The Annals of Applied Statistics
2010, Vol. 4, No. 4, 2150–2180
DOI: 10.1214/10-AOAS355
© Institute of Mathematical Statistics, 2010

SPARSE MODELING OF CATEGORIAL EXPLANATORY VARIABLES

BY JAN GERTHEISS[1] AND GERHARD TUTZ

Ludwig-Maximilians-Universität Munich

Shrinking methods in regression analysis are usually designed for metric predictors. In this article, however, shrinkage methods for categorial predictors are proposed. As an application we consider data from the Munich rent standard, where, for example, urban districts are treated as a categorial predictor. If independent variables are categorial, some modifications to usual shrinking procedures are necessary. 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. Besides applying them to the Munich rent standard, methods are illustrated and compared in simulation studies.

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-dimensional 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)].

This article provides a regularized regression analysis of Munich rent standard data. All larger German cities publish so-called rent standards for having guidelines available to tenants, landlords, renting advisory boards and experts. These rent standards are used, in particular, to determine the local comparative rent. For the composition of 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 come 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 1. The data can be downloaded from the data archive of the Department of Statistics at the University of

Received June 2009; revised April 2010.

[1]Supported in part by DFG project TU62/4-1 (AOBJ: 548166).

Key words and phrases. Categorial predictors, fused lasso, ordinal predictors, rent standard, variable fusion.

TABLE 1
Explanatory variables for monthly rent per square meter

Urban district	Nominal, labeled by numbers $1, \ldots, 25$
Year of construction	Given in ordered classes [1910, 1919], [1920, 1929], ...
Number of rooms	Taken as ordinal factor with levels $1, 2, \ldots, 6$
Quality of residential area	Ordinal, with levels "fair," "good," "excellent"
Floor space (in m^2)	Given in ordered classes $(0, 30)$, $[30, 40)$, $[40, 50)$, $\ldots$, $[140, \infty)$
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)

Munich (http://www.stat.uni-muenchen.de/service/datenarchiv). The direct link is found there.

For example, the urban district is given as a nominal predictor with 25 possible values. The decade of construction can be interpreted as ordinal with 10 levels. Usually such data are analyzed via standard linear regression modeling, with (for example) dummy coded categorial explanatory variables. In the present situation such modeling is possible, since the number of observations (2053) is quite high. Nevertheless, from the viewpoint of interpretation, model selection is desired with the focus on reducing model complexity.

In selection problems for categorical predictors as in the Munich rend data example, 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 of 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.

When investigating which of the 25 urban districts of Munich are to be distinguished with respect to the local rent, the number of possible combinations is huge. If only urban districts are used as categorial predictor in a regression model to explain the monthly rent, and districts are potentially fused (without further

restrictions), the number of possible models—which just follow from different fusion results—is greater than 10^{18}. In cases like that—that is, when the number of possible models is large—regularization techniques which induce sparsity are a promising approach for model selection. The extent of regularization—and hence sparsity—is typically controlled by a tuning parameter. Via choosing this parameter, the model is also implicitly selected.

Most of the regularization techniques developed so far focus on the selection of variables in the case where the effect of one variable is determined by one coefficient. That means coefficients are selected 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 and model selection. This is different when categorical variables have to be included because then a whole group of coefficients refers to one variable.

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 $\mathrm{Var}(\epsilon) = \sigma^2$. 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.

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 coefficients that refer to one variable. However, while the Group Lasso only attacks the problem of factor selection, for categorical predictor variables with many categories a useful strategy is to (additionally) search for clusters of categories with similar effects. As already described, in the presented application (among other things) we try to model the influence of the urban district where a person lives on the rent she/he has to pay. In Figure 1 a map of Munich is drawn with color coded urban districts. Colors correspond to dummy coefficients if an ordinary least squares model is fitted with (dummy coded) explanatory variables from Table 1 and response monthly rent per square meter (in

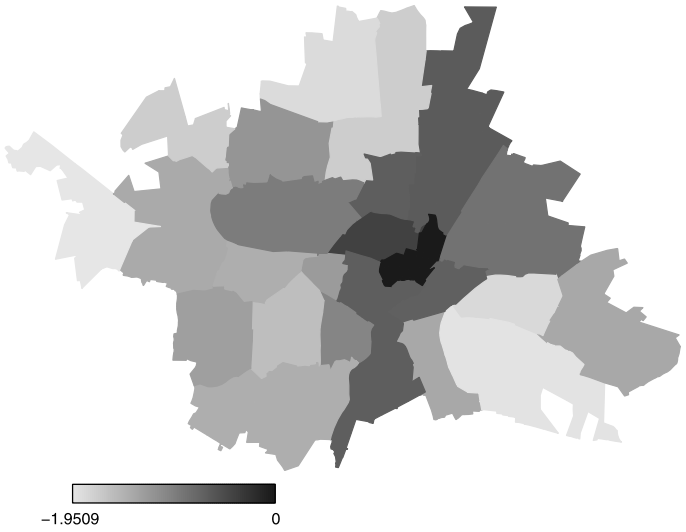

FIG. 1. *Map of Munich indicating urban districts*; *colors correspond to estimated dummy coefficients if an ordinary least squares model is fitted with predictors from Table* 1 *and response monthly rent per square meter* (*in Euro*).

Euro). Some districts are hard to distinguish. That means it can be expected that not all districts do differ substantially. If an ordinary least squares model is fitted, however, estimated dummy coefficients (almost surely) differ. Therefore, the aim is to combine districts which (on average) do not substantially differ in terms of rent per square meter. Generally speaking, that means 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. Therefore, in a regression model corresponding dummy coefficients should be equal. Since, however, the number of possible clustering results—and hence the number of models—tends to be very large (as already mentioned), model selection via regularization is quite attractive.

Clustering or *fusion* of metric predictors may, for example, be obtained by so-called Variable Fusion [Land and Friedman (1997)] and the Fused Lasso proposed by Tibshirani et al. (2005). If predictors can be ordered, by putting a L_1-penalty on differences of adjacent coefficients many of these differences are set to zero, yielding a piecewise constant coefficient function. Recently, Bondell and Reich (2009) adapted this methodology for factor selection and level fusion in ANOVA, to obtain dummy coefficients that are constant over some of the categories. The main focus of Bondell and Reich (2009), however, was on ANOVA typical identification of differences, not on model building as in our case, where prediction accuracy is also an important aspect. So in the following the method is reviewed and adapted to regression type problems. Some modifications are proposed and an

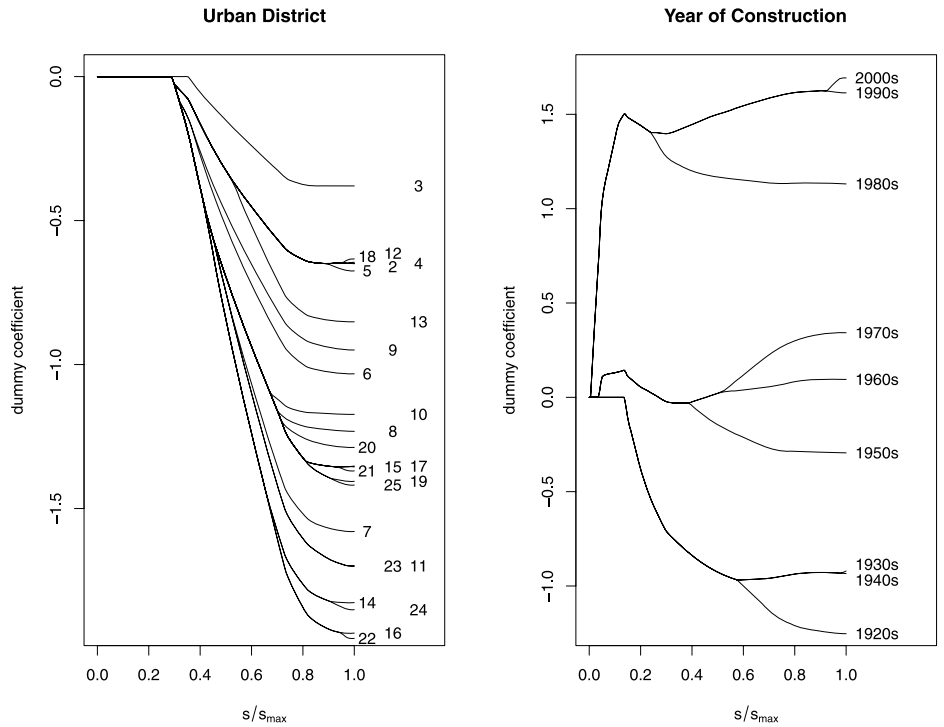

FIG. 2. *Paths of dummy coefficients of two categorial predictors obtained by the regularization technique proposed here for the analysis of rent standard data.*

approximate solution is presented which allows for easy computation of coefficient paths. In addition, the method is adapted to the modeling of ordinal predictors.

Figure 2 shows paths of dummy coefficients for the rent data obtained by the method used in this article. The coefficients at value $s/s_{\max} = 1$ correspond to the ordinary least squares model. It is seen that with decreasing tuning parameter s, categories are successively fused, that is, coefficients are set equal. Besides the urban district, several other covariates are given, among them the (categorized) year of construction. Corresponding paths of dummy coefficients are also shown in Figure 2.

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 design matrix X, coefficient vector β and penalty $J(\beta)$; y contains the observed response values. The estimate of β is given by

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\{Q_p(\beta)\}. \tag{2}$$

The decisive point is a suitable choice of penalty $J(\beta)$. We start with the case of one categorial explanatory 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 like selection and clustering consistency 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 $\hat{\beta}_j^{(LS)}$. For details see Proposition 1 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 Table 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; see the Appendix for details.

2.3. *Computational issues.* For the actual application of the proposed method a fitting algorithm is needed. For that purpose 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). For estimation purposes we consider transformed parameters $\theta_{ij} = \beta_i - \beta_j$ which yield vector $\theta = (\theta_{10}, \theta_{20}, \ldots, \theta_{k,k-1})^T$. 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, that is,

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

with

$$\theta_{ij}^+ \geq 0, \qquad \theta_{ij}^- \geq 0,$$

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)].

The problem with quadratic programming is that the solution can only be computed for a single value s. To obtain a coefficient path (as in Figure 2), the procedure needs to be applied repeatedly. Moreover, when applying the method to our data, we found numerical problems, especially when s was small. To attack these problems, we propose an approximate solution which 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

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

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 higher γ must also 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

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

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 of U 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, Feinstein and Wells (1987)], and ordinary Lasso estimation is applied. Split-coding means that dummies $\widetilde{x}_i$ are defined by splits at categories $i = 1, \ldots, k$, that is,

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

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, Feinstein and Wells (1987) intended the use of classical tests for such identification of substantial "*between-strata differences*."

2.4. *Multiple inputs.* In our application, as usual in statistical modeling, a set of (potential) regressors is available (see Table 1) and only the relevant predictors should be included into the model. 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 in general, say, $x_1, \ldots, x_p$, with levels $0, \ldots, k_l$ for variable x_l ($l = 1, \ldots, p$, and fixed 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

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

depending on the scale level of predictor x_l. The first expression refers to nominal covariates, the second to ordinal ones.

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 that 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 3 in the Appendix, or when predictors differ in the number of levels, as in the rent standard application (see Table 1). 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 (6) 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 considered, 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}.

2.6. *Refitting procedures.* The most attractive features of the methods described above are variable selection and clustering. However, due to penalization, estimates are obviously biased. In the usual ANOVA case, this is not a problem, since the focus is on the identification of differences, and not on quantification. In our case—as in regression analysis in general—we are also interested in parameter estimation and prediction accuracy. In order to reduce the bias, refitting procedures have been proposed by several authors, for example, by Efron et al. (2004) under the name "Lars-OLS hybrid," or by Candes and Tao (2007) as "Gauss-Dantzig Selector." In our setting, that means that the penalty in (1) is only used for variable selection and clustering. After the identification of relevant predictors and clusters, parameters are refitted by ordinary least squares. If variable selection and clustering are based on the already mentioned adaptive weights, asymptotic behavior is obtained which is similar to the nonrefitting case; for details, see the remarks on

Proposition 1 in the Appendix. However, before we apply the refitting method to the rent data (where $n < \infty$), its effect is also tested in simulation studies (see Section 3.2).

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 $i = 0, \ldots, 8$ classes. In class i the response is $N(\mu_i, 4)$-distributed, where the means form three distinct groups of categories, that is, $\mu_0 = \mu_1 = \mu_2$, $\mu_3 = \mu_4 = \mu_5$, $\mu_6 = \mu_7 = \mu_8$. Figure 3 (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 nonadaptive 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

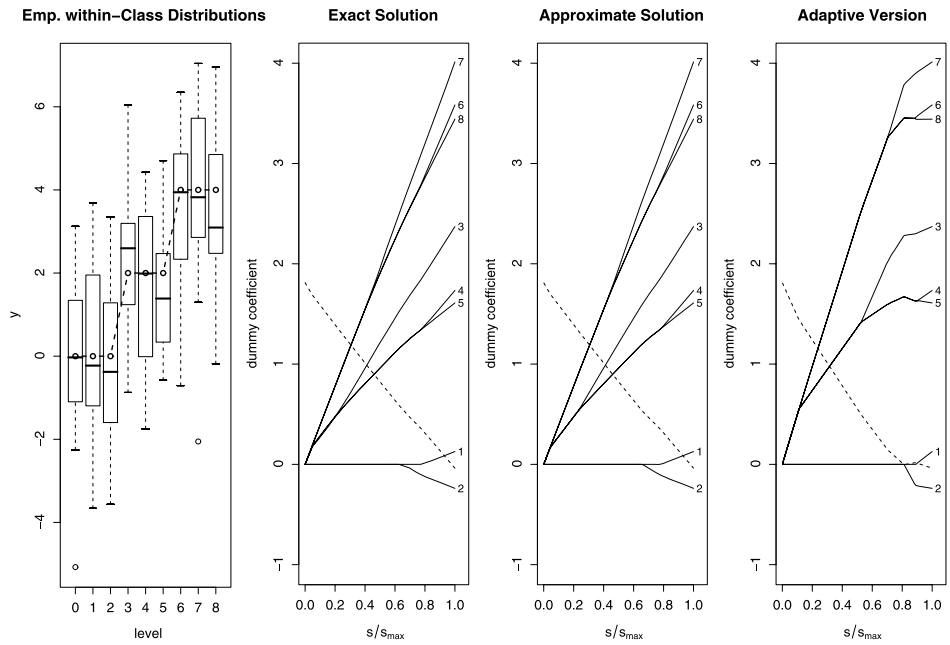

FIG. 3. *Empirical within-class distributions* (left), *exact and approximate coefficient paths* (middle), *as well as results of the adaptive version* (right); *constant* α *is marked by the dashed line.*

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 3, 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 nonadaptive version, bias toward zero is much smaller at the point of perfect grouping.

In a second scenario, settings and data visualized in Figure 3 (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 4. Even for the nonadaptive version (left), grouping is quite good. Moreover, before optimal grouping is reached, bias toward 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.

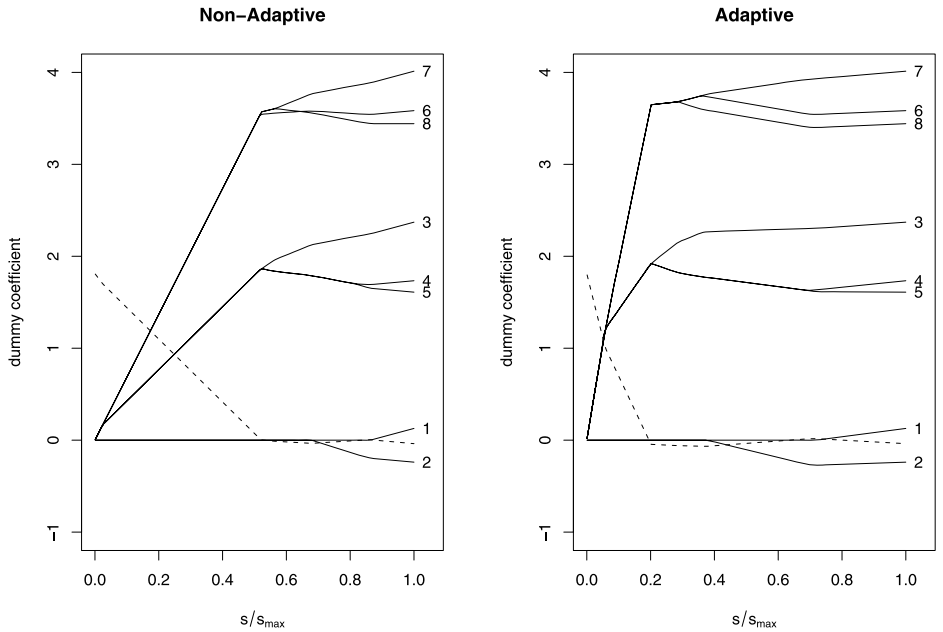

FIG. 4. *Paths of dummy coefficients for data as in Figure 3, but assuming an ordinal class structure, nonadaptive* (left) *and adaptive* (right) *version*; *constant* α *is marked by the dashed line.*

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 nonzero 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 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 nonadaptive weights. In addition, the effect of taking into account marginal class frequencies $n_i^{(l)}$ is investigated, which means we check what happens if $((n_i^{(l)} + n_j^{(l)})/n)^{1/2}$ is omitted in (6). Moreover, refitting is tested (as already mentioned), that is, the penalization is only used for variable selection and clustering. After the identification of relevant predictors and clusters, parameters are refitted by ordinary least squares.

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 nonzero dummy coefficients of predictor x_l and 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 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 nonzero; if clustering is considered, it means that a difference within a nonnoise factor which is truly zero is set to nonzero. By contrast, "false negative" means that all dummy coefficients of a truly relevant factor are set to zero, or that a truly nonzero difference is set to zero, respectively. Figure 5 shows the results for 100 simulation runs; labels are defined in Table 2.

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 the performance of the ordinary least squares (OLS) estimate is also 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

TABLE 2
Definition of labels used in Figures 5 and 6

adapt	Adaptive version, i.e., weighting terms $\lvert\hat{\beta}_i^{(LS)} - \hat{\beta}_j^{(LS)}\rvert^{-1}$ are used
stdrd	Standard (nonadaptive) 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

OLS. Concerning FPR and FNR, differences between pure adaptive/nonadaptive approaches and refitting are caused by the fact that not necessarily the same models are selected, because in cross-validation already refitted coefficients are used.

As already illustrated by Bondell and Reich (2009) and supported by Propositions 1 and 3 in the Appendix, 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 more distinct. However, the most important effect of refitting is on variable selection and clustering—in both the adaptive and the nonadaptive 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 6) 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 before. 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, given a regression problem, refitting has the potential to distinctly improve selection and clustering results in the

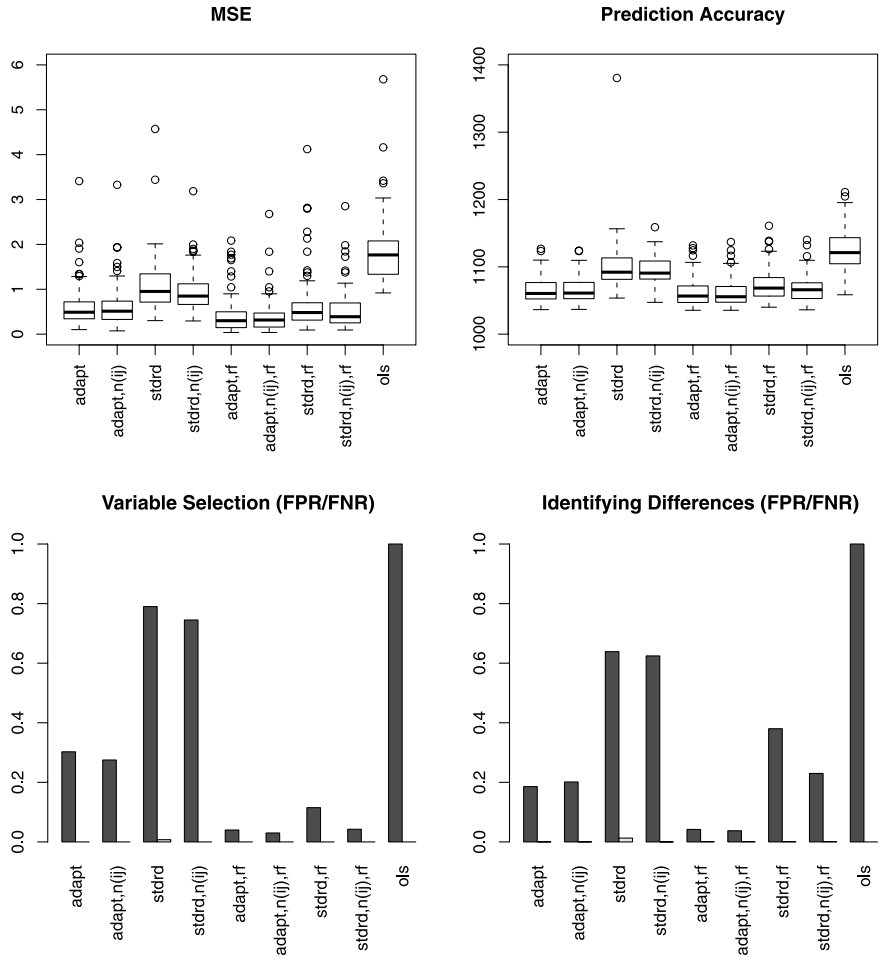

FIG. 5. *Evaluation of adaptive and nonadaptive* (*standard*) *as well as refitting* (*rf*) *approaches, taking into account class sizes* (n_i, n_j) *or not, for comparison the results for the ordinary least squares* (*ols*) *estimator are also given*; *considered are the mean squared error of parameter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

$n < \infty$ case, while providing accurate parameter estimates (if n is not too small compared to p). Therefore, it can be assumed to be a suitable approach for our regression analysis. Moreover, taking into account marginal class frequencies seems to (slightly) improve estimation results.

4. Regularized analysis of Munich rent standard data. For the estimation of regression coefficients with predictors from Table 1, we consider the approaches which performed best in the previous section; more concrete, both the adaptive

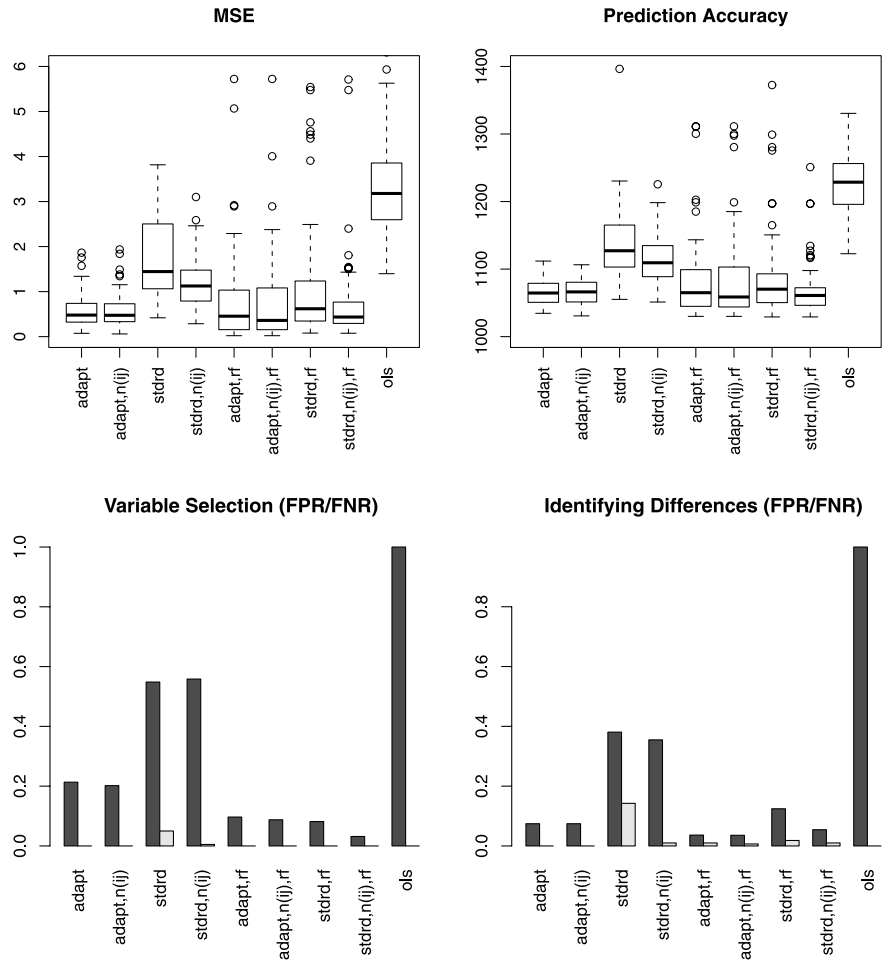

FIG. 6. *Evaluation of different approaches in the presence of many noise variables*: *adaptive and nonadaptive* (*standard*) *as well as refitting* (*rf*), *taking into account class sizes* (n_i, n_j) *or not*, *for comparison also the ordinary least squares* (*ols*) *estimator*; *considered are the 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*.

as well as the standard (nonadaptive) version remain candidates, but each with refitting only and taking marginal class frequencies into account. In the following we first analyze the data and then evaluate the performance of the approach (using the rent data) comparing it to ordinary least squares and Group Lasso estimates, which do not provide variable selection and/or clustering of categories.

4.1. *Data analysis*. In the considered application more than 2000 observations are available for the estimation of 58 regression parameters. Thus, OLS estimation

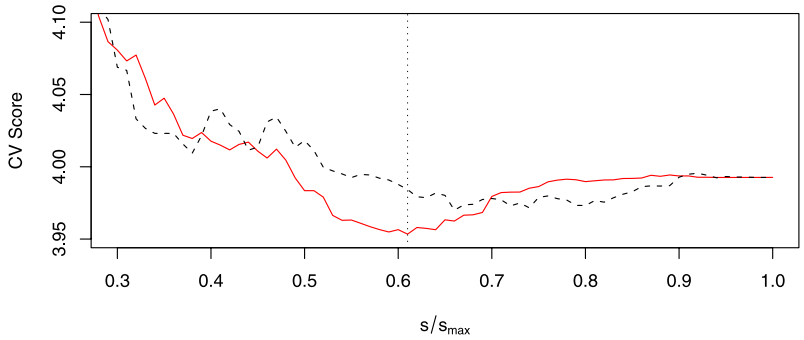

FIG. 7. *Cross-validation score as a function of* $s/s_{\max}$ *if refitting with standard* (*dashed black*) *or adaptive* (*solid red*) *weights is used for the analysis of Munich rent standard data.*

works, and (in the light of the simulation study before) it is to be expected that refitting distinctly improves estimation accuracy as well as variable selection and clustering performance of the proposed penalized approach.

Figure 7 shows the (10-fold) cross-validation score as a function of $s/s_{\max}$, for the refitted model with nonadaptive (dashed black) as well as adaptive weights (solid red). It is seen that penalized estimates, in particular, refitting with adaptive weights, may improve the ordinary least squares estimate (i.e., $s/s_{\max} = 1$) in terms of prediction accuracy. It is not surprising that adaptive weights show better performance than nonadaptive ones, since sample size is high, 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 7). 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. It is interesting that rents of houses constructed shortly before the Second World War and those constructed within or shortly after the war do not substantially differ.

The biggest cluster, which contains 8 categories, is formed within the 25 districts. A map of Munich with color coded clusters (Figure 8) 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. After inspection of OLS estimates (e.g., in Figure 2), it could be expected that rather cheap districts 14 and 24 are fused. It was not clear, however, if they are additionally collapsed with any other districts, and if so, whether fused with $\{16, 22\}$ or $\{11, 23\}$. Based on our regularized analysis, it can now be stated with good reason that rents in districts 14, 16, 22 and 24 are comparatively low and do not substantially differ, which is in agreement with judgements from experts and feelings of

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

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

laymen, because Munich's deprived areas are primarily located in these (nonadjacent) districts. The cluster that contains district 12, however, partly contradicts experiences of experts and tenants. The problem is that this district is very large and reaches from the city center to the outskirts in the north. So very expensive

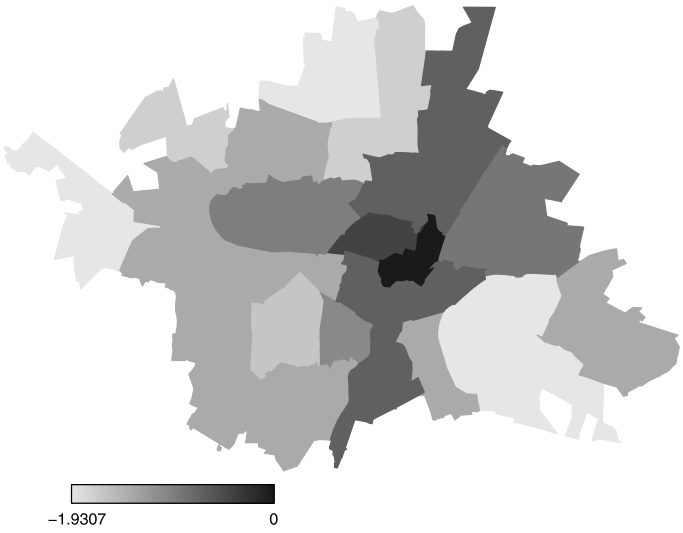

FIG. 8. *Map of Munich indicating clusters of urban districts*; *colors correspond to estimated dummy coefficients from Table* 3.

flats which are close to the city center are put together with cheaper ones on the outskirts. But, on average, rents are rather high in this district, which causes it to be clustered with other expensive but more homogeneous areas. In an ordinary least squares model, district 12 is even identified as belonging to the three most expensive districts (see also Figures 1 and 2). Penalized estimation ranks it only among the top seven. But it should be noted that 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, 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 found 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 effect that for larger flats the rent per square meter is lower. Starting with small apartments, the decrease of rents is quite apparent (between ca. 20 and 60 m^2), then it is much slower. Between 90 and 140 m^2, for example, no differences are identified with respect to rent per square meter. The fact that the covariate which indicates the number of rooms is not completely excluded from the model, although the flat's floor space is also considered, shows that there are dependencies between rent and the number of rooms which do not only refer to the flat's size. If covariate floor space is held constant, but the number of rooms is increased, rents tend to go down.

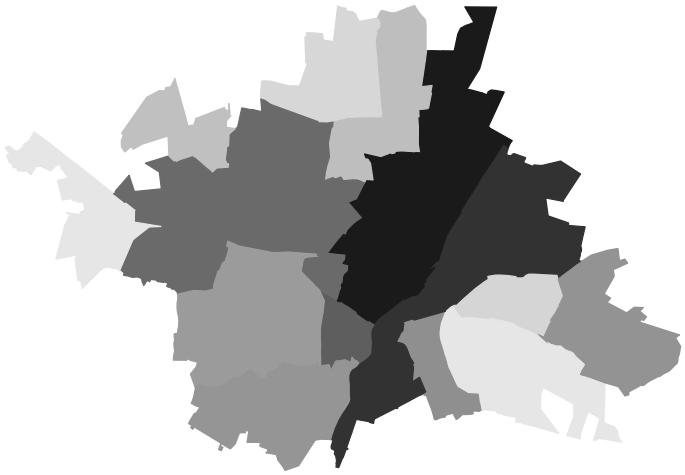

FIG. 9. *Map of Munich indicating clusters of urban districts, if just differences between dummy coefficients of neighboring districts are penalized.*

All in all, the selected model has 32 degrees of freedom, that is, 32 unique nonzero coefficients (including the intercept), which means that the complexity of the unrestricted model (58 df) is reduced by about 45%.

4.2. *Using spatial information.* A possible alternative to treating the urban district as a nominal predictor is to include geographical information. One can use distance measures or neighborhood effects when looking for clusters. A simple approach we used is to penalize—in analogy to the ordinal predictor case—only differences between dummy coefficients of neighboring districts. In Figure 9 a map of Munich is shown which results from such neighborhood penalization. One problem with this map is that district 12 (which reaches from the center to the north) is now fused with three expensive adjacent districts in the center. We also fitted a more advanced neighborhood weighting scheme, which uses the length of the boundary between the corresponding districts as weights. Then the difference between district 12 and a neighboring (and cheap) district in the north would get some higher weight. However, even that modification does not solve the second problem linked with that kind of spatial information based regularization: Two nonadjacent districts will not be fused if they are not also fused with a whole set of districts building a chain that connects them. In Figure 9 the two light-colored districts in the west and southeast (22 and 16) seem quite similar. In contrast to Figure 8 and Table 3, however, they are not fused. The corresponding difference of dummy coefficients is about 0.007—close to, but not exactly zero. Generally speaking, districts which are not neighbors may also be quite similar. Therefore, fusion of such districts should be possible, too. Hence, we prefer an approach like

our initial modeling where all pairwise differences of districts' dummy coefficients have been penalized.

A more general procedure to include spatial information is to incorporate this information into the weights w_{ij} in (3). For that purpose weights may be additionally multiplied by factors ζ_{ij}, where ζ_{ij} contains spatial information. As long as $0 < \zeta_{ij} < \infty$ for all i, j, consistency as given in Proposition 1 is not affected, and all pairwise differences are still penalized as desired in our application. Factor ζ_{ij} can, for example, be defined as a decreasing function of the distance between districts i and j. A special case of such an approach is to penalize only differences of neighboring districts as already done before. This, however, does not guaranty $\zeta_{ij} > 0$ for all i, j, and did not produce good results in our application (as shown above). Furthermore, it seems sensible to assume that differences (concerning rents) between the city center and the outskirts tend to be larger than differences between outskirts in the west and the east of a city. So for defining ζ_{ij} we may use the information whether a district is rather central or peripheral. If ς_i denotes the distance of (the center of) district i to the city center (in km), we define

$$\zeta_{ij} = K\left(\frac{\varsigma_i - \varsigma_j}{h}\right),$$

with a fixed kernel K and bandwidth h. For K we use the Epanechnikov kernel, and $h = 15$ (km), which is roughly the radius of the smallest circle around the city center which contains the whole city of Munich. Incorporating spatial information this way, however, yields exactly the same clustering as already given in Table 3, where urban districts have just been treated as a nominal predictor. So we can keep interpretations given above, and will just use the districts' categorial character in the following. The finding that results do not change if ζ_{ij} are included is obviously due to the fact that weights are decisively influenced by the ols terms (see Proposition 1 in the Appendix).

4.3. *Evaluation of prediction accuracies and sparsity.* The proposed methods provide clustering of categories, which results in a sparser model and facilitates interpretation in the considered application. In order to evaluate their actual prediction accuracies, we perform repeated random splitting of the data into training and test sets. That means coefficients are estimated on the training data (including determination of tuning parameters and weights), and then used to predict the test data. As test set size we choose 100, and the procedure is independently repeated 100 times. Results are shown in Figure 10. Performance is measured in terms of the mean squared error of prediction (MSEP). We investigate the refitted adaptive as well as the nonadaptive version of the presented regularization technique. For comparison, we also give prediction accuracies for the (most complex) ordinary least squares model, and for Group Lasso estimates as proposed by Yuan and Lin (2006) or Meier, Van de Geer and Bühlmann (2008). In the case of ordinal

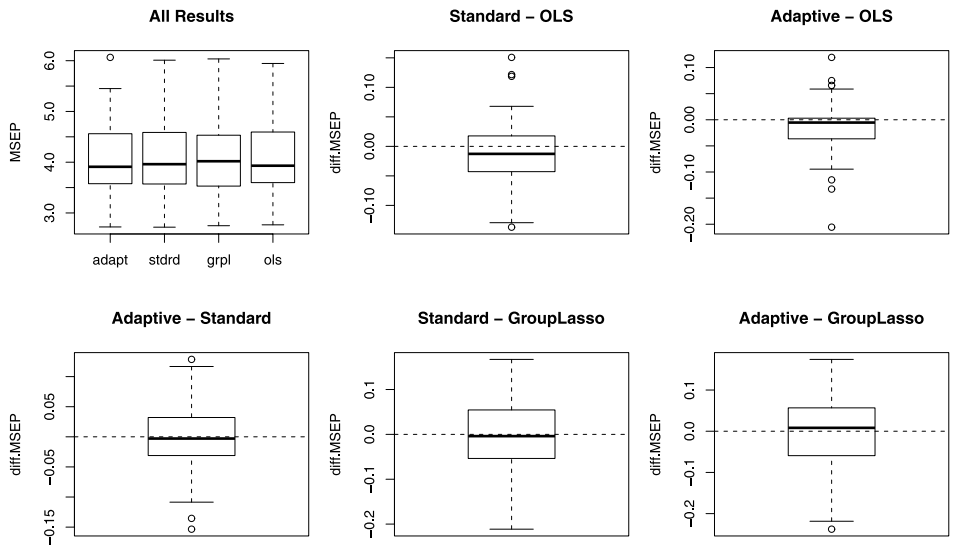

FIG. 10. *Prediction performance of the refitted adaptive* (*adapt*) *as well as nonadaptive* (*stdrd*) *sparse modeling of Munich rent standard data*, *the Group Lasso* (*grpl*) *and ordinary least squares* (*ols*) *fitting*; *all results* (top left) *as well as selected pairwise comparisons*.

predictors, the usual within groups simple ridge penalty is replaced by a difference penalty as proposed in Gertheiss and Tutz (2009) and Gertheiss et al. (2009). For practical estimation of Group Lasso estimates R add-on package `grplasso` [Meier (2007)] was used.

The first plot (top left) in Figure 10 shows boxplots of the observed MSEPs for all four methods. It is seen that all methods perform almost equally. This finding is confirmed by pairwise comparisons. Since in each iteration MSEPs of different methods are observed on the same test data, we report pairwise differences of corresponding MSEPs. Boxplots which tend to be below zero indicate superior performance of the method which is quoted first—and vice versa. It just seems that the proposed adaptive version is slightly superior to the ordinary least squares estimate. Between the different penalization techniques—the presented sparse modeling (adaptive/nonadaptive) and the Group Lasso—there can hardly be observed any difference concerning prediction accuracy on the rent standard data.

It is a quite positive result, however, that prediction accuracies of the considered methods are almost identical, because sparsity is the great advantage of the modeling which has been applied above to analyze the data. While the ordinary least squares model has 58 degrees of freedom, the (refitted adaptive) model which has been chosen on the basis of all data just has 32 df (see Table 3). In Figure 11 we now show kernel density estimates of the model complexities observed during random splitting of the data. It is seen that the adaptive models (solid red) tend to have less degrees of freedom than the nonadaptive version (dashed black). But

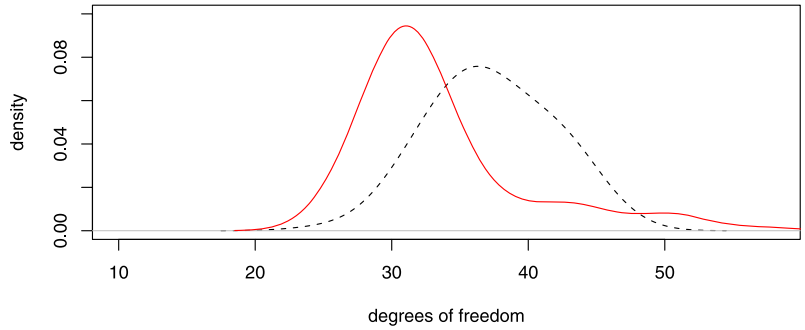

FIG. 11. *Kernel density estimates of chosen degrees of freedom for the adaptive* (*solid red*) *and nonadaptive* (*dashed black*) *model after repeated random splitting of Munich rent standard data.*

also the latter is far away from the 58 df of the OLS model. Furthermore, the Group Lasso can only perform variable selection, but no clustering of single categories. However, in each of the considered random splits all factors were selected (not shown), which means that none of the dummy coefficients estimated by the Group Lasso were set to zero. Hence, with the (via cross-validation) chosen tuning parameters, the effect of the Group Lasso penalty was just shrinkage/smoothing of groups of dummy coefficients, but no variable selection. That means in the case of the analyzed rent standard data the Group Lasso does not result in a sparser parametrization than the OLS model. In summary, on the rent data our model can be expected to be as accurate as competing models, while complexity is distinctly reduced and interpretability is increased.

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 penalties were investigated. Given just nominal covariates, all pairwise differences of dummy 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 their influence on the response. In the analysis of the rent standard data this meant that, for example, certain urban districts were identified where rents do not substantially differ on average. If all dummy coefficients which belong to a certain predictor are set to zero, the corresponding covariate is completely removed from the model.

In particular, it was shown that the usually applied (and accurate) ordinary least squares fitting of rent standard data can be improved if categorial predictors are adequately penalized. Such improvement is primarily in terms of interpretability and model complexity. Via repeated random splitting of the data at hand, it could be shown that model complexity could be reduced by about 40–50% while prediction

accuracies did not deteriorate. As simulation studies showed, in cases of smaller sample sizes estimation and prediction accuracies can also be distinctly improved via the presented L_1-difference-penalization.

An alternative approach would be to apply clustering methods on ols estimates, which may give similar results for the considered rent data (see Figure 2, though it is not clear, for example, in which way districts {14, 24} should be fused with other districts). However, this would be a two-step procedure and hence less elegant than a penalty based regularization technique. Moreover, in case of smaller sample sizes it would severely suffer from instability of ols estimates.

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 the 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 the case of ordinal predictors, computation of the proposed estimator 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 approximate coefficient paths.

APPENDIX

Asymptotic properties for the unordered case. 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 nonzero, and $\mathcal{C}_n$ denote the set corresponding to those difference which are estimated to be nonzero with sample size n, and based on estimate $\hat{\beta}$ from (2) with penalty (3). If $\theta_{\mathcal{C}}^*$ denotes the true vector of pairwise differences included in $\mathcal{C}$, $\hat{\theta}_{\mathcal{C}}$ denotes the corresponding estimate based on $\hat{\beta}$, and $\hat{\beta}_i^{(LS)}$ the ordinary least squares estimate of β_i, 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* $w_{ij} = \phi_{ij}(n)|\hat{\beta}_i^{(LS)} - \hat{\beta}_j^{(LS)}|^{-1}$, *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) $\lim_{n\to\infty} P(\mathcal{C}_n = \mathcal{C}) = 1$.

REMARKS. The proof closely follows Zou (2006) and Bondell and Reich (2009), and is given below. 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 fixed, and that λ_n is determined by $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n \to \infty$. The latter is needed 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 $\lim_{n\to\infty} P(\mathcal{C}_n = \mathcal{C}) = 1$. $\phi_{ij}(n)$ only needs to converge toward a positive finite value (denoted by q_{ij}). 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). The concrete form of Σ results from the asymptotic marginal distribution of a set of nonredundant truly nonzero differences as specified in the proof.

Due to the (additive) form of the penalty (5), theoretic results from above directly generalize to the case of multiple categorial inputs, given the number p of predictors and the number k_l of levels of each predictor x_l are fixed.

Simple consistency $\lim_{n\to\infty} P(\|\hat{\beta} - \beta^*\|^2 > \varepsilon) = 0$ for all $\varepsilon > 0$ 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. This behavior is formally described in Proposition 2.

If adaptive weights are used and refitting is applied after the identification of clusters and relevant variables, asymptotic behavior is obtained which is comparable to Proposition 1. Since clustering and variable selection are directly based on the penalty with adaptive weights, part (b) of Proposition 1 is still valid. Asymptotic normality results from asymptotic normality of the ordinary least squares refit.

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 $u = \sqrt{n}(\beta - \beta^*)$, that is, $\beta = \beta^* + u/\sqrt{n}$, 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

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

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

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

with

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

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

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

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

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

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

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

and

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

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,

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

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 $n_i/n \to c_i$, $0 < c_i < 1\ \forall i$),

$$\lim_{n\to\infty} P\big(\sqrt{n}|\hat{\beta}_i^{(LS)}| \le \lambda_n^{1/2}\big) = 1, \qquad \text{respectively,}$$

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

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

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

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

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 $(\lambda_n/\sqrt{n})\widetilde{J}(u) \to_p \infty$. 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, that is, columns of dummy variables with equal coefficients are added and collapsed, and columns corresponding to zero coefficients are removed. Since $\forall i\ n_i/n \to c_i$ $(0 < c_i < 1)$,

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

Let $\theta_{\mathcal{C}^c}$ denote the vector of differences $\theta_{ij} = \beta_i - \beta_j$ which are truly zero, that is, 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, $V_n(u) \to_d V(u)$ for every u, where

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

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)]

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

Hence, $\hat{u}_{\mathcal{C}} \to_d N(0, \sigma^2 C^{-1})$. By changing the reference category, that is, 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 $\lim_{n\to\infty} P((i, j) \in \mathcal{C}_n) = 1$, if $(i, j) \in \mathcal{C}$, follows from part (a). We will now show that if $(i, j) \notin \mathcal{C}$, $\lim_{n\to\infty} P((i, j) \in \mathcal{C}_n) = 0$. 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 $\hat{\beta}_1 \le \cdots \le \hat{\beta}_z \le 0 \le \hat{\beta}_{z+1} \le \cdots \le \hat{\beta}_k$. That means estimate $\hat{\beta}$ from (2) with penalty (3) is equivalent to

$$\hat{\beta} = \underset{\{\beta_1 \le \cdots \le \beta_z \le 0 \le \beta_{z+1} \le \cdots \le \beta_k\}}{\operatorname{argmin}} \{(y - X\beta)^T (y - X\beta) + \lambda_n J(\beta)\},$$

with

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

Since $\hat{\beta}_q \ne 0$ 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 $\mathcal{C}_n$ 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

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

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

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

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

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

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

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

As a consequence,

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

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

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

PROPOSITION 2. *Suppose $0 \le \lambda < \infty$ 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, that is, $\lim_{n\to\infty} P(\|\hat{\beta} - \beta^*\|^2 > \varepsilon) = 0$ for all $\varepsilon > 0$.*

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 $0 < c_i < 1$ $\forall i$. $\square$

Asymptotic properties for the ordered case. Let now $\mathcal{C} = \{i > 0 : \beta_i^* \ne \beta_{i-1}^*\}$ denote the set of indices corresponding to differences of neighboring (true) dummy coefficients β_i^* which are truly nonzero, and again, $\mathcal{C}_n$ denote the set corresponding to those difference which are estimated to be nonzero, based on estimate $\hat{\beta}$ from (2) with penalty (4). The vector of first differences $\delta_i = \beta_i - \beta_{i-1}$, $i = 1, \ldots, k$, is now denoted as $\delta = (\delta_1, \ldots, \delta_k)^T$. 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 3. *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) $\lim_{n\to\infty} P(\mathcal{C}_n = \mathcal{C}) = 1$.

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

PROOF OF PROPOSITION 3. In Section 2.3 it 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. □

Precision of the approximate solution.

PROPOSITION 4. *If restriction $\theta_{ij} = \theta_{i0} - \theta_{j0}$ is represented by $A\theta = 0$, define $\hat{\theta}_{\gamma,\lambda} = \operatorname{argmin}_\theta\{(y - Z\theta)^T(y - Z\theta) + \gamma(A\hat{\theta})^T A\hat{\theta} + \lambda|\theta|\}$, 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 $\lambda \geq 0$,

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

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

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

we have

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

REFERENCES

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

BÜHLMANN, P. and YU, B. (2003). Boosting with the L2 loss: Regression and classification. *J. Amer. Statist. Assoc.* **98** 324–339. MR1995709

CANDES, E. and TAO, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

GERTHEISS, J., HOGGER, S., OBERHAUSER, C. and TUTZ, G. (2009). Selection of ordinally scaled independent variables. Technical Report 62, Dept. Statistics, Ludwig-Maximilians-Univ. München. To appear.

GERTHEISS, J. and TUTZ, G. (2009). Penalized regression with ordinal predictors. *Int. Statist. Rev.* **77** 345–365.

KARATZOGLOU, A., SMOLA, A., HORNIK, K. and ZEILEIS, A. (2004). kernlab—an S4 package for kernel methods in R. *J. Statist. Soft.* **11** 1–20.

KNEIB, T., HEINZL, F., BREZGER, A. and SABANÉS BOVÉ, D. (2009). BayesX: R utilities accompanying the software package BayesX. R package version 0.2.

LAND, S. R. and FRIEDMAN, J. H. (1997). Variable fusion: A new adaptive signal regression method. Technical Report 656, Dept. Statistics, Carnegie Mellon Univ. Pittsburg.

MEIER, L. (2007). grplasso: Fitting user specified models with Group Lasso penalty. R package version 0.2-1.

MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2008). The group lasso for logistic regression. *J. Roy. Statist. Soc. Ser. B* **70** 53–71. MR2412631

MEINSHAUSEN, N. (2007). Relaxed lasso. *Comput. Statist. Data Anal.* **52** 374–393. MR2409990

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. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNEIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. Roy. Statist. Soc. Ser. B* **67** 91–108. MR2136641

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

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B* **68** 49–67. MR2212574

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B* **67** 301–320. MR2137327

DEPARTMENT OF STATISTICS
LUDWIG-MAXIMILIANS-UNIVERSITÄT MUNICH
AKADEMIESTRASSE 1
80799 MUNICH
GERMANY
E-MAIL: jan.gertheiss@stat.uni-muenchen.de
gerhard.tutz@stat.uni-muenchen.de