Stat Methods Appl manuscript No.
(will be inserted by the editor)

Robust variable selection with application to quality of life research

Andreas Alfons · Wolfgang E. Baaske · Peter Filzmoser · Wolfgang Mader · Roland Wieser

Received: date / Accepted: date

Abstract A large database containing socioeconomic data from 60 communities in Austria and Germany has been built, stemming from 18 000 citizens' responses to a survey, together with data from official statistical institutes about these communities. This paper describes a procedure for extracting a small set of explanatory variables to explain response variables such as the cognition of quality of life. For better interpretability, the set of explanatory variables needs to be very small and the dependencies among the selected variables need to be low. Due to possible inhomogeneities within the data set, it is further required that the solution is robust to outliers and deviating points. In order to achieve these goals, a robust model selection method, combined with a strategy to reduce the number of selected predictor variables to a necessary minimum, is developed. In addition, this context-sensitive method is applied to obtain responsible factors describing quality of life in communities.

Keywords Robustness · Model selection · Success factors · Quality of life

The research was supported by a grant of the Austrian Research Promotion Agency (FFG), Project Ref. No. 813000/10345.

A. Alfons · P. Filzmoser
Department of Statistics and Probability Theory, Vienna University of Technology
Wiedner Hauptstraße 8-10, 1040 Vienna, Austria
Tel.: +43 1 58801 10772
Fax: +43 1 58801 10798
E-mail: alfons@statistik.tuwien.ac.at

W.E. Baaske · R. Wieser
STUDIA-Schlierbach, Studienzentrum für internationale Analysen

W. Mader
SPES Academy

1 Introduction

The research project *ErfolgsVision* (English: *vision of success*) is a joint cooperation of the Austrian institutions SPES Academy (a regional developer), STUDIA-Schlierbach (an applied social researcher) and the Department of Statistics and Probability Theory at Vienna University of Technology. For this project, data from screening processes carried out by SPES in 60 communities in Austria and Germany during the period of 2000 to 2006 were used. In total, 18 748 questionnaires were collected, on average 312 per municipality. The survey was subject to individual adaptations towards the needs of the municipalities. It usually comprised about 250 questions, most of them multiple choice. In this project, we were interested in comparing the communities, therefore indicators referring to the questions were computed jointly from the questionnaires of each community. These data were merged with statistics on demography and economy. After removing observations with more than 50% and variables with more than 20% of missing values, a data matrix with 43 (out of 60) observations and 153 (out of 250) variables resulted. Some of the observations still included missing values (in one case for 20% of the variables), thus kNN imputation (Troyanskaya et al 2001) was used to obtain a complete data matrix.

Although the goal of the project was much broader, this paper is focused on finding the factors controlling quality of life. Since an easy interpretation of the results was a major objective, the number of explanatory variables should be limited to about 5 to at most 10. Moreover, the analysis needed to be robust against outliers and deviating data points because of possible inhomogeneities within the data set.

Various methods for model selection have been proposed to date. Here we are interested in robust approaches, as they are less sensitive to outliers in the data. Such methods have gained increasing attention in the literature (e.g., Ronchetti and Staudte 1994; Ronchetti et al 1997; Wisnowski et al 2003; Müller and Welsh 2005; Khan et al 2007a,b; McCann and Welsch 2007; Salibian-Barrera and Van Aelst 2008; Choi and Kiefer 2010; Riani and Atkinson 2010; Van Aelst et al 2010). However, robust variable selection is especially difficult if the number of observations is smaller than the number of variables. In that case it is no longer possible to directly apply robust regression methods (Maronna et al 2006) in order to select the most significant variables. On the other hand, various techniques for variable selection in high dimensions have been introduced, which are based on the non-robust least squares criterion (see, e.g., Hastie et al 2009; Varmuza and Filzmoser 2009). An example is least angle regression (LARS; Efron et al 2004), which selects the regressor variables in the order of their importance for predicting the response variable. LARS has been robustified in Khan et al (2007b) by two different approaches: the *plug-in* method and the *cleaning* method. In the plug-in method, the non-robust estimators mean, variance and correlation in classical LARS are replaced by robust counterparts. The idea of the cleaning method, on the other hand, is to shrink outliers and to apply classical LARS to the cleaned

data. Both methods use the so-called *winsorization* technique to estimate the correlations and shrink the outliers, respectively. Thus the influence of potential outliers on computing the sequence of predictors is reduced. Since the plug-in approach is computationally faster and more widely applicable, it is the basis of our algorithm for robust variable selection. In the following, the plug-in method will be referred to as RLARS. Khan et al (2007b) illustrated that the sequence of predictors returned by RLARS can be stabilized with the help of the bootstrap. The resulting procedure is called *bootstrapped* RLARS, for short B-RLARS.

A reduced set of the B-RLARS sequence of candidate predictors is then used for building a more refined regression model. For this purpose we suggest to use MM-regression (Yohai 1987; Maronna et al 2006). MM-estimators have many desirable properties. Most importantly, they combine a maximum breakdown point of 0.5 with high efficiency. Salibian-Barrera and Zamar (2002) further studied the distribution of MM-estimates using a robust bootstrap method. We apply MM-regression to filter out the non-significant variables at a certain significance level. Since in general the resulting number of the resulting variables is still too high for a reasonable interpretation, all possible subsets of size k are examined (see, e.g., Furnival and Wilson 1974; Miller 2002; Gatu and Kontoghiorghes 2006), which is sometimes referred to as k-subset regression. In our case, a robustified version of k-subset regression is applied by using the weights obtained from MM-regression. Thus strong dependencies among the regressor variables are eliminated and the smaller models are highly interpretable, which is required in the context of social sciences. This approach will therefore be called *context-sensitive* and can be considered a trade-off between quality of the model and interpretability.

The rest of this paper is organized as follows. In Section 2, we will describe the complete algorithm in more detail. Section 3 outlines how the procedure can be applied to obtain a small set of explanatory variables determining quality of life, and a simulation study is performed in Section 4. The final Section 5 concludes.

2 Context-sensitive model selection

Let $\boldsymbol{y} = (y_1, \ldots, y_n)^t$ be the response variable and $\boldsymbol{x}_1 = (x_{11}, \ldots, x_{n1})^t$, $\ldots$, $\boldsymbol{x}_p = (x_{1p}, \ldots, x_{np})^t$ the candidate predictors. Thus n denotes the number of observations and p the number of candidate predictors. Furthermore, let $J = \{1, \ldots, p\}$ be the set of indices referring to the candidate predictor variables. Our method aims to find a model for the response variable $\boldsymbol{y}$ that contains a very low number of predictors, at most $k \ll p$, in order to achieve high interpretability. Since the predictor variables should contain potentially new information, an additional requirement is that strong dependencies among the regressor variables should be avoided. These goals of easy-to-interpret models and low or only moderate dependencies between the predictors reflect the context-sensitivity of our method.

2.1 Description of the algorithm

For a start, the response variable $\boldsymbol{y}$ and the candidate predictors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ are robustly centered and scaled using median and MAD, according to

$$y_i^* = \frac{y_i - med(y_1, \ldots, y_n)}{MAD(y_1, \ldots, y_n)}, \qquad i = 1, \ldots, n \tag{1}$$

$$x_{ij}^* = \frac{x_{ij} - med(x_{1j}, \ldots, x_{nj})}{MAD(x_{1j}, \ldots, x_{nj})}, \qquad i = 1, \ldots, n,\ j = 1, \ldots, p. \tag{2}$$

Hence all predictor variables $\boldsymbol{x}_j^* = (x_{1j}^*, \ldots, x_{nj}^*)^t$, $j = 1, \ldots, p$, are on an equal scale. Our algorithm then proceeds in three steps. The first step seeks a drastic reduction of the number of candidate predictors such that the following steps become computationally feasible. For this purpose, B-RLARS (Khan et al 2007b) is applied to $\boldsymbol{y}^* = (y_1^*, \ldots, y_n^*)^t$ and $\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_p^*$ to find a sequence $(\boldsymbol{x}_j^*)_{j \in J_1}$, $J_1 \subset J$, of candidate predictors for $\boldsymbol{y}^*$ with $k < |J_1| \ll p$. Clearly, J_1 contains the indices of the $|J_1|$ most important predictor variables returned by B-LARS. In order to allow for an interpretation of the final model, $|J_1|$ should be in the range of 10 to 20.

In the second step, the covariates $\boldsymbol{x}_j^*$, $j \in J_1$, are entered as predictors for $\boldsymbol{y}^*$ in MM-regression (Yohai 1987; Maronna et al 2006). We apply MM-regression to filter out the non-significant variables. Let $J_2 \subseteq J_1$ be the set of indices of the significant variables at a given significance level α. The choice of α should not be too strict (we used $\alpha = 0.3$) in order not to exclude important variables. Note that this test is robust because it is based on robust estimates of the standard errors (Croux et al 2008). The second step thus concludes with fitting another MM-regression model to $\boldsymbol{y}^*$, using only the significant predictors $\boldsymbol{x}_j^*$, $j \in J_2$. Thus we consider the regression model

$$y_i^* = (\mathbf{x}_i^*)^t \boldsymbol{\beta} + e_i, \qquad i = 1, ..., n, \tag{3}$$

where $\mathbf{x}_i^*$ denotes the i-th observation of the predictor variables $\boldsymbol{x}_j^*$, $j \in J_2$, extended by 1 in the first component to account for the intercept. Furthermore, $\boldsymbol{\beta}$ is the vector of length $|J_2| + 1$ of the unknown regression coefficients, and e_i denotes the error terms, which are assumed to be i.i.d. random variables. MM-regression minimizes a function of the scaled residuals. Denoting the residuals by $r_i(\boldsymbol{\beta}) = y_i^* - (\mathbf{x}_i^*)^t \boldsymbol{\beta}$, MM-regression solves the problem

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} \rho\left(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}}\right), \tag{4}$$

where $\rho(r)$ is a bounded function, and $\hat{\sigma}$ is a robust scale estimator of the residuals, derived from a robust (but inefficient) S-estimator (for more details, see Maronna et al 2006). Differentiating (4) with respect to $\boldsymbol{\beta}$ yields

$$\sum_{i=1}^{n} \psi\left(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}}\right) \mathbf{x}_i^* = 0 \tag{5}$$

where $\psi = \rho'$. Using the notation

$$w_i = \frac{\psi(r_i(\boldsymbol{\beta})/\hat{\sigma})}{r_i(\boldsymbol{\beta})/\hat{\sigma}}, \qquad i = 1, ..., n, \tag{6}$$

allows (5) to be rewritten as

$$\sum_{i=1}^{n} w_i r_i(\boldsymbol{\beta})\, \mathbf{x}_i^* = 0. \tag{7}$$

Equation (7) is a weighted version of the normal equations. Hence the estimator can be considered a weighted least squares estimator with weights w_i from (6), which depend on the data. For an estimator to be robust, observations with large residuals should receive small weights. Thus the function ρ was chosen as the bisquare function (see Maronna et al 2006), which ensures that $\psi(r)$ is decreasing towards zero for increasing $|r|$. The resulting weights $\hat{w}_i$, $i = 1, \ldots, n$, for the MM-regression estimator $\hat{\boldsymbol{\beta}}$ will be used in the third step of the algorithm.

The third step is based on k-subset regression (see, e.g., Furnival and Wilson 1974; Miller 2002; Gatu and Kontoghiorghes 2006). Thus we want to find the best subset of maximum size k of the predictor variables that optimizes a criterion such as Mallows' C_p (Mallows 1973) or the BIC (Schwarz 1978). Although k-subset regression is not feasible even for moderate numbers of predictors, our method does not suffer from this problem since the number of predictors has been drastically reduced with B-RLARS in the first step and MM-regression in the second step. Another problem with k-subset regression is that it is not robust. However, a simple robustification is to use the weights computed in the second step during MM-regression, i.e., to enter the procedure with the response variable $\tilde{\boldsymbol{y}} = (\hat{w}_1 y_1^*, \ldots, \hat{w}_n y_n^*)^t$ and the candidate predictors $\tilde{\boldsymbol{x}}_j = (\hat{w}_1 x_{1j}^*, \ldots, \hat{w}_n x_{nj}^*)^t$, $j \in J_2$. Since the data are robustly standardized, multiplying the observations with the weights results in shrinking the outliers towards the main body of the data. This robustified version of k-subset regression yields the optimal subset $\{\boldsymbol{x}_j^* : j \in J_3\}$ with $J_3 \subseteq J_2, |J_3| \leq k$, of the set of candidate predictors $\{\boldsymbol{x}_j^* : j \in J_2\}$.

Instead of using the weights computed in the second step, other robust versions of k-subset regression might be considered. One example is fitting MM-regression models to all possible subsets of maximum size k and using m-fold cross-validation to estimate a robust prediction loss function, e.g., the root trimmed mean squared error of prediction (RTMSEP), for choosing the optimal submodel. In m-fold cross validation, the data are split randomly in m blocks of approximately equal size. Each block is left out once for fitting the model, and the left-out block is used as test data. Thus a prediction is obtained for each observation. Let $b(i)$ be the block to which observation $i = 1, \ldots, n$ belongs, then the prediction for y_i is denoted by $\hat{y}_i^{-b(i)}$. For a trimming factor

$0 \leq \gamma < 0.5$, the RTMSEP is defined as

$$\text{RTMSEP} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} r_{(i)}^2} \tag{8}$$

where $r_i = y_i - \hat{y}_i^{-b(i)}$, $i = 1, \ldots, n$, are the residuals using the predictions from cross-validation, $r_{(1)}^2 \leq \ldots \leq r_{(n)}^2$ are the sorted squared residuals, and $N = n - \lfloor n\gamma \rfloor$ (here $\lfloor a \rfloor$ denotes the integer part of a). Whereas such procedures are certainly more robust than the simple weighted approach, they are computationally expensive even for small problems. On the other hand, using the weights computed in the second step of the procedure results in a cleaned data set, thus reducing the influence of atypical observations in both fitting the submodels and computing classical criteria for deciding on the best submodel. Even though the weights might not be optimal for each submodel, this approach is a reasonable compromise between computational complexity and robustness. It is fast for small problems and worked very well in our studies (see the example in Section 3).

2.2 Summary of the algorithm

The response variable and all candidate predictor variables are robustly centered and scaled using median and MAD. The resulting response variable is denoted by $\boldsymbol{y}^* = (y_1^*, \ldots, y_n^*)^t$, and the resulting candidate predictors by $\boldsymbol{x}_1^* = (x_{11}^*, \ldots, x_{n1}^*)^t, \ldots, \boldsymbol{x}_p^* = (x_{1p}^*, \ldots, x_{np}^*)^t$. Let $J = \{1, \ldots, p\}$ be the set of indices for the candidate predictors, and $k \ll p$ the desired maximum number of predictors for the model. Then the algorithm can be summarized as follows:

1. Perform B-RLARS on $\boldsymbol{y}^*$ and $\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_p^*$ to compute a sequence $(\boldsymbol{x}_j^*)_{j \in J_1}$, $J_1 \subset J$, of candidate predictors with $k < |J_1| \ll p$.
2. Use $\boldsymbol{x}_j^*$, $j \in J_1$, as predictors for $\boldsymbol{y}^*$ in MM-regression. Let $J_2 \subseteq J_1$ be the set of indices of the significant variables at a given significance level α. Fit another MM-regression model to $\boldsymbol{y}^*$ with only the significant predictors $\boldsymbol{x}_j^*$, $j \in J_2$, and let $\hat{w}_1, \ldots, \hat{w}_n$ denote the resulting weights for the observations.
3. Apply k-subset regression with the response variable $\tilde{\boldsymbol{y}} = (\hat{w}_1 y_1^*, \ldots, \hat{w}_n y_n^*)^t$ and the candidate predictors $\tilde{\boldsymbol{x}}_j = (\hat{w}_1 x_{1j}^*, \ldots, \hat{w}_n x_{nj}^*)^t$, $j \in J_2$. This robustified version of k-subset regression yields the optimal subset $\{\boldsymbol{x}_j^* : j \in J_3\}$ with $J_3 \subseteq J_2, |J_3| \leq k$, of the set of candidate predictors $\{\boldsymbol{x}_j^* : j \in J_2\}$.

A more visual summary of the algorithm is given by the following diagram:

$$J \quad \overset{\text{B-RLARS}}{\longrightarrow} \quad J_1 \quad \overset{\text{MM-regression}}{\longrightarrow} \quad J_2 \quad \overset{k\text{-subset regression}}{\longrightarrow} \quad J_3$$

2.3 Diagnostics

The elimination of high dependencies among the predictor variables is a major demand for our context-sensitive method. In the social sciences, such a model has potential for an interesting interpretation. Correlated predictor variables, on the other hand, are likely to describe more or less the same factors, which are just expressed with different variables in the data set. The resulting model will not be as interesting with respect to interpretation, even if it has a high prediction ability of the response variable. Hence a graphical tool to check whether the procedure succeeded in fulfilling this demand would be useful. A dendrogram (e.g., Everitt and Dunn 2001) based on robust correlations seems suitable for this purpose.

Since the number of candidate predictors is in general too large for an informative plot, only the variables $\boldsymbol{x}_j$, $j \in J_1$, from the initial B-RLARS sequence will be used. The correlation matrix of this reduced set of candidate predictors can be estimated with a high-breakdown estimator such as the *minimum covariance determinant* (MCD; Rousseeuw and Van Driessen 1999) or the *orthogonalized Gnanadesikan-Kettenring* estimator (OGK; Maronna and Zamar 2002). Note that the correlations used here do not need not come from an affine equivariant or orthogonal equivariant method, the Spearmann or Kendall correlation could also be used (for their robustness properties, see Croux and Dehon 2010). Let $\boldsymbol{R} = (r_{ij})_{i,j \in J_1}$ denote such a robust estimate of the correlation matrix. Then the dissimilarity matrix $\boldsymbol{D} = (d_{ij})_{i,j \in J_1}$ given by

$$d_{ij} = 1 - |r_{ij}|, \qquad i, j \in J_1, \tag{9}$$

is used for clustering the variables. *Complete linkage* clustering (e.g., Everitt and Dunn 2001) is well suited for our purposes, as the dissimilarity measure is based on robust correlations. In this method, the dissimilarity of two clusters A and B is defined as

$$d(A, B) = \max_{\boldsymbol{x}_i \in A, \boldsymbol{x}_j \in B} d_{ij}. \tag{10}$$

Using (9), this can be written as

$$d(A, B) = 1 - \min_{\boldsymbol{x}_i \in A, \boldsymbol{x}_j \in B} |r_{ij}|. \tag{11}$$

In each step, the two clusters with minimum dissimilarity are merged. Thus complete linkage clustering in our case yields that variables with low correlations will not belong to the same cluster if an appropriate cut-off point is chosen. Hence the resulting dendrogram is a convenient way of exploring the robust correlation structures among the candidate predictor variables. If the selected variables belong to different clusters, then the procedure performed well in the context-sensitive sense. Such a dendrogram may also reveal potential problems due to strong correlations among all predictor variables. In this case, it would probably be difficult to decide on which variables should be eliminated for a highly interpretable model.

Table 1 Explanation of important variables.

Variable	Explanation
qualityLife	quality of life
agriculture	state of local agriculture
beauty	beauty of the community
contrFarmers	contribution of local farmers to quality of life
futureComm	future development of the community
impOrganic	importance of organic products
impTrad	importance of traditional festivities
interesting	interestingness of the community
medCare	state of medical care
merchAssort	assortment of local merchants
merchComm	contribution of local merchants to the development of the community
parish	state of local parish
percAdolesc	percentage of adolescents
publicServ	state of public services
eduProTraining	educational and professional training opportunities
view	state of the community's view

2.4 Implementation

An implementation of our algorithm in the statistical environment R (R Development Core Team 2010) and detailed documentation can be downloaded from `http://www.statistik.tuwien.ac.at/public/filz/programs.html`. The required R code for B-RLARS by Khan et al (2007b) can be obtained from `http://users.ugent.be/~svaelst/software/RLARS.html`. In addition, the R packages `robustbase` (Rousseeuw et al 2009) and `leaps` (Lumley and Miller 2009), which are available on CRAN (the Comprehensive R Archive Network, `http://cran.r-project.org`), need to be installed.

3 Example: driving factors behind quality of life

In this section, we will attempt to find the driving factors behind quality of life in communities, using the data collected by SPES (see Section 1 for a general description of the data). Table 1 contains explanations for the most important variables. In order to ensure an easy-to-interpret model, the response variable *qualityLife* should be explained by at most 10 predictors. Note that some variables, which are too discontinuous or clearly redundant in the context of quality of life, are removed from the data set, resulting in 138 remaining candidate predictors. Hence all variables are continuous, which is important for applying the developed robust method.

Furthermore, we will compare our robust context-sensitive method, in the following referred to as RCS, with B-RLARS.

3.1 Results

RCS is carried out with parameter settings as described in the following. As mentioned above, the maximum number of variables in the final model is set to $k = 10$. In the initial B-RLARS step, 15 variables are sequenced with 50 bootstrap repetitions. These candidate predictors are then filtered at significance level $\alpha = 0.3$ in MM-regression. This unusually high significance level will prevent the exclusion of potentially important variables. For deciding on the optimal submodel in the robustified version of k-subset regression, the BIC is used as criterion. With these parameters, RCS returns the following six predictors: *agriculture*, *medCare*, *merchAssort*, *eduProTraining*, *beauty* and *parish* (see Table 1).

In addition to the simple weighted k-subset regression in the third step of RCS, we also apply a more sophisticated robust version for comparison. In this version, we fit MM-regression models to the subsets and use fivefold cross-validation to estimate the root trimmed mean squared error of prediction (RTMSEP) with 20% trimming, see (8). Fivefold cross-validation seems to be a reasonable choice given the number of observations in the data set. Furthermore, the choice of the trimming proportion is based on the weights returned by the MM-regression in the second step, which indicate some outliers. With a lower value, these outliers may still influence the RTMSEP, whereas a higher value may result in some bias. The submodel with the lowest RTMSEP is then chosen as the optimal submodel. While this procedure yields the same six variables as the simple weighted approach, it is computationally much more expensive.

In order to compare RCS with B-RLARS, we start with the B-RLARS sequence of length 15 that we computed in the first step of RCS. Then we proceed as in the examples in Section 6 of Khan et al (2007b) to obtain the final B-RLARS model. There it is suggested to start with the first variable and to increase the number of variables along the sequence, while fitting a robust regression model in each step. For each model, the robust R^2 measure

$$R^2_{rob} = 1 - \left(\frac{med(|y_1 - \hat{y}_1|, \ldots, |y_n - \hat{y}_n|)}{MAD(y_1, \ldots, y_n)} \right)^2, \tag{12}$$

is computed, where y_i, $i = 1, \ldots, n$, are the observed values of the response variable and $\hat{y}_i$, $i = 1, \ldots, n$, are the fitted values (see Rousseeuw and Leroy 1987). Finally, these robust R^2 values are plotted against the model size to obtain a *learning curve* (c.f. Croux et al 2003). Note that the robust R^2 is not always monotonically increasing with the number of variables since algorithms for robust regression yield only approximate solutions. Keeping in mind that the number of predictors should be at most 10, the learning curve in Figure 1 (left) suggests using the first 8 variables of the sequence: *contrFarmers*, *agriculture*, *medCare*, *merchComm*, *impOrganic*, *merchAssort*, *percAdolesc* and *interesting* (see Table 1). These variables are further examined by fitting MM-regression models to all possible subsets. Deciding on the best subset is done by minimizing the RTMSEP with 20% trimming, which is

Table 2 MM-regression results for the RCS model for quality of life.

	Estimate	Standard error	t-Value	p-Value
(Intercept)	-2.302	11.227	-0.205	0.839
agriculture	0.251	0.053	4.713	$3.6 \cdot 10^{-5}$
medCare	0.076	0.023	3.228	0.003
merchAssort	0.177	0.064	2.751	0.009
eduProTraining	0.117	0.026	4.450	$8.0 \cdot 10^{-5}$
beauty	0.292	0.113	2.588	0.014
parish	0.216	0.035	6.226	$3.5 \cdot 10^{-7}$
Robust residual standard error: 1.705				

Table 3 MM-regression results for the B-RLARS model for quality of life.

	Estimate	Standard error	t-Value	p-Value
(Intercept)	8.795	7.079	1.242	0.221
agriculture	0.337	0.064	5.278	$5.2 \cdot 10^{-6}$
merchAssort	0.277	0.065	4.297	$1.1 \cdot 10^{-4}$
interesting	0.409	0.082	5.009	$1.2 \cdot 10^{-5}$
Robust residual standard error: 2.419				

estimated using fivefold cross-validation. The final model resulting from this procedure contains the predictors *agriculture*, *merchAssort* and *interesting*.

Tables 2 and 3 show the results of MM-regression with the predictor variables selected by RCS and B-RLARS, respectively. In both models, the included variables are highly significant. Containing only three predictor variables, the B-RLARS model is on the one hand somewhat simpler than the RCS model, which consists of six predictors. Two of the three variables selected by B-RLARS are also selected by RCS (*agriculture* and *merchAssort*). On the other hand, the robust residual standard error indicates that the B-RLARS model might be too simple. The RCS model is a better fit due to the much lower robust residual standard error.

However, in order to decide on which model is preferable, it is necessary to estimate the prediction quality of the models. For this purpose, repeated fivefold cross-validation with 1,000 repetitions is applied. In each repetition, the RTMSEP with 20% trimming is estimated. Figure 1 (right) displays the resulting density curves for the RCS model, the final B-RLARS model (B-RLARS-3) and the B-RLARS model with the first 8 variables as suggested by the learning curve (B-RLARS-8). It is clearly visible from this plot that the average RTMSEP is significantly smaller for RCS than for the other two models. Even though the variance of the RTMSEP is slightly larger for RCS than for B-RLARS-3, it is comparable for the two methods. Thus the RCS model performs much better than the two B-RLARS models, while the B-RLARS-8 model clearly leads to the worst prediction performance.

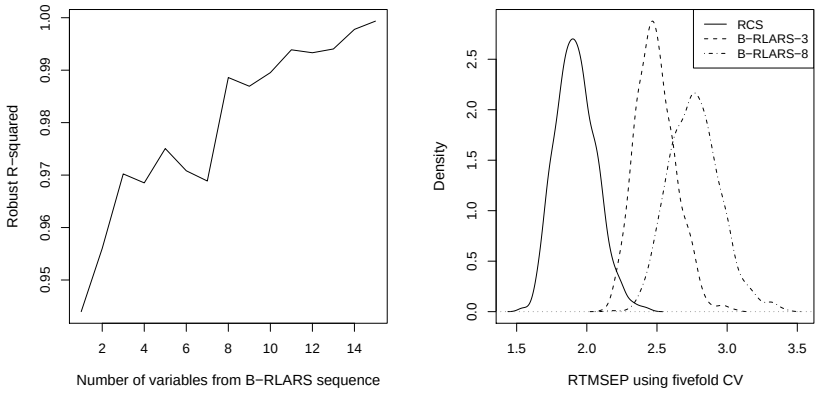

Fig. 1 Learning curve for the B-RLARS sequence (left). Densities of the RTMSEP for the RCS model, the final B-RLARS model (B-RLARS-3) and the B-RLARS model with the first 8 variables as suggested by the learning curve (B-RLARS-8), estimated with repeated fivefold cross-validation (right).

One of the main requirements concerning context-sensitivity was that the resulting model should be simple. Nevertheless, while succeeding in finding a few important predictor variables, the B-RLARS model turns out to be too simple. By only moving along the computed sequence of candidate predictors for finding the optimal size of the model, variables such as *medCare* and *eduProTraining* were completely neglected, even though they are clearly very important in the context of quality of life. Since RCS manages to include these variables in the selected model, the key step for context-sensitivity in the RCS procedure may be selecting the variables of the initial B-RLARS sequence at a certain significance level in MM-regression.

Another main requirement was that the dependencies among the selected variables should be rather low. Therefore, a dendrogram is constructed according to Section 2.3 and shown in Figure 2. It includes the 15 most important candidate predictors for quality of life, which were sequenced with B-RLARS in the first step of our context-sensitive procedure. The robust correlations for the dendrogram were computed with the reweighted MCD. The trimming parameter for the size of the subsets was thereby set to 75%. Furthermore, the finite sample correction factor and the asymptotic consistency factor were used. The dendrogram shows that RCS was able to fulfill this demand of low variable dependencies. In addition, every group in the dendrogram is represented in the RCS model, but not in the B-RLARS model.

The results seem to be significant in terms of theoretical concepts for quality of life assessments. Our selection procedure definitely moves beyond producing inconsistent lists of indicators, it creates a set of meaningful empirical measures. In quality of life research (e.g., Diener et al 1999), individualistic and subjective indicators prevail, but recent concepts combine them with features of the external world. The model of Renwick et al (1994), followed by

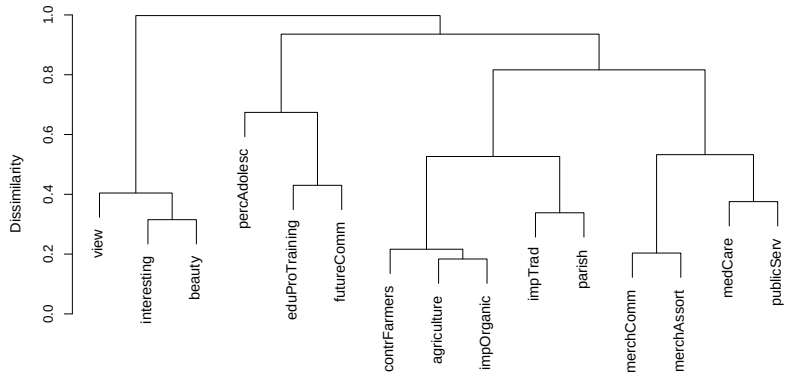

Fig. 2 Dendrogram (based on robust correlations) of the initial B-RLARS sequence of candidate predictors for quality of life.

Tichbon and Newton (2002), allows subjective states (*being*—e.g., health, nutrition, beliefs, values), as well as objective states (*belonging*—e.g., services, activities, leisure) and development (*becoming*—e.g., acquisition of skills and knowledge). Meaningful variables of all three types are included in the empirical results presented in this article (with some of the variables loading on different types): *medCare* and *merchAssort* are *being*-indicators, *agriculture* and *beauty* are *belonging*-indicators, while *parish*, *interesting* and *eduProTraining* signify development (*becoming*). The studies of this project are insofar unique, as they combine internal and external world features on a solid data base with appropriate analysis techniques. We recommend to incorporate the results into the design of agricultural policies. Municipalities often underestimate the role of the "lagging-behind" sector agriculture, whereas our analysis shows that the state of local agriculture constitutes a significant share of quality of life. On a world-wide level, producing quality of life as an external effect within the proximity may cause agriculture to be be respected and handled differently from a mere producer of tradable commodities (Baaske et al 2009).

3.2 CPU times

The computation times presented in this section are average times over 50 runs, carried out on a machine with an Intel® Core^{TM}2 Quad 2.66GHz processor and 8GB main memory. Keep in mind that the computations were carried out with R (and thus only one of the four available processors was effectively used), and that the data set consists of 43 observations and 138 candidate predictor variables. With the parameter settings as described in the beginning of Section 3.1, RCS completed after 20.61 seconds. The running time was thereby dominated by computing the initial B-RLARS sequence, which took 20.54 sec-

onds. This example indicates that RCS is still feasible whenever computing the initial B-RLARS sequence is feasible.

For finding the final B-RLARS model, the learning curve had to be inspected graphically to find the optimal number of predictors. Afterwards, all subsets of the reduced sequence were examined using MM-regression and five-fold cross-validation, which was very time-consuming for such a small problem. Since RCS uses the simple weighted version of robust k-subset regression and does not require manual interaction, obtaining the RCS model was much faster than obtaining the final B-RLARS model.

4 Simulations

For further investigation of the proposed RCS procedure, simulations are carried out using a simulation setting similar to that from Khan et al (2007b). With k latent independent standard normal variables $l_1, \ldots, l_k$ and an independent standard normal variable e, a linear model is constructed as

$$y := l_1 + \ldots + l_k + \sigma e, \tag{13}$$

where σ is chosen so that the signal-to-noise ratio is 5, i.e.,

$$\sqrt{\mathrm{var}(l_1 + \ldots + l_k)/\mathrm{var}(\sigma e)} = \sqrt{k}/\sigma = 5. \tag{14}$$

Using independent standard normal variables $e_1, \ldots, e_p$, a set of p candidate predictors is then constructed as

$$\begin{aligned}
x_1 &:= l_1 + \tau e_1, \\
x_2 &:= l_1 + \tau e_2, \\
x_3 &:= l_1 + \tau e_3, \\
&\vdots \\
x_{3k-2} &:= l_k + \tau e_{3k-2}, \\
x_{3k-1} &:= l_k + \tau e_{3k-1}, \\
x_{3k} &:= l_k + \tau e_{3k}, \\
x_{3k+1} &:= l_1 + \delta e_{3k+1}, \\
x_{3k+2} &:= l_1 + \delta e_{3k+2}, \\
&\vdots \\
x_{5k-1} &:= l_k + \delta e_{5k-1}, \\
x_{5k} &:= l_k + \delta e_{5k}, \\
x_i &:= e_i, \qquad i = 5k+1, \ldots, p,
\end{aligned} \tag{15}$$

where $\tau = 0.2$ and $\delta = 5$ so that $x_1, \ldots, x_{3k}$ form k groups of low-noise perturbations of the latent variables, $x_{3k+1}, \ldots, x_{5k}$ are noise covariates that are correlated with the latent variables, and $x_{5k+1}, \ldots, x_p$ are independent noise covariates.

Regarding contamination, the following scenarios are investigated (similar to a subset of the scenarios investigated in Khan et al 2007b), where ε denotes the fraction of outliers in the data:

1. No contamination.
2. Contamination in y given by $e \sim (1-\varepsilon)N(0,1) + \varepsilon N(0,1)/U(0,1)$.
3. Same as 2., but contaminated observations contain outliers in $x_1, \ldots, x_p$ coming from $N(5,1)$.

Note that in the last scenario, the contamination is not more extreme because the outliers in the data for which the proposed method has been designed (see Section 1) are moderate as well.

In the simulation experiments in Khan et al (2007b), B-RLARS is compared to other methods using *recall curves*, i.e., the average numbers of target variables included in the first m sequenced variables are plotted, with m varying within a certain range. However, our procedure does not produce a sequence of predictor variables, instead it is designed to obtain a final model from an initial sequence of candidate predictors. Hence a comparison with B-RLARS using recall curves is not meaningful.

Moreover, one requirement for our procedure is that strong correlations between variables should be avoided. For each latent variable, a group of low-noise perturbations is thus defined in (15). Variables in the same group are highly correlated, while the correlations between variables from different groups are low. The procedure is successful in the context-sensitive sense if the final model contains exactly one predictor variable from each of these groups. Nevertheless, the success of the procedure of course also depends on the initial B-RLARS sequence. If no variables of one group exist in the initial sequence, the final model cannot contain a variable of this group either.

In the simulations, $k = 5$ latent variables are used to construct the linear model for the response as in (13) and $p = 100$ candidate predictors as in (15). Concerning the number of observations, two situations are investigated: $n = 50$ ($n < p$, high-dimensional data) and $n = 150$ ($n > p$). In both cases, the contamination level is set to $\varepsilon = 0.1$. The number of predictors in the final RCS model is limited to the number of latent variables $k = 5$. For the remaining parameters of RCS, the same settings as in the example from Section 3 are used, i.e., 15 variables are sequenced in the initial B-RLARS step with 50 bootstrap repetitions, the significance level for MM-regression in the second step is set to $\alpha = 0.3$, and the BIC used as criterion for k-subset regression in the third step. In addition, the simulations are performed with the R package `simFrame` (Alfons et al 2009; Alfons 2010), which is a general framework for statistical simulation.

The results from 100 simulation runs are presented in Table 4. Averages of certain quantities of interest are thereby computed. The final RCS model is evaluated by the number of groups of low-noise perturbations that are represented by exactly one variable (#target), the number of noise variables (#noise), and the total number of variables (#total). Ideally, the final model would consist of $k = 5$ target predictors—exactly one from each group and

Table 4 Average results from 100 simulation runs with contamination level $\varepsilon = 0.1$. For RCS, the number of target groups represented by exactly one variable (#target), the number of noise variables (#noise), and the total number of variables (#total) are shown. For the first k variables of B-RLARS, the number of target groups represented by exactly one variable (#target) and the number of noise variables (#noise) are displayed. The full B-RLARS sequence is evaluated using the number of represented groups (#groups) and the number of noise variables (#noise).

		RCS			First k of B-RLARS		B-RLARS	
n	Scenario	#target	#noise	#total	#target	#noise	#groups	#noise
50	1	4.84	0.06	4.90	3.91	0.43	4.99	5.25
	2	4.79	0.09	4.88	3.82	0.52	5	5.58
	3	4.28	0.68	4.96	3.48	1.02	4.82	7.47
150	1	5	0	5	3.86	0	5	1.19
	2	5	0	5	4.12	0.02	5	1.81
	3	4.89	0.11	5	3.91	0.49	5	4.84

no noise variables. Since the success of the procedure depends on the initial B-RLARS step, the initial sequence from this step is evaluated as well. As discussed in the example in Section 3, the first part of the sequence may not contain some important predictors. In order to further investigate this issue, the number of groups that are represented by exactly one variable (#target) and the number of noise variables (#noise) are computed for the first k variables in the initial B-RLARS sequence as well. In the complete B-RLARS sequence, as many of the low-noise perturbations as possible should be included. It is essential that all groups occur in the sequence so that it is possible to extract one variable for each group in the remaining steps of the procedure. Therefore, the initial B-RLARS sequence is evaluated using the number of represented groups (#groups) and the number of noise variables (#noise). The initial B-RLARS step performs well in this setting if all variables from the groups of low-noise perturbations and no additional noise variables are sequenced.

The simulation results from Table 4 indicate that the RCS procedure performs very well. In particular in the case of $n > p$, the results are excellent. Only in some instances for the scenario with contamination in the candidate predictors, the final model does not contain exactly one variable from each group of low-noise perturbations. In these instances, the final model also contains one noise variable, which may be due to the considerably higher number of noise variables in the initial B-RLARS sequence compared to the other scenarios. In the case of $n < p$ (low sample size, high-dimensional data), variable selection is much more difficult, which is also reflected in the simulation results. For all scenarios, the number of noise variables in the initial B-RLARS sequence is much higher than in the case of $n > p$. The RCS procedure still gives excellent results if the data are not contaminated or if contamination is only present in the response. Merely in some cases, the final model consists of less than $k = 5$ predictors or contains a noise variable. But even if the candidate predictors are contaminated as well, the results are very reasonable

considering that on average about half of the variables in the initial B-RLARS sequence are noise variables.

Furthermore, the results from the simulations show that the first parts of the B-RLARS sequence may not contain some important variables for data of a certain structure. In all investigated scenarios, the first k variables in the initial B-RLARS sequence often contain more than one variable from the same group of low-noise perturbations, and in some scenarios even noise variables frequently occur.

5 Conclusions and discussion

Motivated by a practical application, we developed a strategy for finding a linear regression model that includes only a necessary minimum of key predictor variables to describe the response. The number of explanatory variables thereby was supposed to be smaller than a given boundary, each of them should contain potentially new information, and the resulting model should be highly interpretable. Moreover, the variable selection procedure needed to be robust with respect to possible data inhomogeneities and outliers. The difficulty with these requirements was that the underlying data set is high-dimensional, with much more variables than observations.

Several methods for model selection in high dimensions are available to date, but only a few proposals for robust model selection have been made due to the much higher request of computation time. Our algorithm is based on bootstrapped robust least angle regression (B-RLARS; Khan et al 2007b), which we apply to find an initial sequence of explanatory variables. In addition to being robust to atypical observations, B-RLARS yields a stable sequence of predictors because of the bootstrap procedure, it is fast to compute, and R code (R Development Core Team 2010) is freely available. Different strategies for further reducing the initial sequence of predictor variables are possible. Since our aim is to extract a small set of highly informative explanatory variables, filtering out the non-significant variables with MM-regression (Yohai 1987; Maronna et al 2006) seems a suitable approach. MM-regression is used because it is both highly efficient and highly robust. Then all subsets of a given maximum size k of the set of significant variables can be examined to find the optimal regression model. However, using robust regression and resampling methods for this purpose is computationally expensive. Therefore, we suggest using k-subset regression based on least squares (e.g., Furnival and Wilson 1974; Miller 2002; Gatu and Kontoghiorghes 2006), which is robustified by using the weights obtained from another MM-regression model with only the significant explanatory variables. This is a simplification because the weights obtained from MM-regression on the significant variables might not be appropriate for a subset of these variables. For this reason, an alternative procedure based on the root trimmed mean squared error of prediction (RTMSEP) has been proposed as well, which nevertheless is computationally much more demanding. Note that also other procedures for robust variable

selection are possible, such as the forward search strategy (see Atkinson and Riani 2002).

In the example of extracting a small set of explanatory variables for quality of life, the suggested strategy succeeded in finding an easy-to-interpret model containing only predictors with potentially new information. The latter was confirmed by a cluster analysis based on robust correlations (see Figure 2). Moreover, the resulting model is an excellent fit and performs well with respect to prediction. Simulation results were presented as further indication of the excellent performance of the proposed procedure. Last but not least, our procedure also gave meaningful answers to other questions and hypotheses related to the project.

A principal question is whether robust methods are really required for a data set at hand. Usually, inspecting high-dimensional data for possible inhomogeneities or outliers is difficult. For our data set, we used the outlier detection method by Filzmoser et al (2008), which identified some clearly outlying observations. In the example for quality of life, the weights obtained by MM-regression with the reduced set of predictor variables indicated that outliers still exist in the much lower-dimensional subset of the data. In any case, even if only minor contamination is present, robust model selection can yield more stable results, as it is less sensitive to small changes in the (high-dimensional) data.

Acknowledgements The authors are grateful to the referees for helpful comments and suggestions.

References

Alfons A (2010) `simFrame`: Simulation framework. R package version 0.3.4

Alfons A, Templ M, Filzmoser P (2009) `simFrame`: An object-oriented framework for statistical simulation. Research Report CS-2009-1, Department of Statistics and Probability Theory, Vienna University of Technology

Atkinson A, Riani M (2002) Forward search added-variable t-tests and the effect of masked outliers on model selection. Biometrika 89(4):939–946

Baaske W, Filzmoser P, Mader W, Wieser R (2009) Agriculture as a success factor for municipalities. In: Jahrbuch der Österreichischen Gesellschaft für Agrarökonomie (ÖGA), vol 18, Facultas Verlag, Vienna, pp 21–30, ISBN 978-3-7089-0432-3

Choi H, Kiefer N (2010) Improving robust model selection tests for dynamic models. Econ J 13(2):177–204

Croux C, Dehon C (2010) Influence functions of the Spearman and Kendall correlation measures. Stat Methods Appl DOI 10.1007/s10260-010-0142-z, to appear

Croux C, Filzmoser P, Pison G, Rousseeuw P (2003) Fitting multiplicative models by robust alternating regressions. Stat Comput 13(1):23–36

Croux C, Dhaene G, Hoorelbeke D (2008) Robust standard errors for robust estimators. Discussion Papers Series 03.16, KU Leuven

Diener E, Suh E, Lucas R, Smith H (1999) Subjective well-being: Three decades of progress. Psychol Bull 125(2):276–302

Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. Ann Stat 32(2):407–499

Everitt B, Dunn G (2001) Applied Multivariate Data Analysis, 2nd edn. Arnold, London, ISBN 0-340-54529-1

Filzmoser P, Maronna R, Werner M (2008) Outlier identification in high dimensions. Comput Stat Data Anal 52(3):1694–1711
Furnival G, Wilson R (1974) Regression by leaps and bounds. Technometrics 16(4):499–511
Gatu C, Kontoghiorghes E (2006) Branch-and-bound algorithms for computing the best-subset regression models. J Comput Graph Stat 15(1):139–156
Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning, 2nd edn. Springer, New York, ISBN 978-0-387-84857-0
Khan J, Van Aelst S, Zamar R (2007a) Building a robust linear model with forward selection and stepwise procedures. Comput Stat Data Anal 52(1):239–248
Khan J, Van Aelst S, Zamar R (2007b) Robust linear model selection based on least angle regression. J Am Stat Assoc 102(480):1289–1299
Lumley T, Miller A (2009) `leaps`: regression subset selection. R package version 2.9
Mallows C (1973) Some comments on C_p. Technometrics 15(4):661–675
Maronna R, Zamar R (2002) Robust estimates of location and dispersion for high-dimensional datasets. Technometrics 44(4):307–317
Maronna R, Martin D, Yohai V (2006) Robust Statistics. Wiley, Chichester, ISBN 978-0-470-01092-1
McCann L, Welsch R (2007) Robust variable selection using least angle regression and elemental set sampling. Comput Stat Data Anal 52(1):249–257
Miller A (2002) Subset Selection in Regression, 2nd edn. Chapman & Hall/CRC, Boca Raton, ISBN 1-58488-171-2
Müller S, Welsh A (2005) Outlier robust model selection in linear regression. J Am Stat Assoc 100(472):1297–1310
R Development Core Team (2010) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL `http://www.R-project.org`, ISBN 3-900051-07-0
Renwick R, Brown I, Raphael D (1994) Quality of life: Linking conceptual approach to service provision. J Dev Disabil 3(2):32–44
Riani M, Atkinson A (2010) Robust model selection with flexible trimming. Comput Stat Data Anal 54(12):3300–3312
Ronchetti E, Staudte R (1994) A robust version of Mallows's C_p. J Am Stat Assoc 89(426):550–559
Ronchetti E, Field C, Blanchard W (1997) Robust linear model selection by cross-validation. J Am Stat Assoc 92(439):1017–1023
Rousseeuw P, Leroy A (1987) Robust Regression and Outlier Detection. Wiley, New York, ISBN 0-471-48855-0
Rousseeuw P, Van Driessen K (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics 41(3):212–223
Rousseeuw P, Croux C, Todorov V, Ruckstuhl A, Salibian-Barrera M, Verbeke T, Maechler M (2009) `robustbase`: Basic Robust Statistics. R package version 0.5-0-1
Salibian-Barrera M, Van Aelst S (2008) Robust model selection using fast and robust bootstrap. Comput Stat Data Anal 52(12):5121–5135
Salibian-Barrera M, Zamar R (2002) Bootstrapping robust estimates of regression. Ann Stat 30(2):556–582
Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464
Tichbon C, Newton P (2002) Life is do-able: Quality of life development in a supportive small group setting. Occasional Paper Series 2, Mental Health Foundation of New Zealand
Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman R (2001) Missing value estimation methods for DNA microarrays. Bioinformatics 17(6):520–525
Van Aelst S, Welsch R, Zamar R, eds (2010) Special issue on variable selection and robust procedures. Comput Stat Data Anal 54(12)
Varmuza K, Filzmoser P (2009) Introduction to Multivariate Statistical Analysis in Chemometrics. CRC Press, Boca Raton, ISBN 978-0-470-98581-6
Wisnowski J, Simpson J, Montgomery D, Runger G (2003) Resampling methods for variable selection in robust regression. Comput Stat Data Anal 43(3):341–355
Yohai V (1987) High breakdown-point and high efficiency robust estimates for regression. Ann Stat 15(20):642–656