LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Jan Gertheiss & Gerhard Tutz

Regularization and Model Selection with Categorial Effect Modifiers

Technical Report Number 073, 2010
Department of Statistics
University of Munich

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

Regularization and Model Selection with Categorial Effect Modifiers

Jan Gertheiss*† & Gerhard Tutz†

January 18, 2010

Abstract

The case of continuous effect modifiers in varying-coefficient models has been well investigated. Categorial effect modifiers, however, have been largely neglected. In this paper a regularization technique is proposed that allows for selection of covariates and fusion of categories of categorial effect modifiers in a linear model. It is distinguished between nominal and ordinal variables, since for the latter more economic parametrizations are warranted. The proposed methods are illustrated and investigated in simulation studies and real world data evaluations. Moreover, some asymptotic properties are derived.

Keywords: Categorial Predictors, Fused Lasso, Linear Model, Variable Selection, Varying-Coefficient Models

1 Introduction

Varying-coefficient models (Hastie and Tibshirani, 1993) offer a quite flexible framework for regression modeling. In a standard linear model (with one effect modifier) regression coefficients β_j are allowed to vary with the values of a variable u – the so-called effect modifier. That means, we have

$$y = \beta_0(u) + x_1\beta_1(u) + \ldots + x_p\beta_p(u) + \epsilon,$$

where functions $\beta_j(u)$ may depend on the effect modifier u, $j = 0, \ldots, p$, $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$.

The case of metric effect modifiers u has been investigated thoroughly, in the linear model as given above or in other situations (see for example Cardot and

*To whom correspondence should be addressed: jan.gertheiss@stat.uni-muenchen.de.
†Department of Statistics, Ludwig-Maximilians-Universität Munich, Germany.

Sarda, 2008; Hoover et al., 1998; Kim, 2007; Mu and Wei, 2009; or Qu and Li, 2006). The classical approach is to estimate functions $\beta_j(\cdot)$ nonparametrically, for example using splines (see e.g. Lu et al., 2008) or localizing techniques (Fan et al., 2003; or Kauermann and Tutz, 2000). Recently, Wang et al. (2008) and Wang and Xia (2009) proposed penalty approaches for selecting relevant predictors x_j, while Leng (2009) used penalized likelihood estimation to investigate which functions $\beta_j(\cdot)$ actually vary over u. The latter problem means to distinguish between the cases where $\beta_j(u)$ is a constant or not, while selection of predictors is equivalent to determine if $\beta_j(u) = 0$. Hofner et al. (2008) proposed a boosting procedure for the selection of time-varying effects in survival models.

In the present paper methods for categorial effect modifiers in the classical linear model are proposed. The main problem with categorial effect modifiers is that the number of parameters to be estimated may become very large. For categorial $u \in \{1, \ldots, k\}$ the varying functions have the form

$$\beta_j(u) = \sum_{r=1}^{k} \beta_{jr} I(u = r),$$

which means that k parameters have to be estimated. Correspondingly the model with p predictors,

$$y = \sum_{r=1}^{k} \beta_{0r} I(u = r) + \sum_{r=1}^{k} x_1 \beta_{1r} I(u = r) + \ldots + \sum_{r=1}^{k} x_p \beta_{pr} I(u = r) + \epsilon,$$

contains $(p+1)k$ parameters. The interpretation is that on level r of u the model

$$y = \beta_{0r} + x_1 \beta_{1r} + \ldots + x_p \beta_{pr} + \epsilon$$

holds. In many situations, however, the number of parameters has to be reduced – in order to stabilize estimation of parameters and/or to facilitate interpretation. For that purpose we propose a penalty approach that accounts for both aspects already mentioned above: variable selection with respect to predictors x_j, and investigation if functions $\beta_j(\cdot)$ are (partially) constant. That means, the aim is to decide if some of the parameters β_{jr} and β_{js} are equal for fixed j. Moreover, the presented method allows for level specific variable selection, which means that predictors may be excluded (i.e. corresponding coefficients are set to zero) for specific values of u only.

The following example illustrates that the approach is also useful in the case of few predictors. Even then it simplifies the assumed structure of the predictors. We consider the data collected by Derek Whiteside, reported by Hand et al. (1994) and analyzed by Venables and Ripley (2002). Given are the weekly gas consumption (in 1000 cubic feet) and average external temperature (in degree C) at Whiteside's own house in south-east England during two 'heating seasons' – one

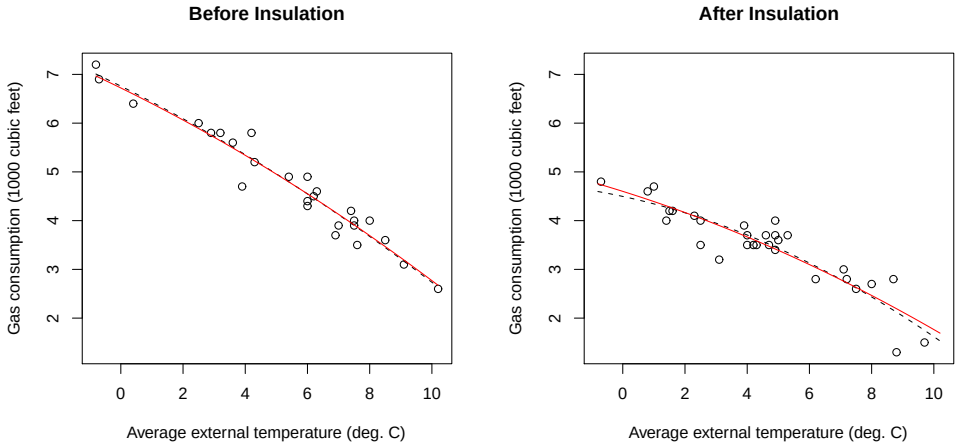

Figure 1: Whiteside's data showing the effect of insulation on household gas consumption with estimated quadratic regression curves; dashed black lines refer to the full model, solid red ones to the regularized model with coefficients of quadratic terms set equal.

before and one after cavity-wall insulation was installed, cf. Venables and Ripley (2002). The most complex model used by Venables and Ripley (2002) fits gas consumption as a quadratic function of temperature separately for both seasons before and after insulation. That means with $u \in \{1, 2\} = \{\text{Before}, \text{After}\}$, x denoting temperature and y gas consumption, one has the linear predictor

$$\eta(x, u) = \beta_0(u) + x\beta_1(u) + x^2\beta_2(u) = E(y|x, u),$$

and for fixed r

$$\eta(x, u = r) = \beta_{0r} + x\beta_{1r} + x^2\beta_{2r} = E(y|x, u = r).$$

In Figure 1 the data are shown together with estimated regression curves. In each plot the dashed black curve refers to the full model; the solid red ones result from the method proposed in this article. Though dashed and solid curves seem quite similar, our model has one degree of freedom less since the parameters of the quadratic term β_{21} and β_{22} are set equal, and hence are not varying over heating seasons. Venables and Ripley's speculation that the quadratic term is possibly needed for the after-insulation group only is not confirmed.

The paper is organized as follows: We first introduce the method and discuss some computational aspects in Section 2. Then large sample properties are investigated (Section 3) and the proposed methods are tested in simulation studies (Section 4). In Sections 5 and 6 real world data are evaluated, and generalizations to multiple effect modifiers are discussed.

2 Penalized Estimation

Let (y_i, x_i, u_i), $i = 1, \ldots, n$ denote the data and identify $\beta_j(u) = \beta_{ju}$. Instability of the ordinary least squares estimate can be avoided by penalized estimation:

$$\hat{\beta} = \text{argmin}_\beta \, Q_p(\beta), \tag{1}$$

with

$$\begin{aligned} Q_p(\beta) &= \sum_{i=1}^{n} \left(y_i - \beta_0(u_i) - \sum_{j=1}^{p} x_{ij} \beta_j(u_i) \right)^2 + \lambda J(\beta) \\ &= (y - Z\beta)^T (y - Z\beta) + \lambda J(\beta), \end{aligned} \tag{2}$$

$y = (y_1, \ldots, y_n)^T$ and $\beta = (\beta_1^T, \ldots, \beta_k^T)^T$, with $\beta_r = (\beta_{0r}, \beta_{1r}, \ldots, \beta_{pr})^T$. The ith row of design matrix Z is $((1, x_i^T) I(u_i = 1), \ldots, (1, x_i^T) I(u_i = k))$. Without penalty $J(\beta)$, i.e. with $\lambda = 0$, ordinary least squares estimation is obtained. With increasing λ the influence of $J(\beta)$ is increased. The crucial point is to chose an adequate penalty $J(\beta)$. Classical penalties are the *Ridge* (Hoerl and Kennard, 1970)

$$J(\beta) = \sum_{j,r} \beta_{jr}^2,$$

or the *Lasso* (Tibshirani, 1996)

$$J(\beta) = \sum_{j,r} |\beta_{jr}|.$$

While the Ridge only shrinks estimates $\hat{\beta}_{jr}$ toward zero, the Lasso additionally allows for variable selection/exclusion, i.e. some $\hat{\beta}_{jr}$ may be set to zero (for details see Tibshirani, 1996). Though variable selection is also included, the pure Lasso penalty is not adequate since it does not enforce $\hat{\beta}_{jr} = \hat{\beta}_{js}$ for some $r \neq s$, which is needed to obtain potentially (piecewise) constant functions $\hat{\beta}_j(u)$. So in the following we present an approach which also allows for such fusion of coefficients. We distinguish between nominal and ordinal effect modifiers because of their different information content.

2.1 Nominal and Ordinal Effect Modifiers

For nominal u we propose penalty

$$J(\beta) = \sum_{j=0}^{p} \sum_{r>s} |\beta_{jr} - \beta_{js}| + \sum_{j=1}^{p} \sum_{r=1}^{k} |\beta_{jr}|. \tag{3}$$

The first term enforces the collapsing of categories of the effect modifier. In the extreme case (i.e. the case of very strong penalization), the effects of covariates

do not depend on the category of u and one obtains $\hat{\beta}_{j1} = \hat{\beta}_{j2} = \ldots = \hat{\beta}_{jk} = \hat{\beta}_{j.}$. The second term in (3) steers selection/exclusion of covariates. In the extreme case $\hat{\beta}_{j1} = \hat{\beta}_{j2} = \ldots = \hat{\beta}_{jk} = 0$ is obtained, and covariate x_j is omitted.

If u is ordinal, levels can be reasonably ordered. Hence the penalty can be modified using this additional information. More concrete, we use

$$J(\beta) = \sum_{j=0}^{p} \sum_{r=2}^{k} |\beta_{jr} - \beta_{j,r-1}| + \sum_{j=1}^{p} \sum_{r=1}^{k} |\beta_{jr}| \tag{4}$$

That means, within each predictor x_j one uses a *Fused Lasso* type penalty (compare Tibshirani et al., 2005), since only differences of 'adjacent' coefficients β_{jr} and $\beta_{j,r-1}$ are penalized. If the effect modifier u is nominal, all pairwise differences of coefficients belonging to covariate x_j are considered ($j = 1, \ldots, p$), as described above.

Following Tibshirani et al. (2005) the selection and the fusion part of the penalty may be differentially weighted. That means, with $\psi \in (0, 1)$, one can also use

$$J(\beta; \psi) = \psi \sum_{j=0}^{p} \sum_{r>s} |\beta_{jr} - \beta_{js}| + (1 - \psi) \sum_{j=1}^{p} \sum_{r=1}^{k} |\beta_{jr}|, \tag{5}$$

or (depending on the scale level of u)

$$J(\beta; \psi) = \psi \sum_{j=0}^{p} \sum_{r=2}^{k} |\beta_{jr} - \beta_{j,r-1}| + (1 - \psi) \sum_{j=1}^{p} \sum_{r=1}^{k} |\beta_{jr}|. \tag{6}$$

The use of flexible ψ means, however, that another tuning parameter (beside penalty parameter λ) is introduced, and it is not clear if this modification really has better performance than penalty (3) and (4) respectively, where $\psi = 0.5$ is fixed. This issue is investigated further in simulation studies in Section 4.

2.2 Computational Issues

When computing estimates it is useful to consider the penalized least squares criterion (2) as a constrained optimization problem

$$\hat{\beta} = \text{argmin}_\beta \sum_{i=1}^{n} \left(y_i - \beta_0(u_i) - \sum_{j=1}^{p} x_{ij} \beta_j(u_i) \right)^2, \text{ subject to } J(\beta) \leq s,$$

where tuning parameter s plays a role which comparable to penalty parameter λ above; see, for example, Hastie et al. (2001). In matrix notation we have

$$\hat{\beta} = \text{argmin}_\beta (y - Z\beta)^T (y - Z\beta), \text{ subject to } J(\beta) \leq s,$$

with y and Z chosen as in (2). With δ defined as

$$\delta^T = (\delta_0^T, \ldots, \delta_p^T), \quad \text{with } \delta_j = (\beta_{j2} - \beta_{j1}, \beta_{j3} - \beta_{j1}, \ldots, \beta_{jk} - \beta_{j,k-1})^T,$$

which can be written as

$$\delta = D\beta,$$

with adequately chosen D, we may set $\nu = (\beta^T, \delta^T)^T$. Then, if penalty (3) is chosen, $\hat{\beta}$ can be computed via

$$\hat{\nu} = \text{argmin}_\nu (y - U\nu)^T (y - U\nu), \text{ subject to } \|\nu\|_1 \leq s \text{ and } A\nu = 0,$$

where $\|\nu\|_1$ denotes the L_1-norm of ν, a possible choice of U is $U = (Z|0)$, and $A = (D|-I)$ since we $A\nu = D\beta - \delta = 0$ has to hold. If every entry of ν is split into positive and negative part, this constrained minimization problem can (in principle) be solved via quadratic programming; for example using methods from the R add-on package `kernlab` (Karatzoglou et al., 2004; R Development Core Team, 2009). If ordinal penalty (4) is chosen, computation can be done in a completely analogous way. With flexible ψ, the constraint becomes $(1-\psi)\|\beta\|_1 + \psi\|\delta\|_1 \leq s$.

The problem with quadratic programming is that the solution can only be computed for a single value s. To obtain a coefficient path (i.e. coefficient values seen as a function of s) the procedure needs to be applied repeatedly. Moreover, in some cases 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 $A\nu \approx 0$ holds. The idea is to 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. We define

$$\hat{\nu}_\gamma = \text{argmin}_\nu \left\{ (y - U\nu)^T (y - U\nu) + \gamma (A\nu)^T A\nu + \lambda \|\nu\|_1 \right\} = \text{argmin}_\nu \left\{ h(\nu, \gamma) \right\}.$$

The first penalty term, which is weighted by γ, penalizes violations of restrictions $A\nu = 0$. Since $\min_\nu h(\nu, \gamma)$ is a monotone function of γ and limited above by $h(\hat{\nu}, \cdot)$, and $h(\nu, \gamma)$ a continuous and convex function of ν, (under the assumption that $\hat{\nu}_\gamma$ is the unique minimizer of $h(\nu, \gamma)$ and also $\hat{\nu}$ is unique) the exact solution of the optimization problem considered here is obtained as the limit

$$\hat{\nu} = \lim_{\gamma \to \infty} \hat{\nu}_\gamma.$$

Hence, with sufficiently high γ an acceptable approximation of $\hat{\nu}$ should be obtained by $\hat{\nu}_\gamma$. Similar approximations have been shown to work quite well (Gertheiss and Tutz, 2009). To judge on precision we use

$$\Delta_\gamma = (A\hat{\nu}_\gamma)^T A\hat{\nu}_\gamma,$$

which also depends on the chosen λ (resp. s). In our analyses we mostly obtained Δ_γ-values of about 10^{-20} or better, which is comparable to results obtained by using the `kernlab` package (note, also if quadratic programming is used to compute "exact" solutions, constraints are just "numerically" met). The advantage of using the estimate $\hat{\nu}_\gamma$ is that its whole path can be computed using `lars`, since it can be formulated as a Lasso solution.

3 Large Sample Properties and Modifications

In the following we will investigate asymptotic properties and introduce a modified version of the proposed estimator that is also consistent in terms of variable selection and the identification of relevant differences $\beta_{jr} - \beta_{js}$. If sample size n tends to infinity it is also assumed that the number of observations n_r made on level r of u tends to infinity for all r. In this case estimator $\hat{\beta}$ as defined in (1) and (2) with penalty (3) is consistent in terms of $\lim_{n\to\infty} P(||\hat{\beta} - \beta^*||^2 > \epsilon) = 0$ for all $\epsilon > 0$, if β^* denotes the vector of true coefficient functions $\beta_j(u)$, resp. true β_{jr}. This behavior is formally described in the following

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

The proof is given in the Appendix. If u is ordinal, penalty (4) is employed and consistency is proven in a completely analogue way. Also employing the generalized version (5) or (6) does not affect consistency results.

However, as pointed out by Zou (2006), regularization as applied so far does not ensure consistency in terms of variable selection. That means, the probability that $\hat{\beta}_{jr} = 0$ if $\hat{\beta}^*_{jr} = 0$ does not tend to one. In our case this inconsistency also applies to differences $\hat{\beta}_{jr} - \hat{\beta}_{js}$.

For solving the problem of selection inconsistency of the original Lasso, Zou (2006) proposed an adaptive version with so-called oracle properties. A corresponding modification is also possible for our estimator. That means, given nominal u, we employ the adaptive penalty

$$J(\beta) = \sum_{j=0}^{p} \sum_{r>s} w_{rs(j)} |\beta_{jr} - \beta_{js}| + \sum_{j=1}^{p} \sum_{r=1}^{k} w_{r(j)} |\beta_{jr}| \tag{7}$$

with adaptive weights

$$w_{rs(j)} = \phi_{rs(j)}(n) |\hat{\beta}^{(LS)}_{jr} - \hat{\beta}^{(LS)}_{js}|^{-1} \quad \text{and} \quad w_{r(j)} = \phi_{r(j)}(n) |\hat{\beta}^{(LS)}_{jr}|^{-1}, \tag{8}$$

with $\hat{\beta}_{jr}^{(LS)}$ denoting the ordinary least squares estimator of β_{jr}. For the sequences $\phi_{rs(j)}(n)$ and $\phi_{r(j)}(n)$ we only need $\phi_{rs(j)}(n) \to q_{rs(j)}$ and $\phi_{r(j)}(n) \to q_{r(j)}$ respectively, with $0 < q_{rs(j)}, q_{r(j)} < \infty$. Though these assumptions are quite general, $\phi_{rs(j)}(n)$ and $\phi_{r(j)}(n)$ will usually be fixed, for example as ψ and $(1-\psi)$ to obtain a generalization like (5). In contrast to Proposition 1, penalty parameter λ from (2) is not fixed now, but increasing with sample size n. More precisely, we need $\lambda = \lambda_n$, with $\lambda_n \to \infty$ for $n \to \infty$, but $\lambda_n/\sqrt{n} \to 0$ for $n \to \infty$.

Before giving the asymptotic properties of the adaptive version, we define $\beta_{-0,r} = (\beta_{1r}, \ldots, \beta_{pr})^T$, i.e. the vector of regression coefficients on level r of u without the intercept, and $\delta_j = (\beta_{j2} - \beta_{j1}, \beta_{j3} - \beta_{j1}, \ldots, \beta_{jk} - \beta_{j,k-1})^T$, i.e. the vector of pairwise differences of regression coefficients belonging to predictor x_j (see also Subsection 2.2). Because also differences of intercepts are considered, δ_j refers to $j = 0, \ldots, p$. Furthermore, we define $\beta_{-0}^T = (\beta_{-0,1}^T, \ldots, \beta_{-0,k}^T)$, $\delta^T = (\delta_0^T, \ldots, \delta_p^T)$, and $\theta^T = (\beta_{-0}^T, \delta^T)$. Now, let $\mathcal{C}$ denote the set of indices corresponding to entries of θ which are truly non-zero, and $\mathcal{C}_n$ denote the set corresponding to those entries which are estimated to be non-zero with sample size n, and based on estimate $\hat{\beta}$ from (1) with penalty (7). If $\theta_{\mathcal{C}}^*$ denotes the true vector of θ-entries included in $\mathcal{C}$, and $\hat{\theta}_{\mathcal{C}}$ denotes the corresponding estimate based on $\hat{\beta}$, then the following 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_r satisfy $n_r/n \to c_r$, where $0 < c_r < 1$. Then penalty (7) with weights (8) ensures 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 uses ideas from Zou (2006), Bondell and Reich (2009) and Gertheiss and Tutz (2009), and is given in the Appendix. The concrete form of Σ results from the asymptotic marginal distribution of a set of non-redundant truly non-zero regression parameters or differences of parameters. Since all estimated differences are (deterministic) linear functions of estimated parameters, covariance-matrix Σ is singular.

If effect modifier u is ordinal, the weighting scheme and the asymptotic behavior of the corresponding estimator (incl. proofs) are completely analogue. The only difference is that just weights $w_{r,r-1(j)}$ (instead of $w_{rs(j)}$) are needed, and that δ only consists of differences of adjacent β-coefficients.

4 Numerical Experiments

Before the presented approach is applied to real-world data, we illustrate and investigate the method in simulation studies where the true underlying model is known.

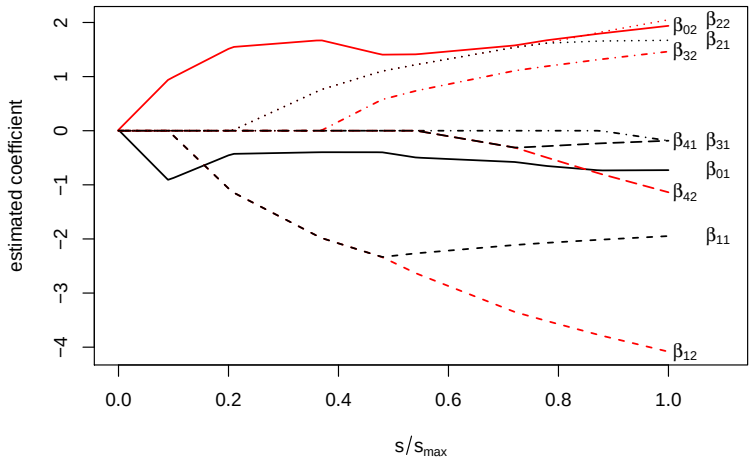

Figure 2: Fitted β-coefficients as functions of tuning parameter s; true values are $\beta_{01} = -1$, $\beta_{02} = 1$, $\beta_{11} = -2$, $\beta_{12} = -4$, $\beta_{21} = \beta_{22} = 2$, $\beta_{31} = 0$, $\beta_{32} = 2$, $\beta_{41} = \beta_{42} = 0$.

4.1 An Illustrative Example

At first, we assume an effect modifier u with only $k = 2$ levels. More precisely, on level $u = 1$ we assume model

$$y = -1 - 2x_1 + 2x_2 + \epsilon,$$

and on level $u = 2$

$$y = 1 - 4x_1 + 2x_2 + 2x_3 + \epsilon.$$

That means, the true intercepts $\beta_{01} = -1$ and $\beta_{02} = 1$ vary with u, as well as $\beta_{11} = -2$ and $\beta_{12} = -4$. Coefficients $\beta_{21} = \beta_{22} = 2$ are constant, i.e. not depending on u. Predictor x_3 is relevant only if $u = 2$, since $\beta_{32} = 2$ but $\beta_{31} = 0$. In addition, a truly pure noise variable x_4 is considered as a potential regressor in both models, i.e. $\beta_{41} = \beta_{42} = 0$. We generate $n = 200$ data points with x_{ij} independently drawn from an $U[0, 1]$ distribution, class levels $u_1 = \ldots = u_{100} = 1$ and $u_{101} = \ldots = u_{200} = 2$, and standard normal error ϵ. Figure 2 shows fitted coefficient paths for all β_{jr} as functions of tuning parameter s, if the standard (non-adaptive) approach with $\psi = 0.5$ is applied (as given in (3)). Black curves correspond to $u = 1$, red ones to $u = 2$. At $s/s_{\max} = 1$ ordinary least squares estimates are obtained. With decreasing s, resp. increasing penalty λ coefficients are successively fused and shrunken toward zero. It is seen that at first coefficient β_{31} is (correctly) set to zero. Then β_{21} and β_{22} are set equal, as well as β_{41} and β_{42}. A little bit later β_{41} and β_{42} are simultaneously set to zero, as desired. In the following steps β_{11} and β_{12} are (wrongly) fused and truly non-zero coefficients are

set to zero. Intercepts β_{01} and β_{02} are not fused until minimal s is chosen. Since only their difference is penalized, at $s = 0$ they equal $\bar{y}$ – the empirical mean of y. In our case we have $\bar{y} = 0.015$; hence, β_{01} and β_{02} seem to be zero at $s = 0$, but actually they are not.

4.2 Comparison of Methods

In order to investigate the potential impact of modifications proposed in Sections 2 and 3, we extend the simulation setting from above. We introduce another level of u and another predictor. More precisely we have:

$$\begin{aligned} y &= -1 - 2x_1 + 2x_2 + 0x_3 + 0x_4 + 0x_5 + \epsilon \text{ on level } u = 1, \\ y &= +1 - 4x_1 + 2x_2 + 2x_3 + 0x_4 + 0x_5 + \epsilon \text{ on level } u = 2, \\ y &= +1 + 2x_1 + 2x_2 + 2x_3 - 4x_4 + 0x_5 + \epsilon \text{ on level } u = 3, \end{aligned} \tag{9}$$

with standard normal error ϵ. We independently generate 300 training and 900 test data data points, both with balanced u, and compare the standard and the adaptive version, both with fixed $\psi = 0.5$ as well as ψ treated as another tuning parameter (which is chosen via cross-validation). This procedure is (independently) repeated 100 times. Results in terms of the (empirical) MSE of parameter estimates and prediction accuracies are shown in Figure 3. Prediction accuracy is measured by the Mean Squared Error of Prediction (MSEP) on the test set. It is seen that all regularized approaches are superior to the ordinary least squares (ols) estimate, which is nicely illustrated by relative errors given in the right panel of Figure 3. Furthermore, if a regularized approach is applied, using adaptive weights as defined in (8) seems to increase accuracy of parameter estimates and prediction. Allowing flexible ψ, by contrast, does not lead to better results.

Beside accuracy of prediction and parameter estimation we examine selection and clustering performance of the considered methods. So in Figure 4 averaged false positive and false negative rates (FPR/FNR) are shown – concerning variable selection and the identification of relevant differences between (potentially) varying coefficients. False positive means that a truly zero coefficient from (9) is set to non-zero, or that a truly zero difference of coefficients belonging to the same predictor is fitted as non-zero, respectively. False negative means that truly non-zero values are estimated to be zero. It is seen, however, that false negatives are hardly observed. In case of the ols estimator false positive rates equal 1, of course, since all coefficients/differences are (almost surly) set to non-zero. Also if the standard regularized approach is applied, false positive rates are rather high. Using adaptive weights, however, substantially reduces error rates, concerning both variable selection and clustering. Differences between fixed $\psi = 0.5$ and ψ chosen via cross-validation are negligible.

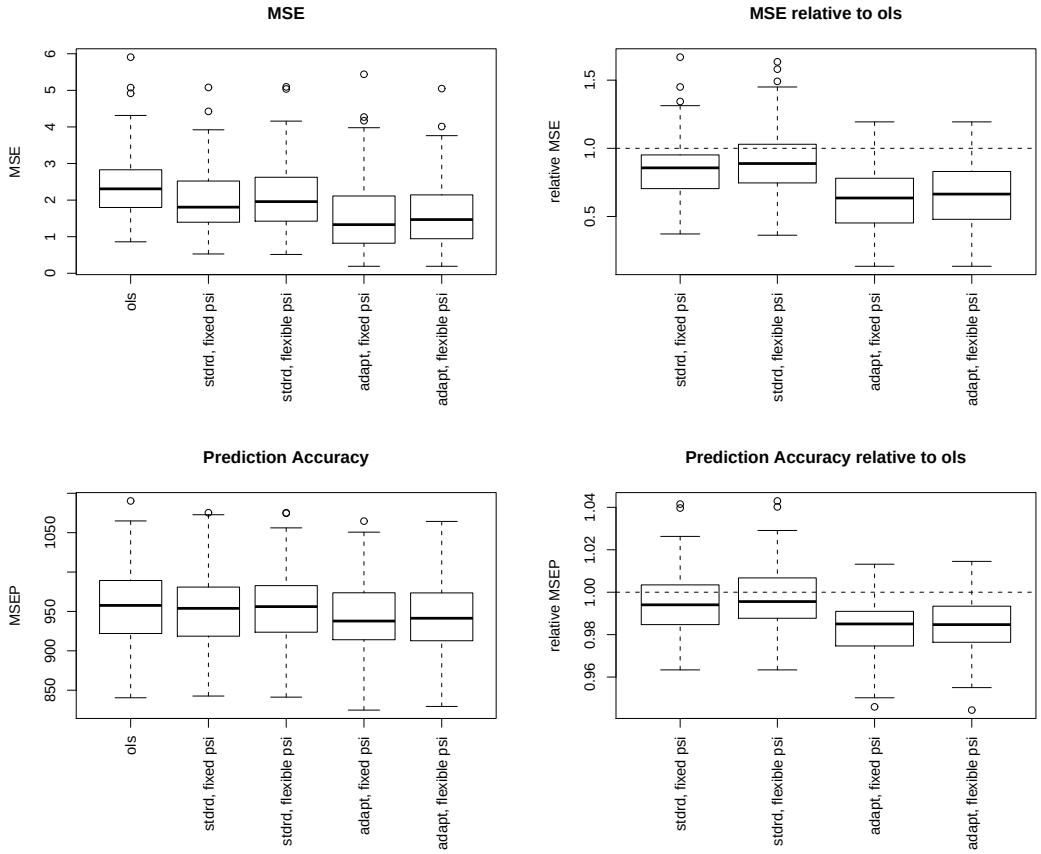

Figure 3: Absolute as well as relative errors of parameter estimates (MSE) and predictions accuracy (MSEP); simulation scenario (9).

As a second scenario we introduce two additional pure noise input variables x_6 and x_7, and repeat the analysis. Results are shown in Figures 5 and 6. Since now a higher number of pure noise variables is given than before, one may think that emphasis should be placed on the penalty's selection part; that is, $\psi < 0.5$ should be chosen in (5). Surprisingly, however, choosing ψ via cross-validation is not superior to using fixed $\psi = 0.5$ (i.e. putting equal weights on the selection and the fusion part). When comparing Figures 3 and 5, only differences between regularized and ordinary least squares estimates seem larger than before. Using the regularized version with adaptive weights, for example, on average the MSE of the ols model is now reduced by more than 50% (see Figure 5, top right).

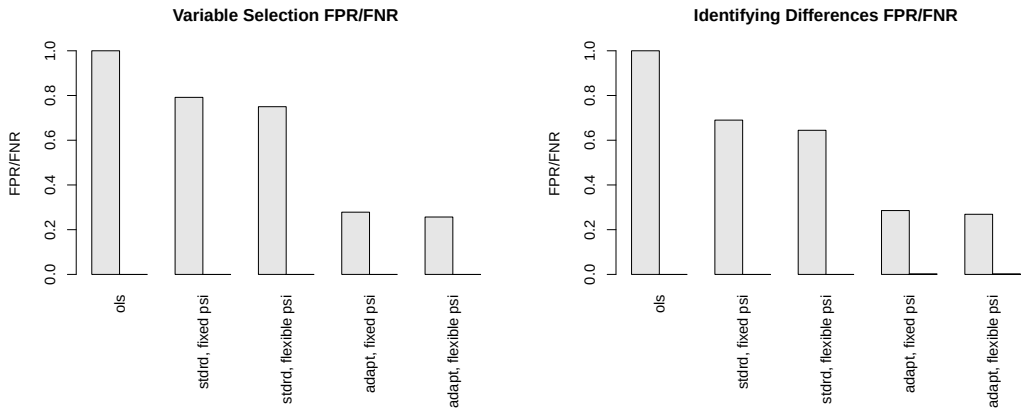

Figure 4: False positive and false negative rates (light/dark-colored bars) concerning variable selection and identification of relevant differences between (potentially) varying coefficients; simulation scenario (9), values are averaged over all simulation runs.

5 Real World Data Evaluation

In the Introduction, results of the analysis of Whiteside's data have already been shown. The fitted model (chosen via 5-fold cross-validation) had one degree of freedom less than the ordinary least squares fit, since the coefficient of the quadratic term was set as constant over both heating seasons. In the following we will analyze two larger data sets, namely income data from Germany, and data collected in Austria during a study on the functioning of lungs of schoolchildren. As the simulation study from above suggests, we will fix $\psi = 0.5$ in the following (i.e. put equal weights on the penalty's selection and fusion part), since flexible ψ did not produce better results. Before our regularized methods are applied variables are scaled to have unit variance to make results independent of the chosen units.

5.1 Analysis of Income Data

We analyze the relationship of monthly income and several (potentially) explanatory variables. The data are taken from the Socio-Economic Panel Study (SOEP) of the year 2002. The SOEP is a representative longitudinal study of private households in Germany, but we only use data from 2002 in a cross-sectional analysis. Table 1 shows the response and predictors we consider for the regression analysis. The so-called *Abitur* is a diploma from German secondary school qualifying for university admission or matriculation. It is comparable to the British A-levels.

We fit (the logarithm of) monthly income using a linear regression model but let coefficients vary with the corresponding person's gender. From former studies

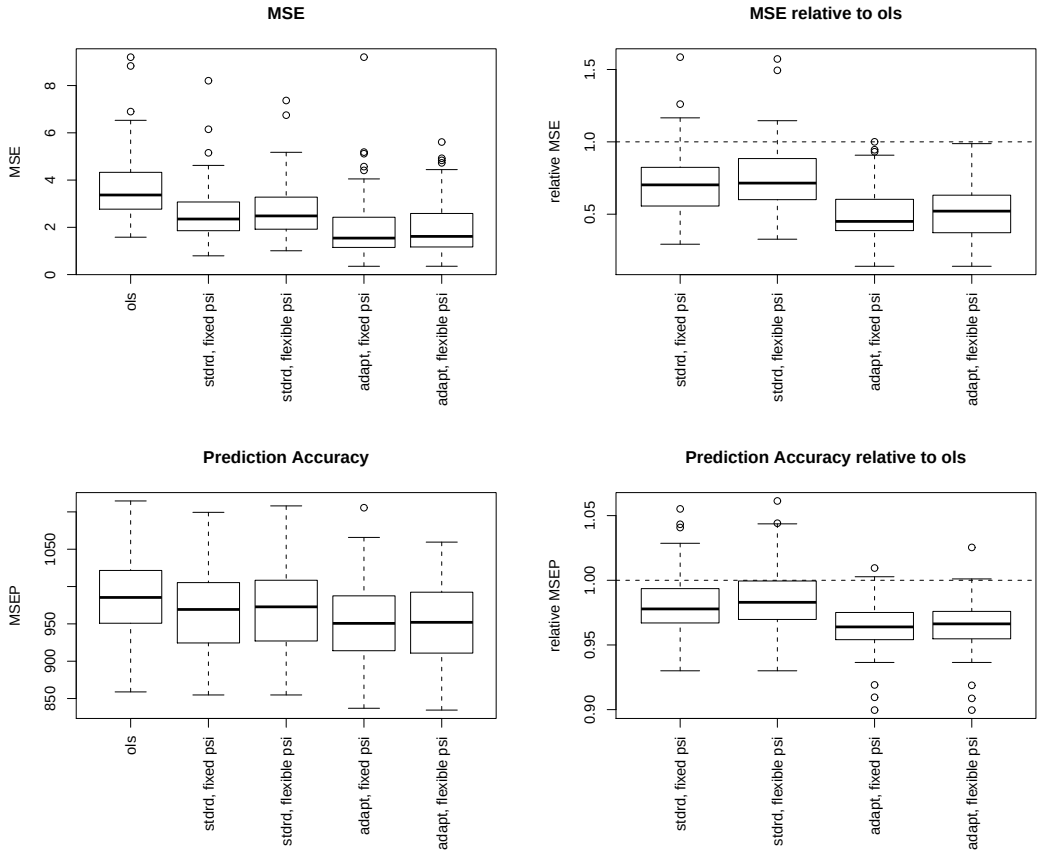

Figure 5: Absolute as well as relative errors of parameter estimates (MSE) and predictions accuracy (MSEP); simulation scenario (9) with two additional pure noise input variables.

it is known that the influence of age is rather quadratic than linear. Therefore Age^2 is also included. That means, we have

$$\begin{aligned} \log(\text{Income}) &= \beta_0(\text{Gender}) + \beta_1(\text{Gender})\text{Age} + \beta_2(\text{Gender})\text{Age}^2 \\ &+ \beta_3(\text{Gender})\text{Tenure} + \beta_4(\text{Gender})\text{Height} \\ &+ \beta_5(\text{Gender})\text{Married} + \beta_6(\text{Gender})\text{Abitur} \\ &+ \beta_7(\text{Gender})\text{Blue-collar} + \epsilon. \end{aligned} \tag{10}$$

Figure 7 shows coefficient paths for all predictors and the intercept. The dashed lines refer to males, the solid ones to females. For small s (i.e. with high penalty λ) regression coefficients are set to zero or equal for males and females. If s is increased, it is seen that gender may play an important role as an effect modifying factor. In particular, it is interesting that earnings of married men tend to be higher than those of unmarried men, while the effect of marriage seems to be contrary in case of women. Qualitatively speaking, effects of job tenure, Abitur

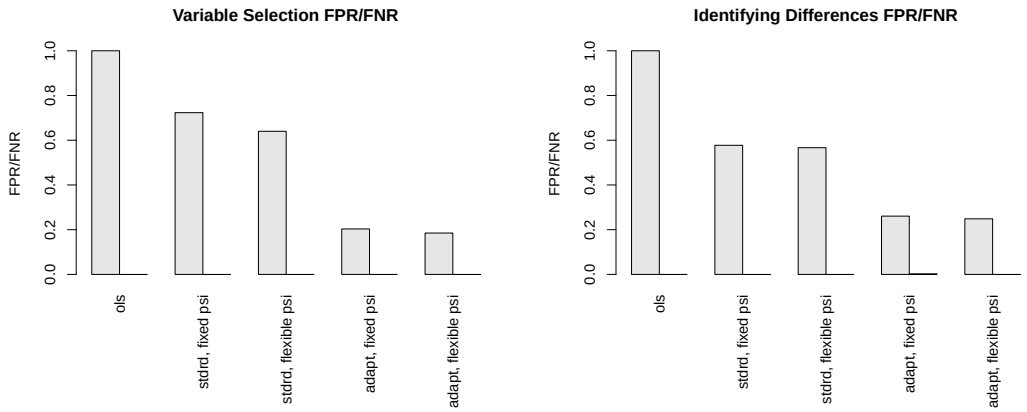

Figure 6: False positive and false negative rates (light/dark-colored bars) concerning variable selection and identification of relevant differences between (potentially) varying coefficients; simulation scenario (9) with two additional pure noise input variables.

Response:	Monthly income	in Euro
Predictors:	Age	in years between 21 and 60
	Job tenure	in months
	Body height	in cm
	Gender	male/female
	Married	no/yes
	Abitur ($\approx$ A-levels)	no/yes
	Blue-collar worker	no/yes

Table 1: Available data for the analysis of the relationship between income and several explanatory variables.

and being a blue-collar worker are similar for males and females, but – particularly in case of job tenure – effects tend to be stronger for females than for males. The phenomenon that taller people earn more than smaller ones is observed for both males and females – with coefficients being set as constant as long as $s/s_{\max} \leq 0.96$.

To evaluate if found differences between men and women can be regarded as substantial, an adequate s-value is chosen via cross-validation. The vertical dotted line in each path plot in Figure 7 indicates the corresponding s with minimum (5-fold) cross-validation score. It is seen that the best solution is found at a point where most coefficients vary with gender. Only intercepts and the effect of body height are fitted as constant over gender. The fact (which is well known for Germany) that earnings of males are (still) higher on average than those of females, is (primarily) modeled via the different influence of age.

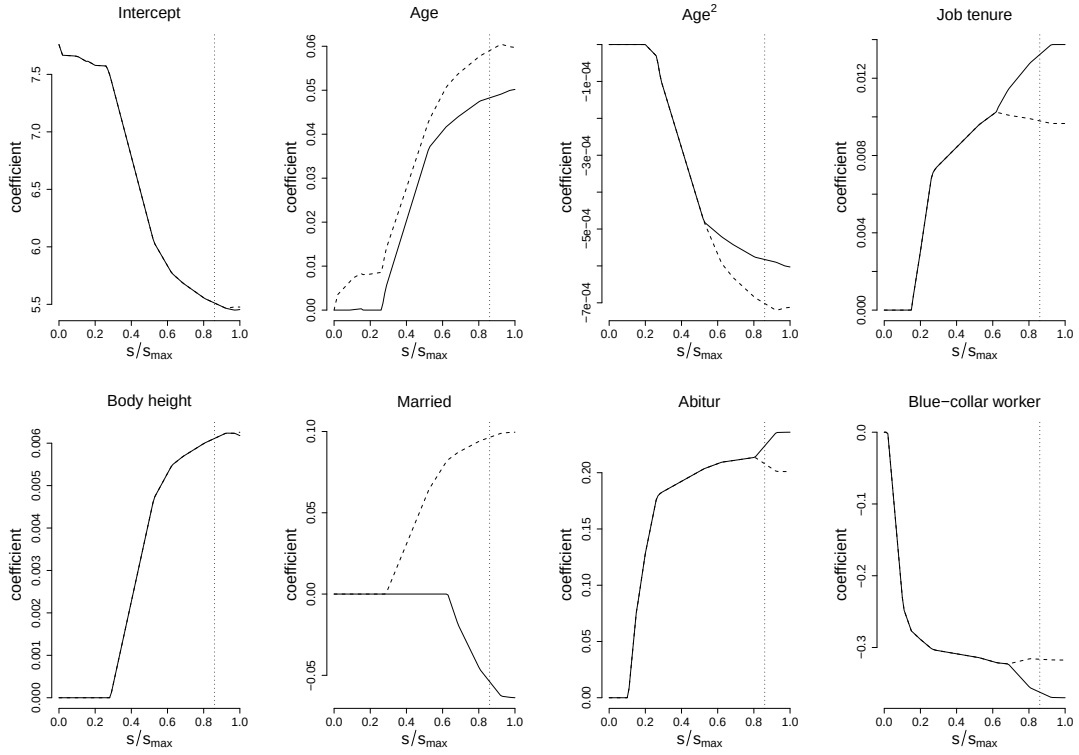

Figure 7: Paths of coefficients (possibly) varying with gender given model (10); dashed lines refer to males, solid ones to females, the vertical dotted line indicates coefficients at cv score minimizing $s/s_{\max} = 0.86$.

5.2 Lung Capacities of Schoolchildren

The data analyzed in the following have been collected by the University of Innsbruck, Austria, during a study on the functioning of lungs and diseases of the respiratory tracts of schoolchildren. The point of interest was the question, whether the functioning of lungs is affected by industry-induced air pollution. The data are based on a cross-sectional study in the district of Brixlegg (Austria). Besides the environmental pollution other covariates are given. A summary of the data used in the following is found in Table 2. We will analyze the relationship between the capacity of the lungs (in liters) and the provided covariates. The degree of environmental pollution at the place of residence is given as a categorial predictor with three levels: highly polluted zone, slightly polluted zone, or with high ozone exposure because of altitude (Brixlegg is located in the Alps). Since levels can only be partially ordered, the degree of pollution is treated as a nominal covariate. All other explantory variables are metric (age, body weight, body height) or binary factors (sex, smoking mother/father, etc.), see Table 2.

Response:	Capacity of the lungs	in liters
Predictors:	Age	in months
	Body weight	in kilograms
	Body height	in cm
	Sex	male/female
	Parental level of education	A-levels etc. (no/yes)
	Existing allergies	no/yes
	Diseases of the respiratory tracts	no/yes
	Does mother smoke?	no/yes
	Does father smoke?	no/yes
	Suffering frequently from colds?	no/yes
	Suffering frequently from coughs?	no/yes
	Lung or bronchial tube diseases	no/yes
	Degree of environmental pollution at place of residence	categorial with zones/levels: 1: highly polluted 2: slightly polluted 3: high ozone exposure

Table 2: Available data for the analysis of the functioning of lungs of schoolchildren.

Since the main interest is on investigating the effect of pollution on the capacity of lungs, a natural first step is to build a model with all predictors except pollution – a so-called confounder model. Then it is to be checked if the model is significantly improved if the degree of pollution is added. If we just fit main effect models – firstly except, and then including pollution – the model is not significantly improved if the degree of pollution is taken into account (F-test based p-value 0.13). By contrast, if pollution is included as an effect modifying factor, the initial model is significantly improved (p-value 0.02). However, most regression parameters are far away from being 'significantly non-zero'. So it can be assumed that the resulting model is unnecessarily complex, and we will use the proposed regularization technique to obtain a sparser representation. A standard stepwise model selection procedure is no alternative, since we do not only aim at excluding covariates or setting parameters to zero, but also look for parameters that are constant over different levels of pollution.

In Figure 8 the (5-fold) cross-validation score is shown as a function of $s/s_{\max}$. It is seen that a rather small s is chosen ($s/s_{\max} = 0.13$). The resulting regression coefficients on the different levels of pollution are shown in Table 3 (values are back-transformed to the original scale for better interpretation). Categories with effects that differ from the other categories are underlined. All predictors are excluded, except age, body weight/height, and sex. The intercept and the effect of body height, however, additionally vary with the degree of pollution. If a child lives in a zone of high pollution his/her lung capacity is identified as being

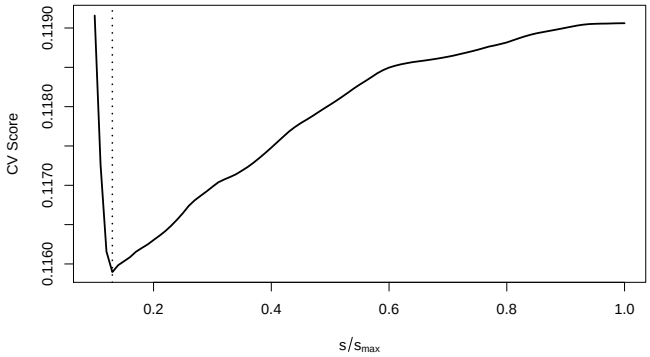

Figure 8: 5-fold cross-validation score as a function of $s/s_{\max}$ given the data from Table 2 and the degree of pollution taken as an effect modifying factor; s with minimum error is marked by the dotted line.

lower. The fitted difference to children being exposed to ozone is about 40 ml, for example. But also if the child is exposed to ozone there is a negative effect, compared to non/slightly polluted zones. According to the model, a child of 1.50 m (for example) that lives in a slightly polluted area has 4.5 ml higher lung capacity than a child that is highly exposed to ozone; the difference to highly polluted zones is even 44.5 ml. Since there are more than 1300 observations available and the minimum of the cross-validation score is well-defined, results can be supposed to be reliable. Moreover, the fit is quite good. The ratio of residual and total sum of squares is just 11.4%.

6 Generalizations to Multiple Effect Modifiers

A problem of the model fitted above to explain lung capacity is that the sex of children is only included as a main effect. However, it is questionable that the difference in lung capacity between boys and girls (of about 160 ml) remains constant, regardless if the children are eight or twelve years old, for example. Therefore the effects of covariates should also be allowed to vary with sex.

Because in many applications there is not only one potential effect modifying factor, in the following it is shown how a model of such type can be specified and regularized. That means, models with multiple categorial effects modifiers are considered.

Suppose there are two predictors x_1 and x_2 given, and two potential (categorial) effect modifiers $u_1 \in \{1, \ldots, k_1\}$ and $u_2 \in \{1, \ldots, k_2\}$. Then a possible model is

$$\eta(x, u) = \beta_{01}(u_1) + \beta_{02}(u_2) + x_1\beta_{11}(u_1) + x_1\beta_{12}(u_2) + x_2\beta_{21}(u_1) + x_2\beta_{22}(u_2).$$

	highly polluted	slightly polluted	ozone exposure
Intercept	-3.35632	-3.31679	-3.31679
Age	0.00017	0.00017	0.00017
Body Weight	0.01689	0.01689	0.01689
Body Height	0.03703	0.03706	0.03703
Sex	-0.15720	-0.15720	-0.15720
Parental Education	0.00000	0.00000	0.00000
Allergies	0.00000	0.00000	0.00000
Respiratory Diseases	0.00000	0.00000	0.00000
Smoking Mother	0.00000	0.00000	0.00000
Smoking Father	0.00000	0.00000	0.00000
Frequent Colds	0.00000	0.00000	0.00000
Frequent Coughs	0.00000	0.00000	0.00000
Lung Diseases	0.00000	0.00000	0.00000

Table 3: Fitted coefficients if the degree of pollution at the place of residence of a child is taken as a (potentially) effect modifying factor when explaining lung capacity; actually varying coefficients are underlined.

That means, for the varying functions $\beta_j(u_1, u_2)$ an additive structure is assumed:

$$\beta_j(u_1, u_2) = \beta_{j1}(u_1) + \beta_{j2}(u_2), \tag{11}$$

with

$$\beta_{j1}(u_1) = \sum_{r=1}^{k_1} \beta_{j1r} I(u_1 = r) \quad \text{and} \quad \beta_{j2}(u_2) = \sum_{s=1}^{k_2} \beta_{j2s} I(u_2 = s).$$

For means of identifiability, functions $\beta_{j2}(\cdot)$ need to be restricted, for example by

$$\beta_{j21} = 0, \ j = 0, \ldots, p.$$

The applied penalty is of the type

$$J(\beta) = \sum_{m \in \{1,2\}} \sum_{j=0}^{p} \sum_{r>s} w_{rs(j,m)} |\beta_{jmr} - \beta_{jms}| + \sum_{j=1}^{p} \sum_{r=1}^{k_1} \sum_{s=1}^{k_2} v_{rs(j,1,2)} |\beta_{j1r} + \beta_{j2s}|,$$

with adequately chosen weights $w_{rs(j,m)}$ and $v_{rs(j,1,2)}$ (for example taking ols estimates into account as done in Section 3). Penalization of terms $|\beta_{j1r}|$ is implicitly included in the second part of penalty J, because of restriction $\beta_{j21} = 0$ for all j. For the same reason, terms $|\beta_{j2r}|$ do not need to be explicitly penalized, since they are implicitly included in the first part of the penalty.

For illustration, we use the data from Table 2 again, but only consider covariates which showed relevant effects in Table 3. In Table 4 fitted coefficients are given if the degree of pollution at the place of residence as well as the child's sex

		highly polluted	slightly polluted	ozone exposure
Intercept	male	-3.52686	-3.49377	-3.49377
	female	-2.82959	-2.79650	-2.79650
Age	male	0.00000	0.00000	0.00000
	female	0.00340	0.00340	0.00340
Body Weight	male	0.02063	0.02088	0.02063
	female	0.01252	0.01277	0.01252
Body Height	male	0.03744	0.03747	0.03744
	female	0.03044	0.03047	0.03044

Table 4: Fitted coefficients as defined in (11) if the degree of pollution as well as the child's sex are taken as (potentially) effect modifying factors when explaining lung capacity. Additive terms coefficients are build of are found in Table 5.

	Intercept	Age	Body Weight	Body Height
highly polluted	-3.52686	0.00000	0.02062	0.03744
slightly polluted	-3.49377	0.00000	0.02088	0.03747
ozone exposure	-3.49377	0.00000	0.02062	0.03744
female	0.69727	0.00340	-0.00811	-0.00700

Table 5: Fitted coefficients $\hat{\beta}_{j,\text{zone},r}$ and $\hat{\beta}_{j,\text{sex},s}$, i.e. the degree of pollution as well as the child's sex are taken as (potentially) effect modifying factors when explaining lung capacity.

are taken as (potentially) effect modifying factors. The value $s/s_{\max} = 0.56$ is chosen via (5-fold) cross-validation again. As before, the estimated intercept is lower if the zone of residence is highly polluted. Moreover, the positive effect of body weight and body height is stronger if the area of residence is just slightly polluted. Because of the additive structure of $\beta_j(u_1, u_2)$ differences between intercepts – as well as differences between other coefficients – are the same for both males and females. For a better understanding, terms $\hat{\beta}_{j1}(u_1)$ and $\hat{\beta}_{j2}(u_2)$ are given in Table 5; u_1 stands for "zone", u_2 for "sex". Since $\beta_{j21} = 0\ \forall j$, for each covariate there is only an u_2-coefficient given for females – which is the difference between females and males (implicitly) already seen in Table 4. Since coefficients given in the last row of Table 5 are negative for body height and weight, resulting coefficients of body weight and body height are higher for males than for females (see also Table 4), which means that the (absolute) difference in lung capacity between boys and girls increases when children grow up. According to the fitted model, changes in lung capacities of boys are well explained by covariates body weight/height and the degree of pollution, whereas for girls there is also a (small)

effect of age. All in all, the ratio of residual and total sum of squares is 10.8%. That means, the fit of our first model from Table 3 is further improved.

7 Summary and Discussion

We showed how regularization can be used to obtain sparser representations of varying-coefficient models with categorial effect modifiers. Via penalizing absolute differences and L_1 norms of regression coefficients coefficients and differences thereof can be set exactly to zero. On the one hand, the proposed regularization technique may lead to stabilization and higher accuracy of estimates (as simulation studies showed); on the other hand, interpretability of the fitted models is increased. Via choosing an adequate penalty parameter, it is implicitly selected which coefficients should be set to zero and which coefficients should actually vary over different levels of the (potentially) effect modifying factor. If weights of penalty terms are included and adaptively chosen (e.g. dependently on ols estimates), selection consistency is obtained, as already shown by Zou (2006) for the original Lasso. But also in the finite case the adaptive version has the potential to outperform the standard (non-adaptive) version in both estimation/prediction accuracy and model selection, as shown in simulation studies.

The analysis of real data sets showed that the proposed method can also be successfully applied in practice. On the one hand, it turned out that gender is an important effect modifying factor when income is explained by several other covariates, as age or marital status. On the other hand, model complexity was distinctly reduced by the presented method when modeling lung capacities of schoolchildren. In the latter case most covariates have been excluded from the model or coefficients have been set as constant over two or more levels of the effect modifying factor 'degree of pollution', but still indicating that environmental pollution or ozone exposure at the place of residence tend to have negative effects on lung capacity. Finally, it was demonstrated how the proposed method can be generalized to the case of multiple (categorial) effect modifiers.

Varying coefficient models are frequently used if measurements are repeatedly taken. If regression coefficients are allowed to vary with time, time-dependent effects of covariates can be studied. Since time t is a continuous quantity it is commonly considered as a continuous effect modifier and $\beta_j(t)$ is fitted as a smooth function. In some cases, however, measurement points are fixed and equal for all observations, for example if a certain quantity is measured every day at the same time, or every week. That means, time points t are fixed, and the model becomes $y_t = \beta_{t0} + x_t^T \beta_t + \epsilon_t$. Then time can be seen as a discrete and ordered effect modifier, and the proposed regularization technique applies. The result is a set of piecewise constant functions $\beta_j(t)$, each comparable to a Fused Lasso (Tibshirani et al., 2005) estimate. The attractive feature of the method is that locations of relevant changes – or 'jumps' – in $\beta_j(\cdot)$ are identified. So

it may be said, for example, that a relevant change occurs between the second and the third day. The only difference to the setting investigated before is in the structure of ϵ_t. Since ϵ_t and ϵ_{t-1} cannot be assumed to be independent, there is within-subject correlation. With the same arguments as given by Wang et al. (2008), this correlation may be disregarded and estimation may be done with 'working independence correlation structure'. A more sensible procedure, however, is to assume a certain correlation structure like autocorrelation. With correlation matrix W, data are transformed to remove within-subject correlation. That means, the penalized least squares criterion from (2) becomes $Q_p(\beta) = (y - Z\beta)^T W^{-1}(y - Z\beta) + \lambda J(\beta)$. If autocorrelation parameter ρ, for example, is unknown, the working independence assumption can be used as a first step to estimate ρ and W respectively; and this procedure is possibly iterated.

Acknowledgements

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

Appendix

Proof of Proposition 1: If $\hat{\beta}$ minimizes $Q_p(\beta)$ from (2), then it also minimizes $Q_p(\beta)/n$. The ordinary least squares estimator $\hat{\beta}^{(LS)}$ minimizes $Q(\beta) = (y-Z\beta)^T(y-Z\beta)$, 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 $n_r/n \to c_r$, with $0 < c_r < 1\ \forall r$.

Proof of Proposition 2: We first show asymptotic normality, which closely follows Zou (2006) and Bondell and Reich (2009). Coefficient vector β (as given in (2)) is represented by $b = \sqrt{n}(\beta - \beta^*)$, i.e. $\beta = \beta^* + b/\sqrt{n}$, where β^* denotes the true coefficient vector. Then we also have $\hat{\beta} = \beta^* + \hat{b}/\sqrt{n}$, with

$$\hat{b} = \text{argmin}_b \Psi_n(b),$$

where

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

with

$$\begin{aligned} J(b) &= \sum_{j=0}^{p}\sum_{r>s}\sqrt{n}\frac{\phi_{rs(j)}(n)}{|\hat{\beta}_{jr}^{(LS)}-\hat{\beta}_{js}^{(LS)}|}\left|\beta_{jr}^{*}-\beta_{js}^{*}+\frac{b_{jr}-b_{js}}{\sqrt{n}}\right| \\ &+ \sum_{j=1}^{p}\sum_{r=1}^{k}\sqrt{n}\frac{\phi_{r(j)}(n)}{|\hat{\beta}_{jr}^{(LS)}|}\left|\beta_{jr}^{*}+\frac{b_{jr}}{\sqrt{n}}\right|. \end{aligned}$$

Furthermore, since $y - Z\beta^{*} = \epsilon$, we have $\Psi_n(b) - \Psi_n(0) = V_n(b)$, where

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

with

$$\begin{aligned} \widetilde{J}(b) &= \sum_{j=0}^{p}\sum_{r>s}\sqrt{n}\frac{\phi_{rs(j)}(n)}{|\hat{\beta}_{jr}^{(LS)}-\hat{\beta}_{js}^{(LS)}|}\left(\left|\beta_{jr}^{*}-\beta_{js}^{*}+\frac{b_{jr}-b_{js}}{\sqrt{n}}\right|-|\beta_{jr}^{*}-\beta_{js}^{*}|\right) \\ &+ \sum_{j=1}^{p}\sum_{r=1}^{k}\sqrt{n}\frac{\phi_{r(j)}(n)}{|\hat{\beta}_{jr}^{(LS)}|}\left(\left|\beta_{jr}^{*}+\frac{b_{jr}}{\sqrt{n}}\right|-|\beta_{jr}^{*}|\right). \end{aligned}$$

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

$$|\hat{\beta}_{jr}^{(LS)}| \to_p |\beta_{jr}^{*}|, \text{ and } \sqrt{n}\left(\left|\beta_{jr}^{*}+\frac{b_{jr}}{\sqrt{n}}\right|-|\beta_{jr}^{*}|\right) = b_{jr}\,\mathrm{sgn}(\beta_{jr}^{*}) \text{ (if } n \text{ large enough)};$$

and similarly, if $\beta_{jr}^{*} \neq \beta_{js}^{*}$,

$$|\hat{\beta}_{jr}^{(LS)}-\hat{\beta}_{js}^{(LS)}| \to_p |\beta_{jr}^{*}-\beta_{js}^{*}|, \text{ and}$$

$$\sqrt{n}\left(\left|\beta_{jr}^{*}-\beta_{js}^{*}+\frac{b_{jr}-b_{js}}{\sqrt{n}}\right|-|\beta_{jr}^{*}-\beta_{js}^{*}|\right) = (b_{jr}-b_{js})\,\mathrm{sgn}(\beta_{jr}^{*}-\beta_{js}^{*});$$

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

$$\frac{\lambda_n}{\sqrt{n}}\frac{\phi_{r(j)}(n)}{|\hat{\beta}_{jr}^{(LS)}|}\sqrt{n}\left(\left|\beta_{jr}^{*}+\frac{b_{jr}}{\sqrt{n}}\right|-|\beta_{jr}^{*}|\right) \to_p 0, \text{ and}$$

$$\frac{\lambda_n}{\sqrt{n}}\frac{\phi_{rs(j)}(n)}{|\hat{\beta}_{jr}^{(LS)}-\hat{\beta}_{js}^{(LS)}|}\sqrt{n}\left(\left|\beta_{jr}^{*}-\beta_{js}^{*}+\frac{b_{jr}-b_{js}}{\sqrt{n}}\right|-|\beta_{jr}^{*}-\beta_{js}^{*}|\right) \to_p 0, \text{ respectively.}$$

If $\beta_{jr}^{*} = 0$ or $\beta_{jr}^{*} = \beta_{js}^{*}$, however,

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

$$\sqrt{n}\left(\left|\beta^*_{jr}-\beta^*_{js}+\frac{b_{jr}-b_{js}}{\sqrt{n}}\right|-|\beta^*_{jr}-\beta^*_{js}|\right)=|b_{jr}-b_{js}|,\ \text{respectively.}$$

Moreover, if $\beta^*_{jr}=0$ or $\beta^*_{jr}=\beta^*_{js}$, due to $\sqrt{n}$-consistency of the ordinary least squares estimate (which is ensured by condition $n_r/n\to c_r$, $0<c_r<1\ \forall r$),

$$\lim_{n\to\infty}P(\sqrt{n}|\hat{\beta}^{(LS)}_{jr}|\le\lambda_n^{1/2})=1,\ \text{resp.}\ \lim_{n\to\infty}P(\sqrt{n}|\hat{\beta}^{(LS)}_{jr}-\hat{\beta}^{(LS)}_{js}|\le\lambda_n^{1/2})=1,$$

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

$$\frac{\lambda_n}{\sqrt{n}}\frac{\phi_{r(j)}(n)}{|\hat{\beta}^{(LS)}_{jr}|}\sqrt{n}\left(\left|\beta^*_{jr}+\frac{b_{jr}}{\sqrt{n}}\right|-|\beta^*_{jr}|\right)\to_p\infty,\ \text{or}$$

$$\frac{\lambda_n}{\sqrt{n}}\frac{\phi_{rs(j)}(n)}{|\hat{\beta}^{(LS)}_{jr}-\hat{\beta}^{(LS)}_{js}|}\sqrt{n}\left(\left|\beta^*_{jr}-\beta^*_{js}+\frac{b_{jr}-b_{js}}{\sqrt{n}}\right|-|\beta^*_{jr}-\beta^*_{js}|\right)\to_p\infty,$$

if $b_{jr}\neq 0$, resp. $b_{jr}\neq b_{js}$. That means, if for any r,s,j with $\beta^*_{jr}=\beta^*_{js}$ $(j\ge 0)$ or $\beta^*_{jr}=0$ $(j>0)$, $b_{jr}\neq b_{js}$ or $b_{jr}\neq 0$, respectively, then $(\lambda_n/\sqrt{n})\widetilde{J}(b)\to_p\infty$. The rest of the proof of part (a) is similar to Bondell and Reich (2009). Let Z^* denote the design matrix corresponding to the correct structure, i.e. columns of variables with equal coefficients on different levels of u are added and collapsed, and columns corresponding to zero coefficients are removed. Since $\forall r\ n_r/n\to c_r$ $(0<c_r<1)$,

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

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

$$V(b)=\begin{cases} b^T_{\mathcal{C}}Cb_{\mathcal{C}}-2b^T_{\mathcal{C}}w & \text{if } \theta_{\mathcal{C}^c}=0\\ \infty & \text{otherwise.}\end{cases}$$

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

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

Hence, $\hat{b}_{\mathcal{C}}\to_d N(0,\sigma^2C^{-1})$. Via a reparametrization of β as, for example, $\tilde{\beta}=(\tilde{\beta}_0^T,\ldots,\tilde{\beta}_p^T)^T$, with $\tilde{\beta}_j=(\beta_{jr}-\beta_{j1},\ldots,\beta_{jr},\ldots,\beta_{jr}-\beta_{jk})^T$, i.e. changing the subset of entries of θ which are part of β, resp. b, asymptotic normality can be proven for all entries of $\hat{\theta}_{\mathcal{C}}$.

To show the consistency part, we first note that $\lim_{n\to\infty}P(\mathfrak{I}\in\mathcal{C}_n)=1$, if $\mathfrak{I}\in\mathcal{C}$, follows from part (a), where $\mathfrak{I}$ denotes a triple of indices (j,r,s) or pair

(j,r). We will now show that if $\mathfrak{I} \notin \mathcal{C}$, $\lim_{n\to\infty} P(\mathfrak{I} \in \mathcal{C}_n) = 0$. A similar proof is found in Bondell and Reich (2009). Let $\mathcal{B}_n$ denote the (nonempty) set of indices $\mathfrak{I}$ which are in $\mathcal{C}_n$ but not in $\mathcal{C}$. With out loss of generality we assume that the largest $\hat{\theta}$-entry corresponding to indices from $\mathcal{B}_n$ is $\hat{\beta}_{lq} > 0$, $l \geq 1$. If a certain difference $\hat{\beta}_{lr} - \hat{\beta}_{ls}$ is the largest $\hat{\theta}$-entry included in $\mathcal{B}_n$ we just need to reparameterize β_l in an adequate way by $\tilde{\beta}_l$ as given above. Since all coefficients and differences thereof are penalized in the same way this can be done without any problems. If $l = 0$, the reparametrization means choosing a reference category whose intercept is not penalized. In this case the proof is analogue to Gertheiss and Tutz (2009).

Moreover, we may order categories such that $\hat{\beta}_{l1} \leq \ldots \leq \hat{\beta}_{lz} \leq 0 \leq \hat{\beta}_{l,z+1} \leq \ldots \leq \hat{\beta}_{lk}$. That means estimate $\hat{\beta}$ from (2) with penalty (7) is equivalent to

$$\hat{\beta} = \text{argmin}_{\mathfrak{B}} \left\{ (y - Z\beta)^T (y - Z\beta) + \lambda_n \sum_j J_j(\beta) \right\}$$

with

$$\mathfrak{B} = \{\beta : \ \beta_{01}, \ldots, \beta_{l-1,k}, \beta_{l1} \leq \ldots \leq \beta_{lz} \leq 0 \leq \beta_{l,z+1} \leq \ldots \leq \beta_{lk}, \beta_{l+1,1}, \ldots, \beta_{pk}\},$$

$$J_j(\beta) = \sum_{r>s} \frac{\phi_{rs(j)}(n)}{|\hat{\beta}_{jr}^{(LS)} - \hat{\beta}_{js}^{(LS)}|} |\beta_{jr} - \beta_{js}| + I(j \neq 0) \sum_{r=1}^{k} \frac{\phi_{r(j)}(n)}{|\hat{\beta}_{jr}^{(LS)}|} |\beta_{jr}|, \ j \neq l$$

and

$$J_l(\beta) = \sum_{r>s} \phi_{rs(l)}(n) \frac{\beta_{lr} - \beta_{ls}}{|\hat{\beta}_{lr}^{(LS)} - \hat{\beta}_{ls}^{(LS)}|} + \sum_{r \geq z+1} \phi_{r(l)}(n) \frac{\beta_{lr}}{|\hat{\beta}_{lr}^{(LS)}|} - \sum_{r \leq z} \phi_{r(l)}(n) \frac{\beta_{lr}}{|\hat{\beta}_{lr}^{(LS)}|}.$$

Since $\hat{\beta}_{lq} \neq 0$ is assumed, at the solution $\hat{\beta}$ this optimization criterion is differentiable with respect to β_{lq}. We may consider this derivative in a neighborhood of the solution where coefficients which are set equal/to zero remain equal/zero. That means, terms corresponding to pairs/triples of indices which are not in C_n can be omitted, since they will vanish in $J(\hat{\beta}) = \sum_j J_j(\hat{\beta})$. If $z_{(l)q}$ denotes the column of design matrix Z which belongs to β_{lq}, due to differentiability, estimate $\hat{\beta}$ must satisfy

$$\frac{Q_q'(\hat{\beta})}{\sqrt{n}} = \frac{2 z_{(l)q}^T (y - Z\hat{\beta})}{\sqrt{n}} = A_n + D_n,$$

with

$$A_n = \frac{\lambda_n}{\sqrt{n}} \left(\sum_{s<q;(l,q,s)\in\mathcal{C}} \frac{\phi_{qs(l)}(n)}{|\hat{\beta}_{lq}^{(LS)} - \hat{\beta}_{ls}^{(LS)}|} - \sum_{r>q;(l,r,q)\in\mathcal{C}} \frac{\phi_{rq(l)}(n)}{|\hat{\beta}_{lr}^{(LS)} - \hat{\beta}_{lq}^{(LS)}|} \right)$$

and

$$D_n = \frac{\lambda_n}{\sqrt{n}} \sum_{s<q;(l,q,s)\in\mathcal{B}_n} \frac{\phi_{qs(l)}(n)}{|\hat{\beta}_{lq}^{(LS)} - \hat{\beta}_{ls}^{(LS)}|} + \frac{\phi_{q(l)}(n)}{|\hat{\beta}_{lq}^{(LS)}|}.$$

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

$$\frac{Q_q'(\hat{\beta})}{\sqrt{n}} = \frac{2z_{(l)q}^T(y - Z\hat{\beta})}{\sqrt{n}} = \frac{2z_{(l)q}^T Z\sqrt{n}(\beta^* - \hat{\beta})}{n} + \frac{2z_{(l)q}^T\epsilon}{\sqrt{n}}.$$

From part (a) and applying Slutsky's theorem, we know that $2z_{(l)q}^T Z\sqrt{n}(\beta - \hat{\beta})/n$ has some asymptotic normal distribution with mean zero, and $2z_{(l)q}^T\epsilon/\sqrt{n}$ 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

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

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

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

As a consequence

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

That means if $\mathfrak{I} \notin \mathcal{C}$, also

$$\lim_{n\to\infty} P(\mathfrak{I} \in \mathcal{C}_n) = 0.$$

References

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

Cardot, H. and P. Sarda (2008). Varying-coefficient functional linear regression models. *Communications in Statistics – Theory and Methods 37*, 3186–3203.

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

Fan, J., Q. Yao, and Z. Cai (2003). Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society B 65*, 57–80.

Gertheiss, J. and G. Tutz (2009). Sparse modeling of categorial explanatory variables. Technical Report 60, Department of Statistics, Ludwig-Maximilians-Universität München. (submitted).

Hand, D. J., F. Daly, K. McConway, D. Lunn, and E. Ostrowski (Eds.) (1994). *A Handbook of Small Data Sets*. London: Chapman & Hall.

Hastie, T. and R. Tibshirani (1993). Varying-coefficient models. *Journal of the Royal Statistical Society B 55*, 757–796.

Hastie, T., R. Tibshirani, and J. H. Friedman (2001). *The Elements of Statistical Learning*. New York: Springer.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*, 55–67.

Hofner, B., T. Hothorn, and T. Kneib (2008). Variable selection and model choice in structured survival models. Technical Report 43, Department of Statistics, Ludwig-Maximilians-Universität München.

Hoover, D. R., J. A. Rice, C. O. Wu, and L.-P. Yang (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika 85*, 809–822.

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.

Kauermann, G. and G. Tutz (2000). Local likelihood estimation in varying coefficient models including additive bias correction. *Journal of Nonparametric Statistics 12*, 343–371.

Kim, M.-O. (2007). Quantile regression with varying coefficients. *Annals of Statistics 35*, 92–108.

Leng, C. (2009). A simple approach for varying-coefficient model selection. *Journal of Statistical Planning and Inference 139*, 2138–2146.

Lu, Y., R. Zhang, and L. Zhu (2008). Penalized spline estimation for varying-coefficient models. *Communications in Statistics – Theory and Methods 37*, 2249–2261.

Mu, Y. and Y. Wei (2009). A dynamic quantile regression transformation model for longitudinal data. *Statistica Sinica 19*, 1137–1153.

Qu, A. and R. Li (2006). Quadratic inference functions for varying-coefficient models with longitudinal data. *Biometrics 62*, 379–391.

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.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (Fourth ed.). New York: Springer.

Wang, H. and Y. Xia (2009). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association 104*, 747–757.

Wang, L., H. Li, and J. Z. Huang (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association 103*, 1556–1568.

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.