LMU
LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR STATISTIK

Jan Gertheiss & Gerhard Tutz

Sparse Modeling of Categorial Explanatory Variables

Technical Report Number 060, 2009
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Sparse Modeling of Categorial Explanatory Variables

Jan Gertheiss & Gerhard Tutz

Ludwig-Maximilians-Universität München

Akademiestraße 1, 80799 München

August 17, 2009

Abstract

Shrinking methods in regression analysis are usually designed for metric predictors. If independent variables are categorial some modifications are necessary. In this article two L_1-penalty based methods for factor selection and clustering of categories are presented and investigated. The first approach is designed for nominal scale levels, the second one for ordinal predictors. All methods are illustrated and compared in simulation studies, and applied to real world data from the Munich rent standard.

Keywords: Fused Lasso, Variable Fusion, Categorial Predictors, Ordinal Predictors

1 Introduction

Within the last decade regularization and in particular variable selection has been a topic of intensive research. With the introduction of the Lasso, proposed by (Tibshirani, 1996), sparse modeling in the high-predictor case with good performance in terms of identification of relevant variables combined with good performance in predictive power became possible. In the following many alternative regularized estimators that include variable selection were proposed, among them the elastic net (Zou and Hastie, 2005), SCAD (Fan and Li, 2001), the Dantzig selector (Candes and Tao, 2007) and boosting approaches (for example Bühlmann and Yu, 2003).

Most of these methods focus on the selection of variables in the case where the effect of one variable is determined by one coefficient, that means one selects coefficients rather than variables. When all predictors are metric and a main effect model is assumed to hold, of course selection of coefficients is equivalent to selection of predictor variables. This is different when categorical variables have to be included because then a group of coefficients refers to one variable. One of the few approaches that explicitly select categorical predictors was proposed by Yuan and Lin (2006) under the name *Group Lasso*. The approach explicitly includes or excludes groups of coefficient that refer to one variable.

In selection problems for categorical predictors it should be distinguished between two problems:

- Which categorical predictors should be included in the model?
- Which categories within one categorical predictor should be distinguished?

The latter problem is concerned with one variable and poses the question which categories differ from one another with respect to the dependent variable. Or, to put it in a different way, which categories should be collapsed? The answer to that question depends on the scale level of the predictor, one should distinguish between nominal and ordered categories because of their differing information content.

To be more concrete let us first consider just one categorial predictor $C \in \{0, \ldots, k\}$ and dummy coding $x_i = I_{\{C=i\}}$. Then the classical linear model is given as

$$y = \alpha + \sum_{i=0}^{k} \beta_i x_i + \epsilon,$$

with $E(\epsilon) = 0$ and $\text{Var}(\epsilon) = \sigma^2$. If category 0 is chosen as reference, coefficient β_0 is fixed to zero. When computing a penalized estimate, for example by use of the simple Lasso (Tibshirani, 1996), the shrinkage effect depends on the coding scheme that is used and the choice of the reference category. With category zero chosen as reference, shrinkage always refers to the difference between category i and zero. Moreover, Lasso type penalties tend to set some coefficients to zero. Usually this feature is seen as a great advantage over methods like Ridge regression, since it can be used for model/variable selection. Applied to dummy coded categorial predictors, however, selection only refers to the currently chosen reference category. In most cases of nominal predictors, class labeling and choice of the reference category is arbitrary, which means that the described selection procedures are not really meaningful. In addition, the estimated model is not invariant against irrelevant permutations of class labels.

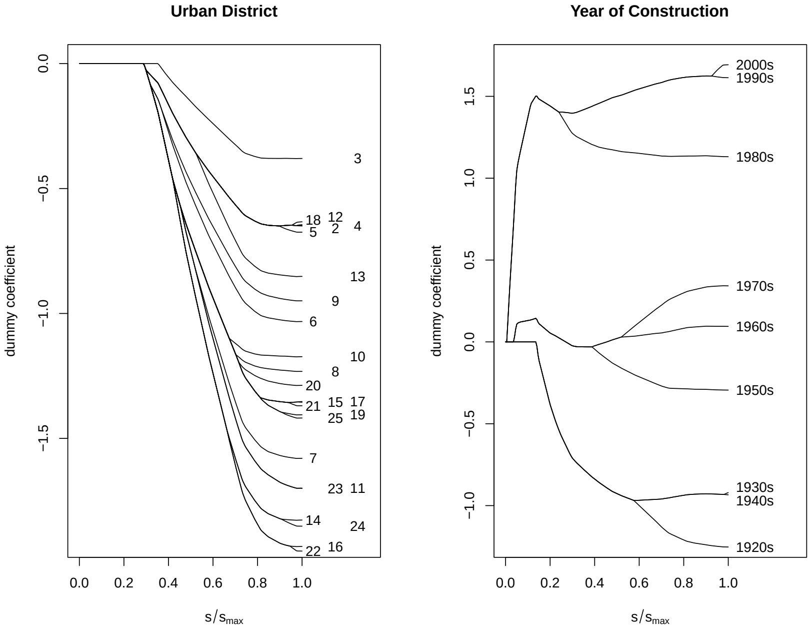

Figure 1: Paths of dummy coefficients of categorial predictors obtained by the proposed method.

For categorical predictor variables with many categories a useful strategy is to search for clusters of categories with similar effects. The objective is to reduce the $k+1$ categories to a smaller number of categories which form clusters. The effect of categories within one cluster is supposed to be the same but responses will differ across clusters. An example, which will be considered in more detail in Section 4, is the modeling of the influence of the urban district where a person lives on the rent she/he has to pay. The data comes from the Munich rent standard, where data is available for all 25 urban districts of Munich. It can be expected that not all districts do differ substantially. Therefore the aim is to combine districts which do not differ in terms of rent per square meter. Hence, in a regression model corresponding dummy coefficients should be equal. Figure 1 shows paths of dummy coefficients obtained by the method proposed in this article. It is seen that with decreasing tuning parameter s, categories are successively fused, i.e. coefficients are set equal. In addition, several other covariates are given, among them the (categorized) year of construction. Corresponding paths of dummy coefficients are also shown in Figure 1.

Clustering or *fusion* of metric predictors was for example realized by so-called Variable Fusion (Land and Friedman, 1997) and the Fused Lasso proposed by Tibshirani et al. (2005). If predictors can be reasonably ordered, by putting a L_1-penalty on differences of adjacent coefficients many of these differences are set to zero which produces a piecewise constant coefficient function. Recently, Bondell and Reich (2009) adapted this methodology for factor selection and level fusion in ANOVA. The result are dummy coefficients being constant over some categories. In the following this method is reviewed and applied to regression problems. Some modifications are proposed and an approximate solution is presented which allows for easy computation of coefficient paths. In addition, the method is adapted for the modeling ordinal predictors.

2 Regularization for Categorical Predictors

In the following we consider the penalized least squares criterion

$$Q_p(\beta) = (y - X\beta)^T(y - X\beta) + \lambda J(\beta), \tag{1}$$

with penalty $J(\beta)$. The estimate of β is given by

$$\hat{\beta} = \text{argmin}_\beta\{Q_p(\beta)\}. \tag{2}$$

The decisive point is a suitable choice of penalty $J(\beta)$. We start with the case of one variable and will distinguish between nominal and ordinal predictors.

2.1 Unordered Categories

If the categorial predictor has only nominal scale level, a modification of Variable Fusion (Land and Friedman, 1997) and the Fused Lasso (Tibshirani et al., 2005), has been proposed by Bondell and Reich (2009) in the form of the penalty

$$J(\beta) = \sum_{i>j} w_{ij}|\beta_i - \beta_j|, \tag{3}$$

with weights w_{ij} and β_i denoting the coefficient of dummy x_i. Since the ordering of $x_0, \ldots, x_k$ is arbitrary, not only differences $\beta_i - \beta_{i-1}$ (as in original fusion methodology), but all differences $\beta_i - \beta_j$ are considered. Since $i = 0$ is chosen as reference, $\beta_0 = 0$ is fixed. Therefore in the limit case, $\lambda \to \infty$, all β_i are set to zero and the categorial predictor C is excluded from the model since no categories are distinguished anymore. For $\lambda < \infty$ the Lasso type penalty (3) sets only some differences $\beta_i - \beta_j$ to zero, which means that categories are clustered. With adequately chosen weights w_{ij} some nice asymptotic properties of $\hat{\beta}$ can be derived. These (adaptive) weights decisively depend on the distance of the ordinary least squares estimates $\hat{\beta}_i^{(LS)}$ and $\hat{\beta}_j^{(LS)}$.

Let $\theta = (\theta_{10}, \theta_{20}, \ldots, \theta_{k,k-1})^T$ denote the vector of pairwise differences $\theta_{ij} = \beta_i - \beta_j$. Furthermore, let $\mathcal{C} = \{(i,j) : \beta_i^* \neq \beta_j^*, i > j\}$ denote the set of indices $i > j$ corresponding to differences of (true) dummy coefficients β_i^* which are truly non-zero, and $\mathcal{C}_n$ denote the set corresponding to those difference which are estimated to be non-zero with sample size n. If $\theta^*_\mathcal{C}$ denotes the true vector of pairwise differences included in $\mathcal{C}$, and $\hat{\theta}_\mathcal{C}$ denotes the corresponding estimate based on $\hat{\beta}$, then a slightly modified version of Theorem 1 in Bondell and Reich (2009) holds:

Proposition 1 *Suppose $\lambda = \lambda_n$ with $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n \to \infty$, and all class-wise sample sizes n_i satisfy $n_i/n \to c_i$, where $0 < c_i < 1$. Then weights $w_{ij} = \phi_{ij}(n)|\hat{\beta}^{(LS)}_i - \hat{\beta}^{(LS)}_j|^{-1}$, with $\phi_{ij}(n) \to q_{ij}$ $(0 < q_{ij} < \infty)$ $\forall i, j$, ensure that*

(a) $\sqrt{n}(\hat{\theta}_\mathcal{C} - \theta^*_\mathcal{C}) \to_d N(0, \Sigma)$,

(b) $\lim_{n\to\infty} P(\mathcal{C}_n = \mathcal{C}) = 1$.

The proof closely follows Zou (2006) and Bondell and Reich (2009), and is given in the Appendix. The main differences to Bondell and Reich (2009) are that a concrete form of the dependence on sample size, specified in $\phi_{ij}(n)$, is not yet chosen, and that λ_n is determined by $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n \to \infty$. The latter is necessary for the proof of asymptotic normality, as given in Zou (2006). Bondell and Reich (2009) used $\lambda_n = O_p(\sqrt{n})$, which also allows $\lambda_n = 0$ and therefore cannot yield $\lim_{n\to\infty} P(\mathcal{C}_n = \mathcal{C}) = 1$. Note, that the covariance matrix Σ of the asymptotic normal distribution is singular due to linear dependencies of pairwise differences, cf. Bondell and Reich (2009).

Simple consistency $\lim_{n\to\infty} P(||\hat{\beta} - \beta^*||^2 > \epsilon) = 0$ for all $\epsilon > 0$, is also reached if λ is fixed and $w_{ij} = \phi_{ij}(n)$, with $\phi_{ij}(n) \to q_{ij}$ $(0 < q_{ij} < \infty)$ $\forall i, j$, is chosen. The proof is given in the Appendix. The issue, how to select concrete weights in the $n < \infty$ case, is further addressed in Sections 2.5 and 3.2.

2.2 Ordered Categories

An interesting case are selection strategies for ordinal predictors, as for example the decade of construction from Figure 1. Ordered categories contain more information than unordered ones but the information has not been used in the penalties considered so far. Since in the case of ordered categories the ordering of dummy coefficients is meaningful, original fusion methodology can be applied, which suggests penalty

$$J(\beta) = \sum_{i=1}^{k} w_i|\beta_i - \beta_{i-1}|, \tag{4}$$

with $\beta_0 = 0$. In analogy to asymptotic properties for the unordered case, with adequately chosen weights w_i similar results can be derived. Let now $\mathcal{C} = \{i > 0 : \beta^*_i \neq \beta^*_{i-1}\}$ denote the set of indices corresponding to differences of neighboring (true) dummy coefficients β^*_i which are truly non-zero, and again, $\mathcal{C}_n$ denote the set corresponding to those difference which are estimated to be non-zero. The vector of first differences $\delta_i = \beta_i - \beta_{i-1}$, $i = 1, \ldots, k$, is now denoted as

$\delta = (\delta_1, \dots, \delta_k)^T$. In analogy to the unordered case, $\delta^*_{\mathcal{C}}$ denotes the true vector of (first) differences included in $\mathcal{C}$, and $\hat{\delta}_{\mathcal{C}}$ the corresponding estimate. With $\hat{\beta}_i^{(LS)}$ denoting the ordinary least squares estimate of β_i, the following proposition holds.

Proposition 2 *Suppose* $\lambda = \lambda_n$ *with* $\lambda_n/\sqrt{n} \to 0$ *and* $\lambda_n \to \infty$, *and all class-wise sample sizes* n_i *satisfy* $n_i/n \to c_i$, *where* $0 < c_i < 1$. *Then weights* $w_i = \phi_i(n)|\hat{\beta}_i^{(LS)} - \hat{\beta}_{i-1}^{(LS)}|^{-1}$, *with* $\phi_i(n) \to q_i$ $(0 < q_i < \infty)$ $\forall i$, *ensure that*

(a) $\sqrt{n}(\hat{\delta}_{\mathcal{C}} - \delta^*_{\mathcal{C}}) \to_d N(0, \Sigma)$,

(b) $\lim_{n\to\infty} P(\mathcal{C}_n = \mathcal{C}) = 1$.

The proof is a direct application of Theorem 2 in Zou (2006), as sketched in the Appendix. As before in the unordered case, simple consistency $\lim_{n\to\infty} P(||\hat{\beta} - \beta^*||^2 > \epsilon) = 0$ for all $\epsilon > 0$, is reached if λ is fixed and $w_i = \phi_i(n)$, with $\phi_i(n) \to q_i$ $(0 < q_i < \infty)$ $\forall i, j$. The proof is completely analogue to the proof of Proposition 3 in the Appendix.

2.3 Computational Issues

For estimation it is useful to consider the penalized minimization problem (2) as a constrained minimization problem. That means, $(y - X\beta)^T(y - X\beta)$ is minimized subject to a constraint. For unordered categories the constraint corresponding to penalty (3) is

$$\sum_{i>j} w_{ij}|\beta_i - \beta_j| \leq s,$$

with $\beta_0 = 0$. There is a one-to-one correspondence between the bound s and penalty parameter λ in (1), cf. Bondell and Reich (2009). As already mentioned, transformed parameters $\theta_{ij} = \beta_i - \beta_j$ yield vector $\theta = (\theta_{10}, \theta_{20}, \dots, \theta_{k,k-1})^T$. If θ is directly estimated (instead of β), one has to take into account that restrictions $\theta_{ij} = \theta_{i0} - \theta_{j0}$ must hold for all $i, j > 0$. For practical estimation, parameters θ_{ij} are additionally split into positive and negative parts, i.e.

$$\theta_{ij} = \theta^+_{ij} - \theta^-_{ij},$$

with

$$\theta^+_{ij} \geq 0, \ \theta^-_{ij} \geq 0,$$

and

$$\sum_{i>j} w_{ij}(\theta^+_{ij} + \theta^-_{ij}) \leq s.$$

Minimization can be done by using quadratic programming methods, we used R 2.9.0 (R Development Core Team, 2009) and the interior point optimizer from add-on package kernlab (Karatzoglou et al., 2004).

A fast approximate solution can be computed using R add-on package `lars` (Efron et al., 2004), where "approximate" means that only $\theta_{ij} \approx \theta_{i0} - \theta_{j0}$ holds. For simplicity we assume that weights $w_{ij} = 1$ are chosen. But results can be generalized easily (see Section 2.5). For the approximation we exploit that the proposed estimator can be seen as the limit of a generalized Elastic Net. The original Elastic Net (Zou and Hastie, 2005) uses a combination of simple Ridge and Lasso penalties. We use a generalized form where the quadratic penalty term is modified. With Z so that $Z\theta = X\beta$, we define

$$\hat{\theta}_{\gamma,\lambda} = \text{argmin}_\theta \left\{ (y - Z\theta)^T (y - Z\theta) + \gamma \sum_{i>j>0} (\theta_{i0} - \theta_{j0} - \theta_{ij})^2 + \lambda \sum_{i>j} |\theta_{ij}| \right\}.$$

A simple choice of Z is $Z = (X|0)$, since $\theta_{i0} = \beta_i$, $i = 1, \ldots, k$. The first penalty term, which is weighted by γ, penalizes violations of restrictions $\theta_{ij} = \theta_{i0} - \theta_{j0}$. The exact solution of the optimization problem considered here is obtained as the limit

$$\hat{\theta} = \lim_{\gamma \to \infty} \hat{\theta}_{\gamma,\lambda}.$$

Hence with sufficiently high γ an acceptable approximation should be obtained. If matrix A represents restrictions $\theta_{ij} = \theta_{i0} - \theta_{j0}$ in terms of $A\theta = 0$, one may define precision by

$$\Delta_{\gamma,\lambda} = (A\hat{\theta}_{\gamma,\lambda})^T A\hat{\theta}_{\gamma,\lambda}.$$

The lower $\Delta_{\gamma,\lambda}$ the better. An upper bound is given by

$$\Delta_{\gamma,\lambda} \leq \frac{\lambda(|\hat{\theta}^{(LS)}| - |\hat{\theta}_{0,\lambda}|)}{\gamma},$$

where $\hat{\theta}^{(LS)}$ denotes the least squares estimate (i.e. $\lambda = 0$) where $A\hat{\theta}^{(LS)} = 0$ holds, and $|\theta| = \sum_{i>j} |\theta_{ij}|$ denotes the L_1-norm of vector θ. (For a proof see the Appendix.) $\hat{\theta}^{(LS)}$ can be computed by $\hat{\theta}_{\gamma,0}$ if any $\gamma > 0$ is chosen. Not surprisingly, for higher λ also higher γ must be chosen to stabilize precision.

The advantage of using the estimate $\hat{\theta}_{\gamma,\lambda}$ is that its whole path can be computed using `lars` (Efron et al., 2004), since it can be formulated as a Lasso solution. With augmented data $\widetilde{Z} = (Z^T, \sqrt{\gamma}A^T)^T$ and $\widetilde{y} = (y^T, 0)^T$, one has

$$\hat{\theta}_{\gamma,\lambda} = \text{argmin}_\theta \left\{ (\widetilde{y} - \widetilde{Z}\theta)^T (\widetilde{y} - \widetilde{Z}\theta) + \lambda \sum_{i>j} |\theta_{ij}| \right\},$$

which is a Lasso type problem on data $(\widetilde{y}, \widetilde{Z})$.

In the case of ordinal predictors the penalty is

$$J(\beta) = \sum_{i=1}^{k} |\beta_i - \beta_{i-1}|,$$

and the corresponding optimization problem can be directly formulated as a simple Lasso type problem. We write

$$Q_p(\beta) = (y - X\beta)^T(y - X\beta) + \lambda J(\beta) = (y - \widetilde{X}\delta)^T(y - \widetilde{X}\delta) + \lambda J(\delta),$$

with $\widetilde{X} = XU^{-1}$, $\delta = U\beta$, $J(\delta) = \sum_{i=1}^{k} |\delta_i|$, and

$$U = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & 1 & \cdots & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & \cdots & -1 & 1 \end{pmatrix}.$$

Simple matrix multiplication shows that the inverse is given by

$$U^{-1} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ 1 & \cdots & \cdots & 1 \end{pmatrix}.$$

In other words, the ordinal input is just split-coded (Walter et al., 1987), and ordinary Lasso estimation is applied. Split-coding means that dummies $\widetilde{x}_i$ are defined by splits at categories $i = 1, \ldots, k$, i.e.

$$\widetilde{x}_i = \begin{cases} 1 & \text{if } C \geq i \\ 0 & \text{otherwise.} \end{cases}$$

Now the model is parameterized by coefficients $\delta_i = \beta_i - \beta_{i-1}$, $i = 1, \ldots, k$. Thus transitions between category i and $i-1$ are expressed by coefficient δ_i. Original dummy coefficients are obtained by back-transformation $\beta_i = \sum_{s=1}^{i} \delta_s$. By applying penalty $\sum_{i=1}^{k} |\delta_i|$ not the whole ordinal predictor is selected, but only relevant transitions between adjacent categories. By contrast, Walter et al. (1987) intended the use of classical tests for such identification of substantial *"between-strata differences"*.

2.4 Multiple Inputs

In statistical modeling, usually a set of (potential) regressors is available and one wants to include the relevant ones into the predictor. In the introduction we already considered two predictors, the urban district where a flat is located and the decade of construction. For the handling of multiple categorial predictors, say $x_1, \ldots, x_p$, with levels $0, \ldots, k_l$ for variable x_l ($l = 1, \ldots, p$), the presented methods can be easily generalized. The corresponding penalty is

$$J(\beta) = \sum_{l=1}^{p} J_l(\beta_l), \tag{5}$$

with

$$J_l(\beta_l) = \sum_{i>j} w_{ij}^{(l)} |\beta_{li} - \beta_{lj}|, \text{ or } J_l(\beta_l) = \sum_{i=1}^{k_l} w_i^{(l)} |\beta_{li} - \beta_{l,i-1}|,$$

depending on the scale level of predictor x_l. The first expression refers to nominal covariates, the second to ordinal ones. Due to the (additive) form of the penalty theoretic results from above directly generalize to the case of multiple categorial inputs.

If multiple predictors are considered, clustering of categories of single predictors as well as selection of predictors is of interest. Penalty (5) serves both objectives, clustering and selection. If all dummy coefficients which belong to a specific predictor are set to zero, the corresponding predictor is excluded from the model. Within each nominal predictor x_l, there is also an L_1-penalty on the differences to the dummy coefficient of the reference category. Since the latter is fixed to zero, clustering of all categories of x_l means that all coefficients which belong to predictor x_l are set to zero. In the ordinal case, this happens if all differences $\delta_{li} = \beta_{li} - \beta_{l,i-1}$ of adjacent dummy coefficients of predictor x_l are set to zero.

2.5 Incorporation of Weights

In many situations weights $w_{ij}^{(l)} \neq 1$ are to be preferred over the simple weights $w_{ij}^{(l)} = 1$; for example to obtain the adaptive versions described in Propositions 1 and 2, or when predictors differ in the number of levels. For nominal variables Bondell and Reich (2009) suggested the weights

$$w_{ij}^{(l)} = (k_l + 1)^{-1} \sqrt{\frac{n_i^{(l)} + n_j^{(l)}}{n}}, \tag{6}$$

where $n_i^{(l)}$ denotes the number of observations on level i of predictor x_l. In the adaptive version the weights contain additionally the factor $|\hat{\beta}_{li}^{(LS)} - \hat{\beta}_{lj}^{(LS)}|^{-1}$. The use of these weights was motivated through standardization of design matrix Z from Section 2.3, in analogy to standardization of metric predictors. In the following these weights are also used, but multiplied by 2. If predictor x_l is nominal, the factor $(k_l + 1)^{-1}$ is necessary to ensure that penalty $J_l(\beta_l)$ in (5) is of order k_l, the number of (free) dummy coefficients. Without these additional weights $J_l(\beta_l)$ would be of order $(k_l+1)k_l$, because the penalty consists of $(k_l+1)k_l/2$ terms if no ordinal structure is assumed. By contrast, if the predictor is ordinal, the penalty is already of order k_l. Hence the factor $2(k_l+1)^{-1}$ is omitted in this case.

In general, if weights $w_{ij}^{(l)} \neq 1$ are included, the model just has to be parameterized by vector $\tilde{\theta} = W\theta$, where W is a diagonal matrix with diagonal elements $w_{ij}^{(l)}$. That means the (centered) design matrix needs to be multiplied by W^{-1}.

3 Numerical Experiments

Before applying the presented methodology to the Munich rent standard data in Section 4, the different approaches are tested and some characteristics are investigated in simulation studies.

3.1 An Illustrative Example

In the first simulation scenario only one predictor and a balanced design are considered with 20 (independent) observations in each of $i = 0, \dots, 8$ classes. In class i the response is N(μ_i,4)-distributed, where the means form three distinct groups of categories, i.e. $\mu_0 = \mu_1 = \mu_2$, $\mu_3 = \mu_4 = \mu_5$, $\mu_6 = \mu_7 = \mu_8$. Figure 2 (left) shows empirical distributions as well as the true μ_i, which are marked by dashed lines. Moreover, exact and approximate paths of dummy coefficients (middle) are shown, where the non-adaptive version of penalty $J(\beta)$ is employed. That means the weighting term $|\hat{\beta}_i^{(LS)} - \hat{\beta}_j^{(LS)}|^{-1}$ is omitted. Since there is only one predictor and the design is balanced simple weights $w_{ij} = 1$ can be used. The x-axis indicates $s/s_{\max}$, the ratio of actual and maximal s value. The latter results in the ordinary least squares (OLS) estimate. With decreasing s (or increasing penalty λ) categories are successively grouped together. First, classes with the same true mean are grouped as desired; for $s = 0$ the model finally consists of the intercept only – the empirical mean of y. For the approximation, $\sqrt{\gamma} = 10^5$ has been chosen. It is hard to see any difference between approximate and exact solution. Indeed, for $s/s_{\max} \geq 10^{-3}$, precision $\Delta_{\gamma,\lambda} < 10^{-17}$ is obtained. Also in the case of the "exact" solution, restrictions are just "numerically" met. In the given example precision of the "exact" solution is about 10^{-18} (or better), which is quite close to the "approximate" solution. So in the following, only approximate estimates are used.

In the right panel of Figure 2, the results of the adaptive version which uses the additional weights $w_{ij} = |\hat{\beta}_i^{(LS)} - \hat{\beta}_j^{(LS)}|^{-1}$ are shown. Grouping is quite good, and compared to the non-adaptive version, bias towards zero is much smaller at the point of perfect grouping.

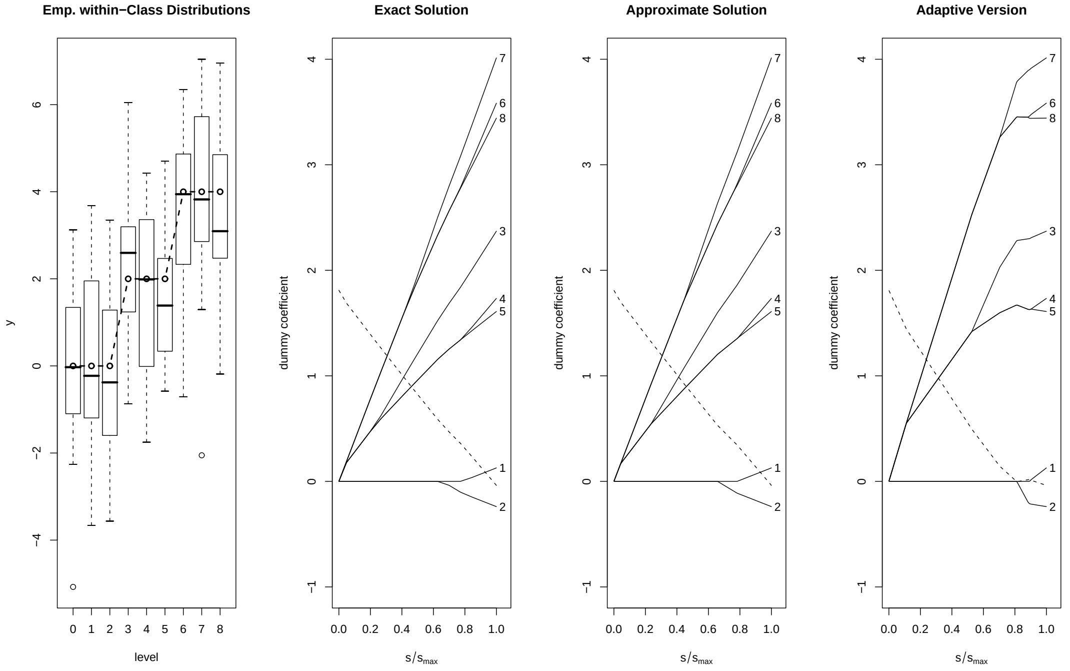

Figure 2: Empirical within-class distributions (left), exact and approximate coefficient paths (middle), as well as results of the adaptive version (right); constant α is marked by the dashed line.

In a second scenario, settings and data visualized in Figure 2 (left) are considered again, but now it is assumed that class labels have an ordinal structure. Hence penalty (4) is employed. Resulting paths of dummy coefficients are plotted in Figure 3. Even for the non-adaptive version (left), grouping is quite good. Moreover, before optimal grouping is reached, bias towards zero seems to be quite low. Of course, assuming an ordinal class structure, which is actually given because all categories with truly equal coefficients are groups of neighbors, makes the estimation problem easier.

3.2 Comparison of Methods

For the comparison of different methods a setting with 8 predictors is considered – 4 nominal and 4 ordinal factors. For both types of variables we use two factors with 8 categories and two with 4, of which in each case only one is relevant. The true non-zero dummy coefficient vectors are $(0, 1, 1, 1, 1, -2, -2)^T$ and $(0, 2, 2)^T$ for the nominal predictors, and $(0, 1, 1, 2, 2, 4, 4)^T$ and $(0, -2, -2)^T$ for the ordinal predictors (constant $\alpha = 1$). A training data set with $n = 500$ (independent) observations is generated according to the classical linear model with standard normal error ϵ. The vectors of marginal a priori class probabilities

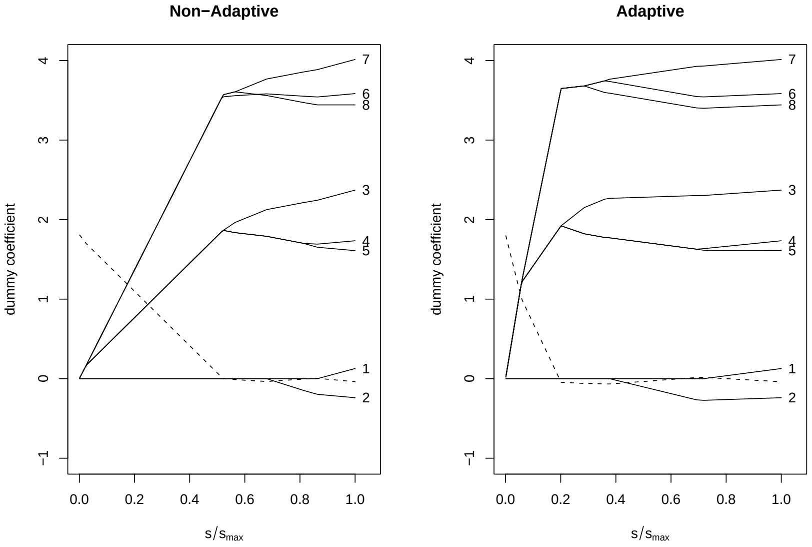

Figure 3: Paths of dummy coefficients for data as in Figure 2, but assuming an ordinal class structure, non-adaptive (left) and adaptive (right) version; constant α is marked by the dashed line.

are $(0.1, 0.1, 0.2, 0.05, 0.2, 0.1, 0.2, 0.05)^T$ and $(0.1, 0.4, 0.2, 0.3)^T$ for 8-level and 4-level factors, respectively. The coefficient vector is estimated by the proposed method, using adaptive as well as non-adaptive weights. In addition, the effect of taking into account marginal class frequencies $n_i^{(l)}$ is investigated, that means we check what happens if $((n_i^{(l)} + n_j^{(l)})/n)^{1/2}$ is omitted in (6). Moreover, refitting is tested, i.e. the penalty in (1) is only used for variable selection and clustering. After the identification of relevant predictors and clusters, parameters are fitted by ordinary least squares. Comparable procedures have already been proposed, for example, by Efron et al. (2004) under the name "Lars-OLS hybrid", or by Candes and Tao (2007) as "Gauss-Dantzig Selector".

For the determination of the right penalty λ, resp. s value, we use 5-fold cross-validation. Of course, any information criterion like AIC or BIC could also be employed. For the latter some measure of model-complexity is needed. In analogy to the Fused Lasso (Tibshirani et al., 2005), the degrees of freedom of a model can be estimated by

$$\widehat{\mathrm{df}} = 1 + \sum_{l=1}^{p} k_l^*,$$

where k_l^* denotes the number of unique non-zero dummy coefficients of predictor x_l, the 1 accounts for the intercept.

After estimation of coefficient vector β the result is compared to the true parameters. The MSE is computed, as well as False Positive and False Negative

adapt	Adaptive version, i.e. weighting terms $\lvert\hat{\beta}_i^{(LS)} - \hat{\beta}_j^{(LS)}\rvert^{-1}$ are used.
stdrd	Standard (non-adaptive) version, i.e. terms $\lvert\hat{\beta}_i^{(LS)} - \hat{\beta}_j^{(LS)}\rvert^{-1}$ are omitted.
n(ij)	Marginal class frequencies are taken into account, i.e. $((n_i^{(l)} + n_j^{(l)})/n)^{1/2}$ are used in (6).
rf	Refitting was performed.

Table 1: Definition of labels used in Figure 4 and 5

Rates (FPR/FNR) concerning variable selection and clustering. As far as variable selection is concerned, "false positive" means that any dummy coefficient of a pure noise factor is set to non-zero; if clustering is considered, it means that a difference within a non-noise factor which is truly zero is set to non-zero. By contrast, "false negative" means that all dummy coefficients of a truly relevant factor are set to zero, or that a truly non-zero difference is set to zero, respectively. Figure 4 shows the results for 100 simulation runs, labels are defined in Table 1.

In addition to the MSE and FPR/FNR, an independent test set of 1000 observations is generated and prediction accuracies are reported in terms of the mean squared error of prediction. For comparison also the performance of the ordinary least squares (OLS) estimate is given. MSE and prediction accuracy are shown as boxplots to give an idea of variability, FPR (dark gray) and FNR (light-colored) are averaged over all simulation runs. It is seen that all methods are superior to the OLS. Concerning FPR and FNR, differences between pure adaptive approaches and refitting are caused by the fact that not necessarily the same models are selected, because in the cross validation already refitted coefficients are used.

As already illustrated by Bondell and Reich (2009) and supported by Propositions 1 and 2, selection and grouping characteristics of the adaptive version are quite good – at least compared with the standard approach. Also accuracies of parameter estimates and prediction of the adaptive version are very high in our simulation study. Via refitting they can only be slightly improved. In the case of standard weights the improvement is much clearer. However, the most important effect of refitting is on variable selection and clustering – in both the adaptive and the non-adaptive case. It can be seen that via refitting error rates are enormously diminished – concerning false variable selection as well as clustering. This finding can be explained by the bias which is caused by shrinking. If tuning parameters are determined via cross validation (as done here), with refitting the chosen penalty parameter λ may be higher than without, because in the latter case a higher penalty directly results in a higher bias which may deteriorate prediction accuracy on the test fold. Since in the case of refitting the penalty is only used for selection purposes, a higher value does not necessarily cause higher coefficient shrinkage and bias. Apparently, however, in many of our simulated cases a higher penalty would have been necessary to obtain accurate variable selection and grouping.

In a modified scenario further noise variables are included, 4 nominal and 4 ordinal, each with 6 levels and constant marginal a priori class probabilities. Qualitatively, results (shown in Figure 5) are similar to those obtained before. However, since the number of independent variables has been considerably increased, the performance of the ordinary least squares estimates is even worse than above. This also explains why (in the adaptive case) the MSE and prediction accuracies cannot be really improved by OLS refitting, and why in the case of refitting variability is higher. Nevertheless, variable selection and clustering results are still distinctly better if refitting is done.

As an overall result, it can be stated that refitting has the potential to distinctly improve selection and clustering results in the $n < \infty$ case, while providing accurate parameter estimates (if n is not to small compared to p). Moreover, taking into account marginal class frequencies seems to (slightly) improve estimation results.

4 Application to Munich Rent Standard

All larger German cities compose so-called rent standards for having a decision making instrument available to tenants, landlords, renting advisory boards and experts. These rent standards are used in particular for the determination of the local comparative rent. For the composition of the rent standards, a representative random sample is drawn from all relevant households, and the interesting data are determined by interviewers by means of questionnaires. The data analyzed here comes from 2053 households interviewed for the Munich rent standard 2003. The response is monthly rent per square meter in Euro. The predictors are ordered as well as unordered and binary factors. A detailed description is given in Table 2. The data can be downloaded from the data archive of the Department of Statistics at the University of Munich. The direct link is `http://www.stat.uni-muenchen.de/service/datenarchiv/miete/miete03_e.html`.

For the estimation of regression coefficients corresponding to Table 2 we consider the approaches which performed best in the previous section; more concrete, both the adaptive as well as the standard (non-adaptive) version remain candidates, but each with refitting only and taking marginal class frequencies into account. In the considered application more than 2000 observations are available for the estimation of at maximum 58 regression parameters. So it can be assumed that OLS estimation is accurate, and hence (in the light of the simulation study before) refitting distinctly improves estimation accuracy as well as variable selection and clustering performance of the penalized approach.

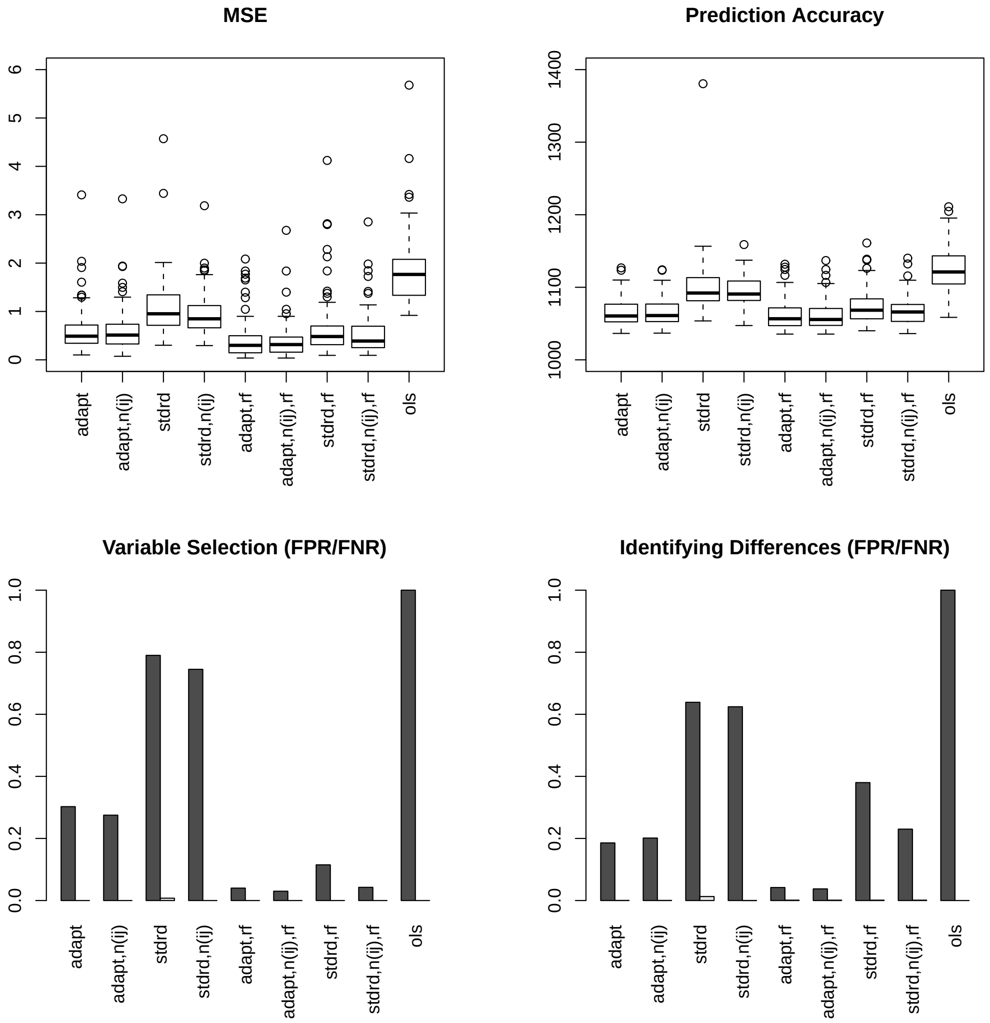

Figure 4: Evaluation of adaptive and non-adaptive (standard) as well as refitting (rf) approaches, taking into account class sizes (n_i,n_j) or not, for comparison also the results for the ordinary least squares (ols) estimator are given; considered are the mean squared error of parameter estimate, prediction accuracy, and false positive/negative rates (FPR/FNR) concerning variable selection and identification of relevant differences (i.e. clustering) of dummy coefficients.

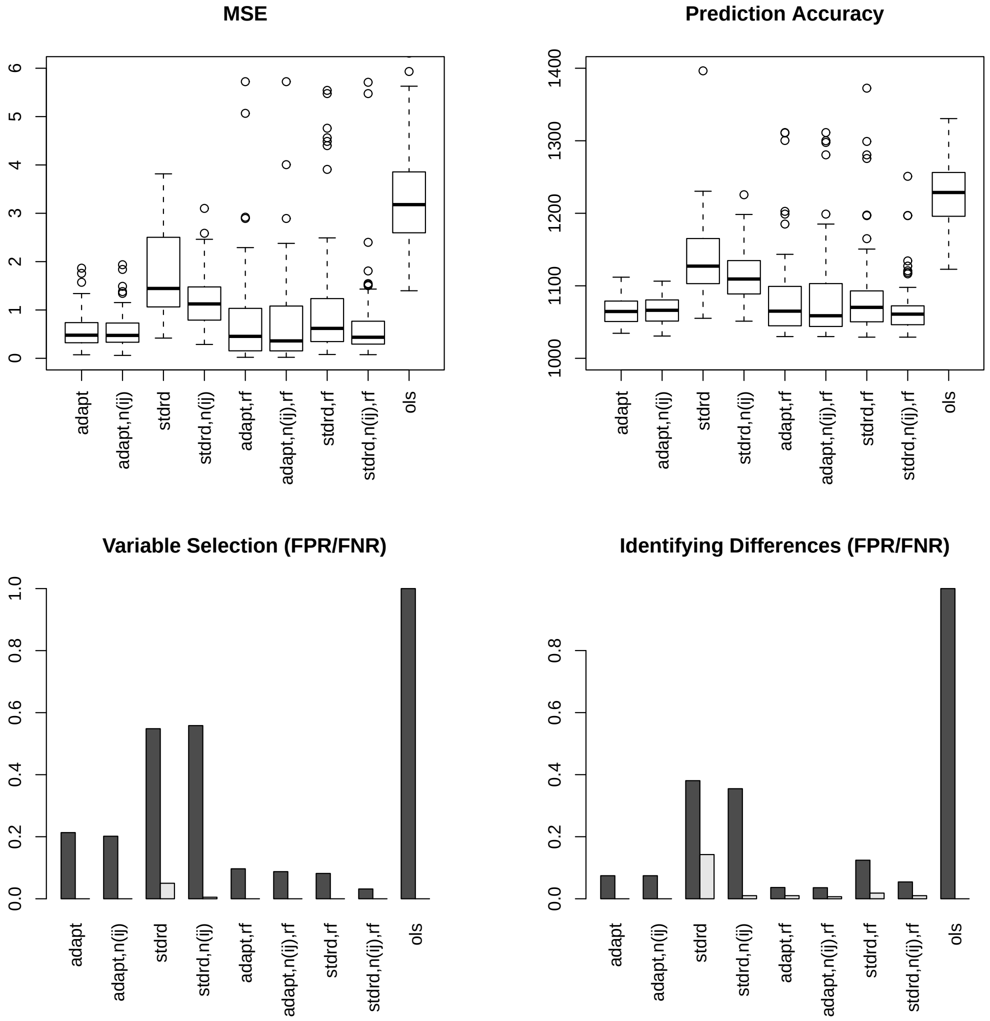

Figure 5: Evaluation of different approaches in the presence of many noise variables: adaptive and non-adaptive (standard) as well as refitting (rf), taking into account class sizes (n_i,n_j) or not, for comparison also the ordinary least squares (ols) estimator; considered are the mean squared error of parameter estimate, prediction accuracy, and false positive/negative rates (FPR/FNR) concerning variable selection and identification of relevant differences (i.e. clustering) of dummy coefficients.

urban district	nominal, labeled by numbers $1, \dots, 25$
year of construction	given in ordered classes $[1910, 1919]$, $[1920, 1929], \dots$
number of rooms	taken as ordinal factor with levels $1, 2, \dots, 6$
quality of residential area	ordinal, with levels "fair", "good", "excellent"
floor space (in m^2)	given in ordered classes $(0, 30)$, $[30, 40)$, $[40, 50)$, $\dots$, $[140, \infty)$
hot water supply	binary (yes/no)
central heating	binary (yes/no)
tiled bathroom	binary (yes/no)
supplementary equipment in bathroom	binary (no/yes)
well equipped kitchen	binary (no/yes)

Table 2: Explanatory variables for monthly rent per square meter.

Figure 6 shows the (10-fold) cross-validation score as a function of $s/s_{\max}$, for the refitted model with non-adaptive (dashed black) as well as adaptive weights (solid red). It is seen that with adequately chosen penalty, refitting with adaptive weights may improve the ordinary least squares estimate (i.e. $s/s_{\max} = 1$) in terms of prediction accuracy, whereas for non-adaptive weights such improvement is less obvious. It is plausible that adaptive weights are better than the non-adaptive ones, since a lot of data is available, which means that ordinary least squares estimates are quite stable, and the latter decisively influence adaptive weights. So we choose adaptive weights at cross-validation score minimizing $s/s_{\max} = 0.61$ (marked by dotted line in Figure 6). The estimated regression coefficients are given in Table 3. There is no predictor which is completely excluded from the model. However, some categories of nominal and ordinal predictors are clustered, for example houses constructed in the 1930s and 1940s, or urban districts 14, 16, 22 and 24. The biggest cluster, which contains 8 categories, is formed within the 25 districts. A map of Munich with color coded clusters (Figure 7) illustrates the 10 found clusters. The map has been drawn using functions from R add-on package `BayesX` (Kneib et al., 2009). The most expensive district is the city center. The fact that districts 16, 22 and 24 are found among the cheapest districts makes good sense, because Munich's deprived areas are primarily located in these (non-adjacent) districts. Concerning district 12, however, results partly contradict the experiences made by experts and tenants. The problem is that this district is very large and reaching from the city center to the outskirts in the north. So very expensive flats which are close to the city center are put together with the cheap ones on the outskirts. On average rents are rather high in this district,

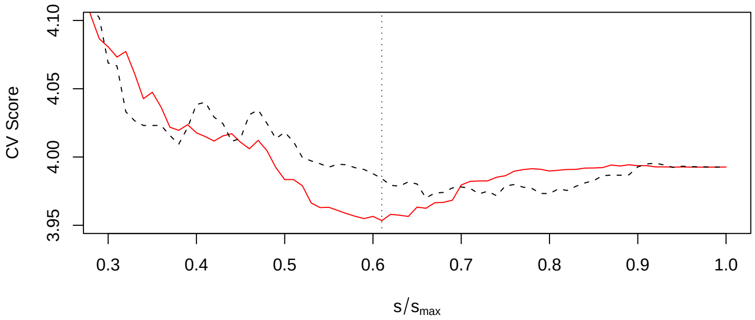

Figure 6: Cross-validation score as a function of $s/s_{\max}$ if refitting with standard (dashed black) or adaptive (solid red) weights is used for the analysis of Munich rent standard data.

which causes that it is clustered with other expensive but more homogeneous areas. So in Figure 7 some actually cheap regions in the north are marked as expensive. But this is a problem of the data, not of the type of penalization used here. In an ordinary least squares model, district 12 is even identified as belonging to the three most expensive districts. In our case it is only among the top seven. But note, in the final regression model there is also an ordinal predictor included which indicates the quality of the residential area and allows for further discrimination between flats which are located in the same district. Not surprisingly by contrast, rent per square meter goes down if the number of rooms increases. Between four, five or more rooms, however, no relevant differences are identified. Flats with two rooms are fused with the reference category, since the corresponding dummy coefficient is set to zero. The fact that no differences between flats with one and two rooms are fitted is caused by the inclusion of floor space into the model. Existing differences are obviously modeled via the variable which directly measures the flat's size, with the result: The larger the flat the lower the rent per square meter. Between 90 and 140 m^2, however, no differences are identified with respect to rent per square meter. All in all the selected model has 32 degrees of freedom, i.e. 32 unique non-zero coefficients (including the intercept), which means that the complexity of the unrestricted model (58 df) is reduced by about 45%.

5 Summary and Discussion

We showed how L_1-penalization of dummy coefficients can be employed for sparse modeling of categorial explanatory variables in multiple linear regression. Depending on the scale level of the categorial predictor two types of penalty were investigated. Given just nominal covariates, all pairwise differences of dummy

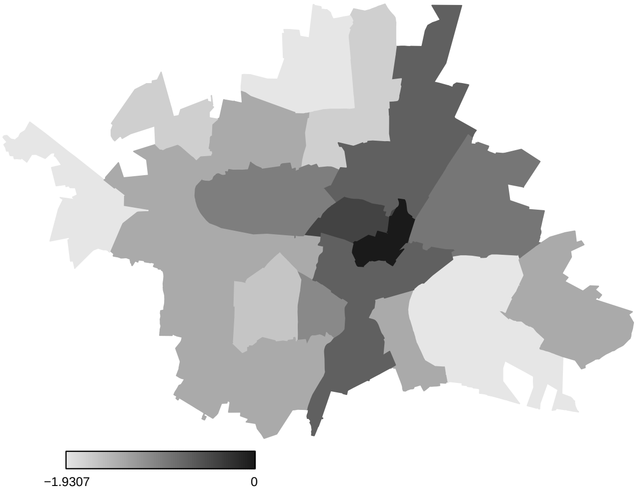

Figure 7: Map of Munich indicating clusters of urban districts; colors correspond to estimated dummy coefficients from Table 3.

coefficients belonging to the same predictor are penalized. If the variable has ordinal scale level differences of adjacent coefficients are considered. L_1-penalization causes that certain differences are set to zero. The interpretation is clustering of categories concerning the influence on the response. In the Munich rent standard example this meant that e.g. certain urban districts were identified where rents do not substantially differ. If all dummy coefficients which belong to a certain predictor are set to zero, the corresponding covariate is completely removed from the model.

Though penalization with adaptive weights has some nice asymptotic properties, simulation studies also showed that in the case of finite n particularly variable selection and clustering performance can even be further improved via ordinary least squares refitting of fused categories. A generalization of refitting is the so-called relaxed Lasso (Meinshausen, 2007), which puts a second penalty on (dummy) coefficients of fused categories. The disadvantage of relaxation is the second tuning parameter. In case of the Munich rent standard, sample sizes are so high that accurate (ordinary) least squares estimation is possible, which means that the second penalty parameter can be omitted.

In case of ordinal predictors computation is easily carried out by the lars algorithm (Efron et al., 2004), since the estimate is just an ordinary Lasso solution, if independent variables are split-coded. If predictors are nominal, we showed how procedures designed for ordinary Lasso problems can also be used to compute an approximate solution of the problem considered here.

Acknowledgements

This work was partially supported by DFG project TU62/4-1 (AOBJ: 548166).

Appendix

Proof of Proposition 1: We first show asymptotic normality, which closely follows Zou (2006) and Bondell and Reich (2009). Coefficient vector β is represented by $u = \sqrt{n}(\beta - \beta^*)$, i.e. $\beta = \beta^* + u/\sqrt{n}$, where β^* denotes the true coefficient vector. Then we also have $\hat{\beta} = \beta^* + \hat{u}/\sqrt{n}$, with

$$\hat{u} = \operatorname{argmin}_u \Psi_n(u),$$

where

$$\Psi_n(u) = \left(y - X\left(\beta^* + \frac{u}{\sqrt{n}}\right)\right)^T \left(y - X\left(\beta^* + \frac{u}{\sqrt{n}}\right)\right) + \frac{\lambda_n}{\sqrt{n}} J(u),$$

with

$$\begin{aligned} J(u) &= \sum_{i>j;i,j\neq 0} \sqrt{n} \frac{\phi_{ij}(n)}{|\hat{\beta}_i^{(LS)} - \hat{\beta}_j^{(LS)}|} \left|\beta_i^* - \beta_j^* + \frac{u_i - u_j}{\sqrt{n}}\right| \\ &+ \sum_{i>0} \sqrt{n} \frac{\phi_{i0}(n)}{|\hat{\beta}_i^{(LS)}|} \left|\beta_i^* + \frac{u_i}{\sqrt{n}}\right|. \end{aligned}$$

Furthermore, since $y - X\beta^* = \epsilon$, we have $\Psi_n(u) - \Psi_n(0) = V_n(u)$, where

$$V_n(u) = u^T \left(\frac{1}{n} X^T X\right) u - 2\frac{\epsilon^T X}{\sqrt{n}} u + \frac{\lambda_n}{\sqrt{n}} \widetilde{J}(u),$$

with

$$\begin{aligned} \widetilde{J}(u) &= \sum_{i>j;i,j\neq 0} \sqrt{n} \frac{\phi_{ij}(n)}{|\hat{\beta}_i^{(LS)} - \hat{\beta}_j^{(LS)}|} \left(\left|\beta_i^* - \beta_j^* + \frac{u_i - u_j}{\sqrt{n}}\right| - |\beta_i^* - \beta_j^*|\right) \\ &+ \sum_{i>0} \sqrt{n} \frac{\phi_{i0}(n)}{|\hat{\beta}_i^{(LS)}|} \left(\left|\beta_i^* + \frac{u_i}{\sqrt{n}}\right| - |\beta_i^*|\right). \end{aligned}$$

As given in Zou (2006) we will consider the limit behavior of $(\lambda_n/\sqrt{n})\widetilde{J}(u)$. If $\beta_i^* \neq 0$, then

$$|\hat{\beta}_i^{(LS)}| \to_p |\beta_i^*|, \text{ and } \sqrt{n}\left(\left|\beta_i^* + \frac{u_i}{\sqrt{n}}\right| - |\beta_i^*|\right) = u_i \operatorname{sgn}(\beta_i^*);$$

and similarly, if $\beta_i^* \neq \beta_j^*$,

$$|\hat{\beta}_i^{(LS)} - \hat{\beta}_j^{(LS)}| \rightarrow_p |\beta_i^* - \beta_j^*|, \text{ and}$$

$$\sqrt{n}\left(\left|\beta_i^* - \beta_j^* + \frac{u_i - u_j}{\sqrt{n}}\right| - |\beta_i^* - \beta_j^*|\right) = (u_i - u_j)\operatorname{sgn}(\beta_i^* - \beta_j^*);$$

Since by assumption $\phi_{ij}(n) \rightarrow q_{ij}$ $(0 < q_{ij} < \infty)$ and $\lambda_n/\sqrt{n} \rightarrow 0$, by Slutsky's theorem, we have

$$\frac{\lambda_n}{\sqrt{n}} \frac{\phi_{i0}(n)}{|\hat{\beta}_i^{(LS)}|}\sqrt{n}\left(\left|\beta_i^* + \frac{u_i}{\sqrt{n}}\right| - |\beta_i^*|\right) \rightarrow_p 0, \text{ and}$$

$$\frac{\lambda_n}{\sqrt{n}} \frac{\phi_{ij}(n)}{|\hat{\beta}_i^{(LS)} - \hat{\beta}_j^{(LS)}|}\sqrt{n}\left(\left|\beta_i^* - \beta_j^* + \frac{u_i - u_j}{\sqrt{n}}\right| - |\beta_i^* - \beta_j^*|\right) \rightarrow_p 0, \text{ respectively.}$$

This also makes clear that assumption $\lambda_n = O_p(\sqrt{n})$ is not enough. If $\beta_i^* = 0$ or $\beta_i^* = \beta_j^*$, however,

$$\sqrt{n}\left(\left|\beta_i^* + \frac{u_i}{\sqrt{n}}\right| - |\beta_i^*|\right) = |u_i|, \text{ and}$$

$$\sqrt{n}\left(\left|\beta_i^* - \beta_j^* + \frac{u_i - u_j}{\sqrt{n}}\right| - |\beta_i^* - \beta_j^*|\right) = |u_i - u_j|, \text{ respectively.}$$

Moreover, if $\beta_i^* = 0$ or $\beta_i^* = \beta_j^*$, due to $\sqrt{n}$-consistency of the ordinary least squares estimate (which is ensured by condition $n_i/n \rightarrow c_i$, $0 < c_i < 1\ \forall i$),

$$\lim_{n\rightarrow\infty} P(\sqrt{n}|\hat{\beta}_i^{(LS)}| \leq \lambda_n^{1/2}) = 1, \text{ resp. } \lim_{n\rightarrow\infty} P(\sqrt{n}|\hat{\beta}_i^{(LS)} - \hat{\beta}_j^{(LS)}| \leq \lambda_n^{1/2}) = 1,$$

since $\lambda_n \rightarrow \infty$ by assumption. Hence,

$$\frac{\lambda_n}{\sqrt{n}} \frac{\phi_{i0}(n)}{|\hat{\beta}_i^{(LS)}|}\sqrt{n}\left(\left|\beta_i^* + \frac{u_i}{\sqrt{n}}\right| - |\beta_i^*|\right) \rightarrow_p \infty, \text{ or}$$

$$\frac{\lambda_n}{\sqrt{n}} \frac{\phi_{ij}(n)}{|\hat{\beta}_i^{(LS)} - \hat{\beta}_j^{(LS)}|}\sqrt{n}\left(\left|\beta_i^* - \beta_j^* + \frac{u_i - u_j}{\sqrt{n}}\right| - |\beta_i^* - \beta_j^*|\right) \rightarrow_p \infty,$$

if $u_i \neq 0$, resp. $u_i \neq u_j$. That means, if for any $i, j > 0$ with $\beta_i^* = \beta_j^*$ or $\beta_i^* = 0$, $u_i \neq u_j$ or $u_i \neq 0$, respectively, then $(\lambda_n/\sqrt{n})\widetilde{J}(u) \rightarrow_p \infty$. The rest of the proof of part (a) is almost identical to Bondell and Reich (2009). Let X^* denote the design matrix corresponding to the correct structure, i.e. columns of dummy variables with equal coefficients are added and collapsed, and columns corresponding to zero coefficients are removed. Then $V_n^*(u_*)$ denotes the value of function V_n based on X^*. Since $\forall i\ n_i/n \rightarrow c_i\ (0 < c_i < 1)$,

$$\frac{1}{n} X^{*T} X^* \rightarrow C > 0 \text{ and } \frac{\epsilon^T X^*}{\sqrt{n}} \rightarrow_d w, \text{ with } w \sim N(0, \sigma^2 C).$$

Let $\theta_{\mathcal{C}^c}$ denote the vector of differences $\theta_{ij} = \beta_i - \beta_j$ which are truly zero, i.e. not from $\mathcal{C}$, and $u_{\mathcal{C}^c}$ the subset of entries of $\theta_{\mathcal{C}^c}$ which are part of u. By contrast, $u_{\mathcal{C}}$ denotes the subset of $\theta_{\mathcal{C}}$ which are in u. As given in Zou (2006), by Slutsky's theorem, $V_n(u) \rightarrow_d V(u)$ for every u, where

$$V(u) = \begin{cases} u_*^T C u_* - 2u_*^T w & \text{if } \theta_{\mathcal{C}^c} = 0 \\ \infty & \text{otherwise.} \end{cases}$$

Since $V_n(u)$ is convex and the unique minimum of $V(u)$ is $(C^{-1}w, 0)^T$, we have (cf. Zou, 2006; Bondell and Reich, 2009)

$$\hat{u}_{\mathcal{C}} \rightarrow_d C^{-1} w, \text{ and } \hat{u}_{\mathcal{C}^c} \rightarrow_d 0.$$

Hence, $\hat{u}_{\mathcal{C}} \rightarrow_d N(0, \sigma^2 C^{-1})$. By changing the reference category, i.e. changing the subset of entries of θ which are part of u, asymptotic normality can be proven for all pairwise differences in $\hat{\theta}_{\mathcal{C}}$.

To show the consistency part, we first note that $\lim_{n\rightarrow\infty} P((i,j) \in \mathcal{C}_n) = 1$, if $(i,j) \in \mathcal{C}$, follows from part (a). We will now show that if $(i,j) \notin \mathcal{C}$, $\lim_{n\rightarrow\infty} P((i,j) \in \mathcal{C}_n) = 0$. The proof is a modified version of the one given by Bondell and Reich (2009). Let $\mathcal{B}_n$ denote the (nonempty) set of pairs of indices $i > j$ which are in $\mathcal{C}_n$ but not in $\mathcal{C}$. Then we may choose reference category 0 such that $\hat{\beta}_q = \hat{\beta}_q - \hat{\beta}_0 > 0$ is the largest difference corresponding to indices from $\mathcal{B}_n$. Moreover, we may order categories such that $\hat{\beta}_1 \leq \ldots \leq \hat{\beta}_z \leq 0 \leq \hat{\beta}_{z+1} \leq \ldots \leq \hat{\beta}_k$. That means estimate $\hat{\beta}$ from (2) with penalty (3) is equivalent to

$$\hat{\beta} = \text{argmin}_{\{\beta_1 \leq \ldots \leq \beta_z \leq 0 \leq \beta_{z+1} \leq \ldots \leq \beta_k\}} \left\{ (y - X\beta)^T (y - X\beta) + \lambda_n J(\beta) \right\}$$

with

$$J(\beta) = \sum_{i>j; i,j\neq 0} \phi_{ij}(n) \frac{\beta_i - \beta_j}{|\hat{\beta}_i^{(LS)} - \hat{\beta}_j^{(LS)}|} + \sum_{i\geq z+1} \phi_{i0}(n) \frac{\beta_i}{|\hat{\beta}_i^{(LS)}|} - \sum_{i\leq z} \phi_{i0}(n) \frac{\beta_i}{|\hat{\beta}_i^{(LS)}|}.$$

Since $\hat{\beta}_q \neq 0$ is assumed, at the solution $\hat{\beta}$ this optimization criterion is differentiable with respect to β_q. We may consider this derivative in a neighborhood of the solution where coefficients which are set equal remain equal. That means, terms corresponding to pairs of indices which are not in C_n can be omitted, since they will vanish in $J(\hat{\beta})$. If x_q denotes the qth column of design matrix X, due to differentiability, estimate $\hat{\beta}$ must satisfy

$$\frac{Q_q'(\hat{\beta})}{\sqrt{n}} = \frac{2x_q^T(y - X\hat{\beta})}{\sqrt{n}} = A_n + D_n,$$

with

$$A_n = \frac{\lambda_n}{\sqrt{n}} \left(\sum_{j<q;(q,j)\in\mathcal{C}} \frac{\phi_{qj}(n)}{|\hat{\beta}_q^{(LS)} - \hat{\beta}_j^{(LS)}|} - \sum_{i>q;(i,q)\in\mathcal{C}} \frac{\phi_{iq}(n)}{|\hat{\beta}_i^{(LS)} - \hat{\beta}_q^{(LS)}|} \right)$$

and

$$D_n = \frac{\lambda_n}{\sqrt{n}} \sum_{j<q;(q,j)\in\mathcal{B}_n} \frac{\phi_{qj}(n)}{|\hat{\beta}_q^{(LS)} - \hat{\beta}_j^{(LS)}|}.$$

If β^* denotes the true coefficient vector, $Q_q'(\hat{\beta})/\sqrt{n}$ can be written as

$$\frac{Q_q'(\hat{\beta})}{\sqrt{n}} = \frac{2x_q^T(y - X\hat{\beta})}{\sqrt{n}} = \frac{2x_q^T X\sqrt{n}(\beta^* - \hat{\beta})}{n} + \frac{2x_q^T\epsilon}{\sqrt{n}}.$$

From part (a) and applying Slutsky's theorem, we know that $2x_q^T X\sqrt{n}(\beta - \hat{\beta})/n$ has some asymptotic normal distribution with mean zero, and $2x_q^T\epsilon/\sqrt{n}$ as well (by assumption, and applying the central limit theorem), cf. Zou (2006). Hence for any $\varepsilon > 0$, we have

$$\lim_{n\to\infty} P(Q_q'(\hat{\beta})/\sqrt{n} \leq \lambda_n^{1/4} - \varepsilon) = 1$$

Since $\lambda_n/\sqrt{n} \to 0$, we also know $\exists\, \varepsilon > 0$ such that $\lim_{n\to\infty} P(|A_n| < \varepsilon) = 1$. By assumption $\lambda_n \to \infty$; due to $\sqrt{n}$-consistency of the ordinary least squares estimate, we know that

$$\lim_{n\to\infty} P(\sqrt{n}|\hat{\beta}_q^{(LS)} - \hat{\beta}_j^{(LS)}| \leq \lambda_n^{1/2}) = 1,$$

if $(q, j) \in \mathcal{B}_n$. Hence

$$\lim_{n\to\infty} P(D_n > \lambda_n^{1/4}) = 1.$$

As a consequence

$$\lim_{n\to\infty} P(Q_q'(\hat{\beta})/\sqrt{n} = A_n + B_n) = 0.$$

That means if $(i, j) \notin \mathcal{C}$, also

$$\lim_{n\to\infty} P((i, j) \in \mathcal{C}_n) = 0.$$

Proof of Proposition 2: In Section 2.3 it is has been shown that the proposed estimate given an ordinal class structure is equivalent to a Lasso type estimate, if ordinal predictors are split-coded. That means, since $\phi_i(n) \to q_i$ $(0 < q_i < \infty)$ $\forall i$ by assumption, and employing Slutsky's Theorem, (the proof of) Theorem 2 about the adaptive Lasso by Zou (2006) can be directly applied. Condition $n_i/n \to c_i$, with $0 < c_i < 1$ $\forall i$, ensures that the ordinary least squares estimate is $\sqrt{n}$-consistent.

Proposition 3 *Suppose* $0 \leq \lambda < \infty$ *has been fixed, and all class-wise sample sizes* n_i *satisfy* $n_i/n \to c_i$*, where* $0 < c_i < 1$*. Then weights* $w_{ij} = \phi_{ij}(n)$*, with* $\phi_{ij}(n) \to q_{ij}$ $(0 < q_{ij} < \infty)$ $\forall i, j$*, ensure that estimate* $\hat{\beta}$ *from (2) with penalty (3) is consistent, i.e.* $\lim_{n\to\infty} P(||\hat{\beta} - \beta^*||^2 > \epsilon) = 0$ *for all* $\epsilon > 0$*.*

Proof: If $\hat{\beta}$ minimizes $Q_p(\beta)$ from (1), then it also minimizes $Q_p(\beta)/n$. The ordinary least squares estimator $\hat{\beta}^{(LS)}$ minimizes $Q(\beta) = (y - X\beta)^T(y - X\beta)$, resp. $Q(\beta)/n$. Since $Q_p(\hat{\beta})/n \to_p Q(\hat{\beta}^{(LS)})/n$ and $Q_p(\hat{\beta})/n \to_p Q(\hat{\beta})/n$, we have $Q(\hat{\beta})/n \to_p Q(\hat{\beta}^{(LS)})/n$. Since $\hat{\beta}^{(LS)}$ is the unique minimizer of $Q(\beta)/n$, and $Q(\beta)/n$ is convex, we have $\hat{\beta} \to_p \hat{\beta}^{(LS)}$, and consistency follows from consistency of the ordinary least squares estimator $\hat{\beta}^{(LS)}$, which is ensured by condition $n_i/n \to c_i$, with $0 < c_i < 1$ $\forall i$.

Proposition 4 *If restriction* $\theta_{ij} = \theta_{i0} - \theta_{j0}$ *is represented by* $A\theta = 0$*, define* $\hat{\theta}_{\gamma,\lambda} = argmin_\theta\{(y - Z\theta)^T(y - Z\theta) + \gamma(A\hat{\theta})^T A\hat{\theta} + \lambda|\theta|\}$*, where* $\theta = (\theta_{10}, \ldots, \theta_{k,k-1})^T$ *and* $|\theta| = \sum_{i>j} |\theta_{ij}|$*. Then with* $\gamma > 0$ *and* $\lambda \geq 0$*,* $\Delta_{\gamma,\lambda} = (A\hat{\theta}_{\gamma,\lambda})^T A\hat{\theta}_{\gamma,\lambda}$ *is bounded above by*

$$\Delta_{\gamma,\lambda} \leq \frac{\lambda(|\hat{\theta}^{(LS)}| - |\hat{\theta}_{0,\lambda}|)}{\gamma},$$

where $\hat{\theta}^{(LS)}$ *denotes the least squares estimate (i.e.* $\lambda = 0$*) where* $A\hat{\theta}^{(LS)} = 0$ *holds.*

Proof: Obviously, for all $\gamma > 0$ and $\lambda \geq 0$,

$$(y - Z\hat{\theta}_{\gamma,\lambda})^T(y - Z\hat{\theta}_{\gamma,\lambda}) + \lambda|\hat{\theta}_{\gamma,\lambda}| + \gamma\Delta_{\gamma,\lambda} \leq (y - Z\hat{\theta}^{(LS)})^T(y - Z\hat{\theta}^{(LS)}) + \lambda|\hat{\theta}^{(LS)}|.$$

Since also

$$(y - Z\hat{\theta}_{0,\lambda})^T(y - Z\hat{\theta}_{0,\lambda}) + \lambda|\hat{\theta}_{0,\lambda}| \leq (y - Z\hat{\theta}_{\gamma,\lambda})^T(y - Z\hat{\theta}_{\gamma,\lambda}) + \lambda|\hat{\theta}_{\gamma,\lambda}|,$$

and

$$(y - Z\hat{\theta}_{0,\lambda})^T(y - Z\hat{\theta}_{0,\lambda}) \geq (y - Z\hat{\theta}^{(LS)})^T(y - Z\hat{\theta}^{(LS)}),$$

we have

$$\gamma\Delta_{\gamma,\lambda} \leq \lambda(|\hat{\theta}^{(LS)}| - |\hat{\theta}_{0,\lambda}|).$$

References

Bondell, H. D. and B. J. Reich (2009). Simultaneous factor selection and collapsing levels in anova. *Biometrics 65*, 169–177.

Bühlmann, P. and B. Yu (2003). Boosting with the L2 loss: Regression and classification. *Journal of the American Statistical Association 98*, 324–339.

Candes, E. and T. Tao (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics 35*, 2313–2351.

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

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*, 1348–1360.

Karatzoglou, A., A. Smola, K. Hornik, and A. Zeileis (2004). kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software 11*(9), 1–20.

Kneib, T., F. Heinzl, A. Brezger, and D. Sabanés Bové (2009). *BayesX: R Utilities Accompanying the Software Package BayesX*. R package version 0.2.

Land, S. R. and J. H. Friedman (1997). Variable fusion: A new adaptive signal regression method. Technical report 656, Department of Statistics, Carnegie Mellon University Pittsburg.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics & Data Analysis 52*, 374–393.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

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

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Kneight (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society B 67*, 91–108.

Walter, S. D., A. R. Feinstein, and C. K. Wells (1987). Coding ordinal independent variables in multiple regression analysis. *American Journal of Epidemiology 125*, 319–323.

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

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*, 1418–1429.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B 67*, 301–320.

predictor	label	coefficient
intercept		12.597
urban district	14, 16, 22, 24	-1.931
	11, 23	-1.719
	7	-1.622
	8, 10, 15, 17, 19, 20, 21, 25	-1.361
	6	-1.061
	9	-0.960
	13	-0.886
	2, 4, 5, 12, 18	-0.671
	3	-0.403
year of construction	1920s	-1.244
	1930s, 1940s	-0.953
	1950s	-0.322
	1960s	0.073
	1970s	0.325
	1980s	1.121
	1990s, 2000s	1.624
number of rooms	4, 5, 6	-0.502
	3	-0.180
	2	0.000
quality of residential area	good	0.373
	excellent	1.444
floor space (m^2)	$[140, \infty)$	-4.710
	$[90, 100)$, $[100, 110)$, $[110, 120)$, $[120, 130)$, $[130, 140)$	-3.688
	$[60, 70)$, $[70, 80)$, $[80, 90)$	-3.443
	$[50, 60)$	-3.177
	$[40, 50)$	-2.838
	$[30, 40)$	-1.733
hot water supply	no	-2.001
central heating	no	-1.319
tiled bathroom	no	-0.562
suppl. equipment in bathroom	yes	0.506
well equipped kitchen	yes	1.207

Table 3: Estimated regression coefficients for Munich rent standard data using adaptive weights with refitting, and (cross-validation score minimizing) $s/s_{\max} = 0.61$.