A context-sensitive method for robust model selection with application to analyzing success factors of communities

A. Alfons, W.E. Baaske, P. Filzmoser, W. Mader, and R. Wieser

Forschungsbericht CS-2008-6
Dezember 2008

Kontakt: P.Filzmoser@tuwien.ac.at
A context-sensitive method for robust model selection with application to analyzing success factors of communities

A. Alfons\textsuperscript{a}, W.E. Baaske\textsuperscript{b}, P. Filzmoser\textsuperscript{*,a}, W. Mader\textsuperscript{c}, R. Wieser\textsuperscript{a}

\textsuperscript{a}Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstraße 8-10, 1040 Vienna, Austria
\textsuperscript{b}STUDIA-Schlierbach, Studienzentrum für internationale Analysen, Panoramaweg 1, 4553 Schlierbach, Austria
\textsuperscript{c}SPES Academy, Panoramaweg 1, 4553 Schlierbach, Austria

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 like 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.

Key words: robustness, model selection, context-sensitivity, success factors, quality of life

1. Introduction

The research project ErfolgsVision (engl. 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. These processes established innovative ideas, alliances and problem solutions in the municipalities. For many of them it was the first time that broad levels of the population were actively involved in local development. The survey was subject to individual adaptations towards the needs of the municipalities. It usually comprised about 250 questions, most of them multiple choice. 134 questions were posed identically in 30 or more municipalities and yield comparable results. Since the screening process was supported by the municipality and of major interest to the community, the data obtained from the questionnaires can be assumed to be of high reliability. The screening process resulted in evaluations of each municipality.

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. Table I contains explanations for the most important variables. Since some of the
observations still included missing values (in one case for 20% of the variables), we used kNN-imputation (Troyanskaya et al., 2001) to obtain a complete data matrix.

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

Table 1: Explanation of important variables.

The goal of the research project was to verify hypotheses and to answer questions of SPES (e.g., What makes communities successful?, What explains quality of life?, What factors encourage optimism?, Sector thinking in the communities disturbs the development of a positive quality of life.). From these questions and hypotheses, we derived target variables. In order to answer the questions, and to confirm or redraw the hypotheses, an appropriate statistical method to select a small number of meaningful variables that are able to explain the target variable was requested. Since the interpretation of the factors controlling the targets was a major goal, the number of explanatory variables was limited to about 5 to at most 10, depending on the context. Moreover, the analysis needed to be robust against outliers and deviating data points because of possible inhomogeneities within the data set.

Classical model selection methods are often based on optimizing a certain criterion, e.g., the AIC [Akaike, 1970], Mallows’ $C_p$ [Mallows, 1973] or the BIC [Schwarz, 1978]. Another approach is to minimize a prediction loss function, which can be estimated by resampling methods like cross-validation [Shao, 1993] or the bootstrap [Shao, 1996]. It is well known that even a small proportion of outliers may have a severe influence on these methods. Therefore, robust approaches towards variable selection have gained increasing attention in the literature (e.g., Ronchetti and Staudte, 1994, Ronchetti et al., 1997, Wisnowski et al., 2003, Miller and Welsh, 2005, Khan et al., 2007a, McCann and Welsch, 2007, Salibian-Barrera and van Aelst, 2008).

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., 2001, Varmuza and Filzmoser, 2009). An example is Lasso regression (Tibshirani, 1996), which is in fact a modification of Ridge regression (Hoerl and Kennard, 1970) by using an $L_1$ penalty for the size of the regression coefficients. Depending on the size constraint, some of the regression coefficients are shrunk towards zero. Least angle regression (LARS, Efron et al., 2004) is closely related to Lasso regression, as it computes the complete Lasso solution simultaneously for all values of the shrinkage parameter, starting from zero, to the least-squares fit. As a result, LARS 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. In addition, MM-estimators are consistent and possess the exact fit property. Salibian-Barrera and Zamar (2002) further studied the distribution of MM-estimates using a robust bootstrap method. We apply MM-regression to filter out variables at a certain significance level. Since in general the resulting number of filtered 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. Therefore, this context-sensitive approach 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. The final Section 4 concludes.

2. Context-sensitive model selection

Let \( y = (y_1, \ldots, y_n)' \) be the response variable and \( x_1 = (x_{11}, \ldots, x_{1n})', \ldots, x_p = (x_{1p}, \ldots, x_{np})' \) 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 \( y \) 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 the (robust) correlations among the selected predictors are low. These goals of easy-to-interpret models and low dependencies between the predictors reflect the context-sensitivity of our method.

2.1. Description of the algorithm

For a start, the response variable \( y \) and the candidate predictors \( x_1, \ldots, x_p \) are robustly centered and scaled using median and MAD, according to

\[
y_i^* = \frac{y_i - \text{med}(y_1, \ldots, y_n)}{\text{MAD}(y_1, \ldots, y_n)}, \quad i = 1, \ldots, n
\]

\[
x_{ij}^* = \frac{x_{ij} - \text{med}(x_{1j}, \ldots, x_{nj})}{\text{MAD}(x_{1j}, \ldots, x_{nj})}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, p.
\]

Hence all predictor variables \( x_j^* = (x_{1j}^*, \ldots, x_{nj}^*)' \), \( j = 1, \ldots, p \), 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 \( y^* = (y_1^*, \ldots, y_n^*)' \) and \( x_1^*, \ldots, x_p^* \) to find a sequence \( (x_j^*)_{j \in J_1} \), \( J_1 \subset J \), of candidate predictors for \( y^* \) 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 the second step, the covariates $x_j^*$, $j \in J_1$, are entered as predictors for $y^*$ in MM-regression (Yohai 1987 [Maronna et al. 2006]. We apply MM-regression to filter out the significant variables. Let $J_2 \subseteq J_1$ be the set of indices of the significant variables at a given significance level $\alpha$. Then the second step concludes with fitting another MM-regression model to $y^*$, using only the significant predictors $x_j^*$, $j \in J_2$. Thus we consider the regression model

$$y^*_i = (x_i^*)^T \beta + e_i, \quad i = 1, \ldots, n,$$

where $x_i^*$ denotes the $i$-th observation of the predictor variables $x_j^*$, $j \in J_2$, extended by 1 in the first component to account for the intercept. Furthermore, $\beta$ 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 $r_i(\beta) = y^*_i - (x_i^*)^T \beta$, then MM-regression solves the problem

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^n \rho\left(\frac{r_i(\beta)}{\hat{\sigma}}\right),$$

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 $\beta$ yields

$$\sum_{i=1}^n \psi\left(\frac{r_i(\beta)}{\hat{\sigma}}\right) x_i^* = 0$$

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

$$w_i = \frac{\psi(r_i(\beta)/\hat{\sigma})}{r_i(\beta)/\hat{\sigma}}, \quad i = 1, \ldots, n,$$

allows (5) to be rewritten as

$$\sum_{i=1}^n w_i \left(y^*_i - (x_i^*)^T \hat{\beta}\right) x_i^* = 0.$$  

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 $\rho$ 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 $\tilde{w}_i$, $i = 1, \ldots, n$, for the MM-regression estimator $\hat{\beta}$ 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 like 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 $\tilde{y} = (\tilde{w}_1 y_1^*, \ldots, \tilde{w}_n y_n^*)^T$ and the candidate predictors $\tilde{x}_j = (\tilde{w}_1 x_{1j}^*, \ldots, \tilde{w}_n x_{nj}^*)^T$, $j \in J_2$. 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 $\{x_j^*: j \in J_3\}$ with $J_3 \subseteq J_2$, $|J_3| \leq k$, of the set of candidate predictors $\{x_j^*: j \in J_2\}$.

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 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. For a trimming factor $0 \leq \gamma < 0.5$, the RTMSEP is defined as
\[
\text{RTMSEP} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} r_i^2(j)}
\]  
where $r_i = y_i - \hat{y}_i$, $i = 1, \ldots, n$, are the residuals using the predictions $\hat{y}_i$ from cross-validation, $r_i^2(j) \leq \cdots \leq r_i^2(n)$ 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 $y^* = (y_1^*, \ldots, y_n^*)'$, and the resulting candidate predictors by $x_i^* = (x_{i1}^*, \ldots, x_{iJ}^*)'$, $i = 1, \ldots, p$. 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. Compute a sequence $\{x_{i}^*\}_{i \in J}$, $J_1 \subset J$, of candidate predictors with $k < |J_1| \ll p$ by performing B-RLARS on $y^*$ and $x_1^*, \ldots, x_p^*$.

2. Use $x_j^*$, $j \in J_1$, as predictors for $y^*$ in MM-regression. Let $J_2 \subseteq J_1$ be the set of indices of the significant variables at a given significance level $\alpha$. Fit another MM-regression model to $y^*$ with only the significant predictors $x_j^*$, $j \in J_2$, 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 $\hat{y} = (\hat{w}_1y_1^*, \ldots, \hat{w}_ny_n^*)'$ and the candidate predictors $\hat{x}_j = (\hat{w}_1x_{1j}^*, \ldots, \hat{w}_nx_{nj}^*)'$, $j \in J_2$. This robustified version of $k$-subset regression yields the optimal subset $\{x_j^* : j \in J_3\}$ with $J_3 \subseteq J_2$, $|J_3| \leq k$, of the set of candidate predictors $\{x_j^* : j \in J_2\}$.

2.3. Diagnostics

Low robust correlations among the selected predictor variables are a major demand for our context-sensitive method. 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 $x_j$, $j \in J_1$, 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 like the minimum covariance determinant (MCD; Rousseeuw and van Driessen [1999]) or the orthogonalized Gnanadesikan-Kettenring estimator (OGK; Maronna and Zamar [2002]). Let $R = (r_{ij})_{i,j \in J}$ denote such a robust estimate of the correlation matrix. Then the dissimilarity matrix $D = (d_{ij})_{i,j \in J}$ is defined as
\[
d_{ij} = 1 - |r_{ij}|, \quad i, j \in J_1,
\]  
which 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
\[
d(A, B) = \max_{x_i \in A, x_j \in B} d_{ij},
\]
Using (9), this can be written as
\[ d(A, B) = 1 - \min_{x_i \in A, x_j \in B} |r_{ij}|. \] (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.

Figure 1: Dendrogram (based on robust correlations) of the initial B-RLARS sequence of candidate predictors for quality of life.

Figure 1 shows such a dendrogram for the example in Section 3. 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 MCD.

2.4. Implementation

An implementation of our algorithm in R (R Development Core Team, 2008), as well as a detailed description of its use, is available at [http://www.statistik.tuwien.ac.at/public/filz/programs.html](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](http://users.ugent.be/~svaelst/software/RLARS.html). In addition, the R-packages robustbase and leaps, which are available at CRAN (the Comprehensive R Archive Network, [http://cran.r-project.org](http://cran.r-project.org)), are required.

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 detailed description of the data, in particular Table 1 for an explanation of the 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 135 remaining candidate predictors. 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. 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).

Instead of 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). 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

\[
R_{rob}^2 = 1 - \frac{\text{med}(|y_1 - \hat{y}_1|, \ldots, |y_n - \hat{y}_n|)}{\text{MAD}(y_1, \ldots, y_n)}^2,
\]

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 2 (left) suggests using the first 8 variables of the sequence: contrFarmers, agriculture, medCare, merchComm, impOrganic, merchAssort, percAdolescent 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 estimated using fivefold cross-validation. The final model resulting from this procedure contains the predictors agriculture, merchAssort and interesting.

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Standard error</th>
<th>t-Value</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-2.302</td>
<td>11.227</td>
<td>-0.205</td>
</tr>
<tr>
<td>agriculture</td>
<td>0.251</td>
<td>0.053</td>
<td>4.713</td>
</tr>
<tr>
<td>medCare</td>
<td>0.076</td>
<td>0.023</td>
<td>3.228</td>
</tr>
<tr>
<td>merchAssort</td>
<td>0.177</td>
<td>0.064</td>
<td>2.751</td>
</tr>
<tr>
<td>eduProTraining</td>
<td>0.117</td>
<td>0.026</td>
<td>4.450</td>
</tr>
<tr>
<td>beauty</td>
<td>0.292</td>
<td>0.113</td>
<td>2.588</td>
</tr>
<tr>
<td>parish</td>
<td>0.216</td>
<td>0.035</td>
<td>6.226</td>
</tr>
</tbody>
</table>

Robust residual standard error: 1.705

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

Table 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
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 2 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, whereas the B-RLARS-8 model clearly leads to the worst prediction performance.

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 like 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 filtering out variables of the initial B-RLARS sequence at a certain significance level in MM-regression.
requirement was that the robust correlations between the selected variables should be low. The dendrogram in Figure 1 shows that RCS was able to fulfill this demand. 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 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™2 Quad 2.66GHz processor and 8GB main memory. Keep in mind that the computations were carried out with R version 2.8.0 (and thus only one of the four available processors was effectively used), and that the data set consists of 43 observations and 135 variables. With the parameter settings as described in the beginning of Section 3.1, RCS completed after 20.61 seconds. Thereby the running time was dominated by computing the initial B-RLARS sequence, which took 20.54 seconds. 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 fivefold 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. Conclusions

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. Thereby the number of explanatory variables 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.

Nowadays several methods for model selection in high dimensions are available, 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, 2008) 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 significant variables with MM-regression \cite{Yohai1987, Maronna2006} 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 \cite{Furnival1974, Miller2002, Gatu2006}, which is robustified by using the weights obtained from another MM-regression model with only the significant explanatory variables.

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 1). Moreover, the resulting model is an excellent fit and performs well with respect to prediction. 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. \cite{Filzmoser2008}, 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.

Acknowledgments

The research was supported by a grant of the Austrian Research Promotion Agency (FFG), project ref. no. 813000/10345.

References


Computational Statistics & Data Analysis 52 (1), 239–248.


URL http://www.r-project.org