gamma) = f_j(\gamma)$. Since (b) holds for $k - 1$, we have $f_0(0) \geq f_j(0)$. But (A.2) implies

$$f_0(\gamma_{k-1}) = R^2(\boldsymbol{z}_{k-1} \sim \tilde{\boldsymbol{x}}_{(k-1)}) < R^2(\boldsymbol{z}_{k-1} \sim \underline{\boldsymbol{X}}_j) = f_j(\gamma_{k-1}).$$

Hence there must exist a positive $\tilde{\gamma} < \gamma_{k-1}$ for which $f_0(\tilde{\gamma}) = f_j(\tilde{\gamma})$. This contradicts the definition of γ_{k-1}, thus (A.2) cannot hold, which proves (b) for k.

Finally, we show that (c) holds for k. It is sufficient to show that there always exists a solution γ_k of (5) with $0 \leq \gamma_k \leq r_k/a_k$. For $\gamma = 0$, we have $r_k^2 = R^2(\boldsymbol{z}_{k-1} \sim \tilde{\boldsymbol{x}}_{(k)}) \geq R^2(\boldsymbol{z}_{k-1} \sim \underline{\boldsymbol{X}}_j)$ since we have already proven (b) for $k-1$. For $\gamma = r_k/a_k$, on the other hand, similar calculations as in (A.1) yield

$$R^2(\boldsymbol{z}_{k-1} - \gamma \boldsymbol{u}_k \sim \tilde{\boldsymbol{x}}_{(k)}) = \operatorname{cor}^2(\boldsymbol{z}_{k-1} - \gamma \boldsymbol{u}_k, \tilde{\boldsymbol{x}}_{(k)}) = \frac{(r_k - \gamma a_k)^2}{\sigma_k^2} = 0 \leq R^2(\boldsymbol{z}_{k-1} - \gamma \boldsymbol{u}_k \sim \underline{\boldsymbol{X}}_j).$$

Hence there must exist a solution γ_k to (5) between 0 and r_k/a_k. □

Proof of Equation (7). The left hand side of (5) can be rewritten as

$$R^2(\boldsymbol{z}_{k-1} - \gamma \boldsymbol{u}_k \sim \tilde{\boldsymbol{x}}_{(k)}) = \operatorname{cor}^2(\boldsymbol{z}_{k-1} - \gamma \boldsymbol{u}_k, \tilde{\boldsymbol{x}}_{(k)}) = \frac{(\operatorname{cov}(\boldsymbol{z}_{k-1}, \tilde{\boldsymbol{x}}_{(k)}) - \gamma \operatorname{cov}(\boldsymbol{u}_k, \tilde{\boldsymbol{x}}_{(k)}))^2}{\operatorname{var}(\boldsymbol{z}_{k-1} - \gamma \boldsymbol{u}_k)} = \frac{(r_k - \gamma a_k)^2}{\operatorname{var}(\boldsymbol{z}_{k-1} - \gamma \boldsymbol{u}_k)}.$$

Here we used (3), (6) and the fact that $\tilde{\boldsymbol{x}}_{(k)}$ is standardized. The right hand side of (5) can be rewritten as

$$R^2(\boldsymbol{z}_{k-1} - \gamma \boldsymbol{u}_k \sim \underline{\boldsymbol{X}}_j) = 1 - \frac{(\boldsymbol{z}_{k-1} - \gamma \boldsymbol{u}_k)'(\boldsymbol{I} - \boldsymbol{H}_j)(\boldsymbol{z}_{k-1} - \gamma \boldsymbol{u}_k)}{(\boldsymbol{z}_{k-1} - \gamma \boldsymbol{u}_k)'(\boldsymbol{z}_{k-1} - \gamma \boldsymbol{u}_k)} = \frac{(\boldsymbol{z}_{k-1} - \gamma \boldsymbol{u}_k)' \boldsymbol{H}_j (\boldsymbol{z}_{k-1} - \gamma \boldsymbol{u}_k)}{(n-1) \operatorname{var}(\boldsymbol{z}_{k-1} - \gamma \boldsymbol{u}_k)},$$

where $\boldsymbol{H}_j = \underline{\boldsymbol{X}}_j(\underline{\boldsymbol{X}}_j' \underline{\boldsymbol{X}}_j)^{-1} \underline{\boldsymbol{X}}_j'$ is the projection matrix on the space spanned by $\underline{\boldsymbol{X}}_j$ and $j \in A^c$. Hence condition (5) is equivalent to

$$\begin{aligned}
\frac{(r_k - \gamma a_k)^2}{\operatorname{var}(\boldsymbol{z}_{k-1} - \gamma \boldsymbol{u}_k)} &= \frac{(\boldsymbol{z}_{k-1} - \gamma \boldsymbol{u}_k)' \boldsymbol{H}_j (\boldsymbol{z}_{k-1} - \gamma \boldsymbol{u}_k)}{(n-1) \operatorname{var}(\boldsymbol{z}_{k-1} - \gamma \boldsymbol{u}_k)} \\
(r_k - \gamma a_k)^2 &= \frac{(\boldsymbol{z}_{k-1} - \gamma \boldsymbol{u}_k)' \boldsymbol{H}_j (\boldsymbol{z}_{k-1} - \gamma \boldsymbol{u}_k)}{n-1} \\
r_k^2 - 2 a_k r_k \gamma + a_k^2 \gamma^2 &= \frac{\boldsymbol{z}_{k-1}' \boldsymbol{H}_j \boldsymbol{z}_{k-1}}{n-1} - 2 \frac{\boldsymbol{u}_k' \boldsymbol{H}_j \boldsymbol{z}_{k-1}}{n-1} \gamma + \frac{\boldsymbol{u}_k' \boldsymbol{H}_j \boldsymbol{u}_k}{n-1} \gamma^2.
\end{aligned}$$

After rearranging the terms and using

$$\begin{aligned}
\frac{\boldsymbol{z}_{k-1}' \boldsymbol{H}_j \boldsymbol{z}_{k-1}}{n-1} &= \frac{\boldsymbol{z}_{k-1}' \boldsymbol{H}_j' \boldsymbol{H}_j \boldsymbol{z}_{k-1}}{n-1} = \frac{\left(\hat{\boldsymbol{z}}_{k-1}^j\right)' \hat{\boldsymbol{z}}_{k-1}^j}{n-1} = \operatorname{var}(\hat{\boldsymbol{z}}_{k-1}^j) = r_{k,j}^2 \\
\frac{\boldsymbol{u}_k' \boldsymbol{H}_j \boldsymbol{z}_{k-1}}{n-1} &= \frac{\boldsymbol{u}_k' \hat{\boldsymbol{z}}_{k-1}^j}{n-1} = \operatorname{cov}(\boldsymbol{u}_k, \hat{\boldsymbol{z}}_{k-1}^j) = \sqrt{\operatorname{var}(\hat{\boldsymbol{z}}_{k-1}^j)} \operatorname{cor}(\boldsymbol{u}_k, \hat{\boldsymbol{z}}_{k-1}^j) = r_{k,j} a_{k,j}, \\
\frac{\boldsymbol{u}_k' \boldsymbol{H}_j \boldsymbol{u}_k}{n-1} &= \frac{\boldsymbol{u}_k' \boldsymbol{H}_j' \boldsymbol{H}_j \boldsymbol{u}_k}{n-1} = \frac{\left(\hat{\boldsymbol{u}}_k^j\right)' \hat{\boldsymbol{u}}_k^j}{n-1} = \operatorname{var}(\hat{\boldsymbol{u}}_k^j) = \tau_{k,j}^2,
\end{aligned}$$

the quadratic equation (7) follows. □

Appendix B. Algorithms

In Algorithm 1, the groupwise LARS algorithm is presented in detail. Note that faster computation of r_k, $r_{k,j}$ and σ_k with iterative formulas is possible such that the current response $\boldsymbol{z}_{k-1}$ is no longer explicitly required in Step 3. However, these minor improvements are of a technical nature and are not further discussed. The robust groupwise LARS algorithm is then obtained performing the data cleaning step from Algorithm 2 before entering Algorithm 1 with the cleaned data.

References

Alfons, A., 2014a. **robustHD**: Robust methods for high-dimensional data. R package version 0.5.0.

Alfons, A., 2014b. **simFrame**: Simulation framework. R package version 0.5.3.

Alfons, A., Baaske, W., Filzmoser, P., Mader, W., Wieser, R., 2011. Robust variable selection with application to quality of life research. Statistical Methods & Applications 20 (1), 65–82.

Alfons, A., Croux, C., Gelper, S., 2013. Sparse least trimmed squares regression for analyzing high-dimensional large data sets. The Annals of Applied Statistics 7 (1), 226–248.

Alfons, A., Templ, M., Filzmoser, P., 2010. An object-oriented framework for statistical simulation: The R package **simFrame**. Journal of Statistical Software 37 (3), 1–36.

Breheny, P., Huang, J., 2009. Penalized methods for bi-level variable selection. Statistics and Its Interface 2 (3), 369–380.

Breiman, L., 1995. Better subset regression using the nonnegative garrote. Technometrics 37 (4), 373–384.

Chen, X., Wang, Z., McKeown, M., 2010. FMRI group studies of brain connectivity via a group robust lasso. In: 2010 IEEE International Conference on Image Processing. Hong Kong, pp. 589–592.

Dudoit, S., Fridlyand, J., Speed, T., 2001. Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association 97 (457), 77–87.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. The Annals of Statistics 32 (2), 407–499.

Hastie, T., Efron, B., 2013. **lars**: Least angle regression, lasso and forward stagewise. R package version 1.2.

Khan, J., Van Aelst, S., Zamar, R., 2007. Robust linear model selection based on least angle regression. Journal of the American Statistical Association 102 (480), 1289–1299.

Khan, J., Van Aelst, S., Zamar, R., 2010. Fast robust estimation of prediction error based on resampling. Computational Statistics & Data Analysis 54 (12), 3121–3130.

Maronna, R., Martin, D., Yohai, V., 2006. Robust Statistics: Theory and Methods. John Wiley & Sons, Chichester, ISBN 978-0-470-01092-1.

McCann, L., Welsch, R., 2007. Robust variable selection using least angle regression and elemental set sampling. Computational Statistics & Data Analysis 52 (1), 249–257.

Meier, L., van de Geer, S., Brühlmann, P., 2008. The group lasso for logistic regression. Journal of the Royal Statistical Society, Series B 70 (1), 53–71.

R Development Core Team, 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Ronchetti, E., Field, C., Blanchard, W., 1997. Robust linear model selection by cross-validation. Journal of the American Statistical Association 92 (439), 1017–1023.

Rousseeuw, P., van Zomeren, B., 1990. Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association 85 (411), 633–639.

Salibian-Barrera, M., Van Aelst, S., 2008. Robust model selection using fast and robust bootstrap. Computational Statistics & Data Analysis 52 (12), 5121–5135.

Schwarz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6 (2), 461–464.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58 (1), 267–288.

Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G., 2003. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. Statistical Science 18 (1), 104–117.

Witten, D., Tibshirani, R., Hastie, T., 2009. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10 (3), 515–534.

Yohai, V., 1987. High breakdown-point and high efficiency robust estimates for regression. The Annals of Statistics 15 (20), 642–656.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B 68 (1), 49–67.

Zhao, P., Rocha, G., Yu, B., 2009. The composite absolute penalties family for grouped and hierarchical variable selection. The Annals of Statistics 37 (6), 3468–3497.

Algorithm 1 Groupwise least angle regression

```
1. Initialization
    $\underline{\boldsymbol{X}}_j \leftarrow$ matrix of standardized variables of predictor group $j = 1, \ldots, m$
    $z_0 \leftarrow$ standardized response

2. Find first predictor group
    $\hat{z}_0^j \leftarrow$ fitted values of least squares regression $z_0 \sim \underline{\boldsymbol{X}}_j$, $j = 1, \ldots, m$
    $(1) \leftarrow \arg\max_{j \in \{1,\ldots,m\}} R_j^2/p_j$ with $R_j^2 = R^2(z_0 \sim \underline{\boldsymbol{X}}_j) = \mathrm{cor}^2(z_0, \hat{z}_0^j)$
    $A = \{(1)\}$

3. Sequence remaining predictor groups
    for $k = 1, \ldots, \min(m, n-1) - 1$ do
        $\tilde{\boldsymbol{x}}_{(k)} \leftarrow$ standardized $\hat{\boldsymbol{z}}_{k-1}^{(k)}$
        $r_k \leftarrow \mathrm{cor}(\boldsymbol{z}_{k-1}, \tilde{\boldsymbol{x}}_{(k)}) / \sqrt{p_{(k)}}$
        $\boldsymbol{R}_A \leftarrow$ correlation matrix of $\tilde{\boldsymbol{X}}_A = (\tilde{\boldsymbol{x}}_{(1)}, \ldots, \tilde{\boldsymbol{x}}_{(k)})$
        $\boldsymbol{q}_k \leftarrow (\sqrt{p_{(1)}}, \ldots, \sqrt{p_{(k)}})'$
        $a_k \leftarrow \left(\boldsymbol{q}_k' \boldsymbol{R}_A^{-1} \boldsymbol{q}_k\right)^{-1/2}$
        $\boldsymbol{w}_k \leftarrow a_k \boldsymbol{R}_A^{-1} \boldsymbol{q}_k$
        $\boldsymbol{u}_k \leftarrow \tilde{\boldsymbol{X}}_A \boldsymbol{w}_k$
        for all $j \in A^c$ do
            $r_{k,j} \leftarrow \mathrm{cor}(\boldsymbol{z}_{k-1}, \hat{\boldsymbol{z}}_{k-1}^j) / \sqrt{p_j}$
            $a_{k,j} \leftarrow \mathrm{cor}(\boldsymbol{u}_k, \hat{\boldsymbol{z}}_{k-1}^j) / \sqrt{p_j}$
            $\hat{\boldsymbol{u}}_k^j \leftarrow$ fitted values of least squares regression $\boldsymbol{u}_k \sim \underline{\boldsymbol{X}}_j$
            $\tau_{k,j} \leftarrow \mathrm{cor}(\boldsymbol{u}_k, \hat{\boldsymbol{u}}_k^j) / \sqrt{p_j}$
            $\overline{\gamma}_j \leftarrow$ smallest positive solution of quadratic equation (7)
        end for
        $(k+1) \leftarrow \arg\min_{j \in A^c} \overline{\gamma}_j$
        $\gamma_k \leftarrow \overline{\gamma}_{(k+1)}$
        $\boldsymbol{z}_k \leftarrow (\boldsymbol{z}_{k-1} - \gamma_k \boldsymbol{u}_k)/\sigma_k$ with $\sigma_k^2 = \mathrm{var}(\boldsymbol{z}_{k-1} - \gamma_k \boldsymbol{u}_k)$
        for all $j \in A^c$ do
            $\hat{\boldsymbol{z}}_k^j \leftarrow (\hat{\boldsymbol{z}}_{k-1}^j - \gamma_k \hat{\boldsymbol{u}}_k^j)/\sigma_k$
        end for
        $A \leftarrow A \cup \{(k+1)\}$
    end for

4. Evaluate submodels along the obtained sequence
```

Algorithm 2 Initial data cleaning step for robust groupwise least angle regression

```
$\underline{\boldsymbol{X}}_j^* \leftarrow$ matrix of robustly standardized variables of predictor group $j = 1, \dots, m$
$\boldsymbol{z}^* \leftarrow$ robustly standardized response
select one of the following:

a. Minimum weight from robust short regressions
    obtain weights $\omega_{ij}$, $i = 1, \dots, n$, from robust regression $\boldsymbol{z}^* \sim \underline{\boldsymbol{X}}_j^*$, $j = 1, \dots, m$
    define weights $\omega_i \leftarrow \min_{j \in \{1,\dots,m\}} \sqrt{\omega_{ij}}$, $i = 1, \dots, n$

b. Weights based on Euclidean distances
    $\boldsymbol{S} \leftarrow$ standardized residuals from robust regressions $\boldsymbol{z}^* \sim \underline{\boldsymbol{X}}_j^*$, $j = 1, \dots, m$
    compute weights $\omega_i \leftarrow \min\left(\sqrt{c/(\mathbf{s}_i' \mathbf{s}_i)}, 1\right)$, $i = 1, \dots, n$

c. Weights based on Mahalanobis distances
    $\boldsymbol{S} \leftarrow$ standardized residuals from robust regressions $\boldsymbol{z}^* \sim \underline{\boldsymbol{X}}_j^*$, $j = 1, \dots, m$
    $\hat{\boldsymbol{V}} \leftarrow$ robust correlation matrix of $\boldsymbol{S}$
    compute weights $\omega_i \leftarrow \min\left(\sqrt{c/(\mathbf{s}_i' \hat{\boldsymbol{V}}^{-1} \mathbf{s}_i)}, 1\right)$, $i = 1, \dots, n$

end select
obtain the cleaned predictor groups $\boldsymbol{X}_j^* \leftarrow \left(\omega_1 \underline{\mathbf{x}}_{1j}^*, \dots, \omega_n \underline{\mathbf{x}}_{nj}^*\right)'$, $j = 1, \dots, m$
obtain the cleaned response $\boldsymbol{y}^* \leftarrow \left(\omega_1 z_1^*, \dots, \omega_n z_n^*,\right)$
```