Statistica Sinica **22** (2012), 957-982
doi:http://dx.doi.org/10.5705/ss.2010.177

REGULARIZATION AND MODEL SELECTION WITH CATEGORIAL EFFECT MODIFIERS

Jan Gertheiss and Gerhard Tutz

Ludwig-Maximilians-Universität Munich

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. A distinction is made between nominal and ordinal variables, since for the latter more economic parameterizations are warranted. The proposed methods are illustrated and investigated in simulation studies and data evaluations. Moreover, some asymptotic properties are derived.

Key words and phrases: Categorial predictors, fused lasso, linear model, variable selection, varying-coefficient models.

1. Introduction

Varying-coefficient models (Hastie and Tibshirani (1993)) offer a 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, say

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

where the $\beta_j(u)$ 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 has been investigated thoroughly, in the linear model as given above and in other situations (see for example Cardot and 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, Zhang, and Zhu (2008)) or localizing techniques (Fan, Yao, and Cai (2003)); or Kauermann and Tutz (2000)). Recently, Wang, Li, and Huang (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 $\beta_j(u) = 0$.

In the present paper methods for categorial effect modifiers are proposed for the classical linear model. The main problem with categorial effect modifiers is that the number of parameters to be estimated and – more importantly – the number of potential models may be 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,

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

contains $(p + 1)k$ parameters. The interpretation is that on level r of u the model $y = \beta_{0r} + x_1\beta_{1r} + \cdots + x_p\beta_{pr} + \epsilon$ 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 variable selection with respect to predictors x_j and investigation if functions $\beta_j(\cdot)$ are (partially) constant; 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 in that predictors may be excluded (i.e. corresponding coefficients are set to zero) for specific values of u only. Of course, all potential models could also be estimated using pure maximum likelihood/ordinary least squares estimation and model selection may be based on information criteria like AIC or BIC. Such complete enumeration, however, is feasible only for data sets with a very small number of covariates $x_1, \ldots, x_p$ and a very small k (denoting the number of levels of u). More precisely, the exact number of potential models is

$$M(p, k) = C(k)\Big(1 + \sum_{s=1}^{k} \binom{k}{s} C(s)\Big)^p, \tag{1.1}$$

with

$$C(s) = \sum_{v=1}^{s} S(s, v), \text{ and } S(s, v) = \frac{1}{v!} \sum_{l=1}^{v} (-1)^{v-l} \binom{v}{l} l^s. \tag{1.2}$$

A derivation of (1.1) is given in the Appendix. For example $M(7, 2) = 156,250$, which may still be feasible. But $M(p, k)$ grows very rapidly with growing p and k; so already $M(10, 3) > 10^{12}$. In cases like this, stepwise procedures like forward or backward selection (based e.g., on the AIC) could be used. However, these suffer from high variability (cf., Hastie, Tibshirani and Friedman (2009)). Hence, regularization techniques that induce sparsity are a promising approach for model

selection. Beside stability, the advantage of regularization techniques is that the extent of regularization – and hence sparsity – is typically controlled by a tuning parameter. Selection of that parameter implicitly determines the model.

A well-known regularization technique designed for grouped variables is the Group Lasso (Yuan and Lin (2006)), which can be used to select pre-specified groups of variables. When applied to the considered regression problem, coefficients $\beta_{j1}, \ldots, \beta_{jk}$ may be set to zero simultaneously, which means variable selection with respect to predictors x_j. However, due to the Ridge-type penalty within groups of coefficients, in Group Lasso estimated coefficients $\hat{\beta}_{jr}$ and $\hat{\beta}_{js}$ cannot be enforced to be equal.

An example shows that our 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 weekly gas consumption (in 1,000 cubic feet) and average external temperature (in degree C) at Whiteside's house in south-east England during two 'heating seasons' – one before and one after cavity-wall insulation was installed, cf., Venables and Ripley (2002). The most complex model used there fits gas consumption as a quadratic function of temperature separately for both seasons before and after insulation. Thus, 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 $\eta(x, u = r) = \beta_{0r} + x\beta_{1r} + x^2\beta_{2r} = E(y|x, u = r)$ for fixed r. In Figure 1 the data are shown along with estimated regression curves. Though results look quite similar, our model has one degree of freedom less since the coefficients of the quadratic term (β_{21} and β_{22}) are set equal for the two 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 introduce the method and discuss some computational aspects in Section 2. Large sample properties are investigated in Section 3, and the proposed methods are tested in simulation studies reported on in Section 4. In Sections 5 and 6, some 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 write $\beta_j(u) = \beta_{ju}$. Instability of the ordinary least squares estimate can be avoided by penalized estimation:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\ Q_p(\beta), \tag{2.1}$$

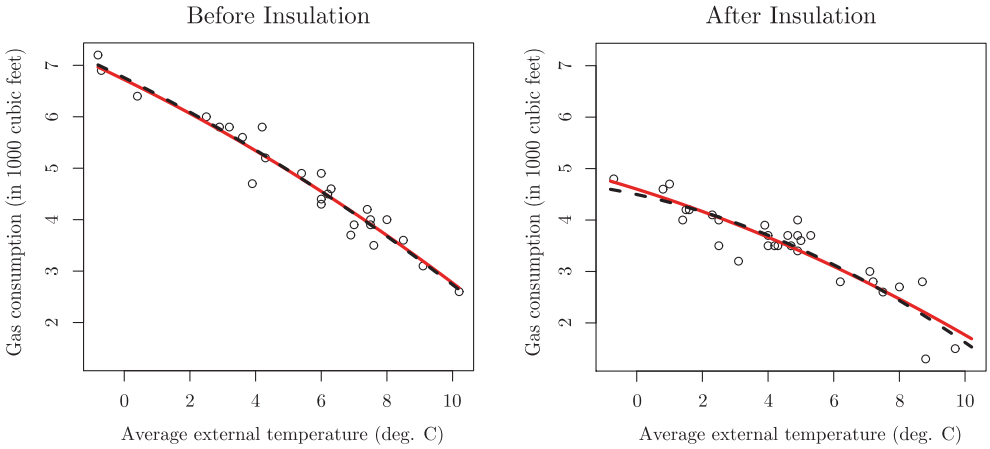

Figure 1. Whiteside's data showing the effect of insulation on household gas consumption with estimated quadratic regression curves; dashed lines refer to the full model, solid ones to the regularized model with coefficients of quadratic terms set equal.

with

$$Q_p(\beta) = \sum_{i=1}^{n} \Big(y_i - \beta_0(u_i) - \sum_{j=1}^{p} x_{ij}\beta_j(u_i) \Big)^2 + \lambda J(\beta)$$
$$= (y - Z\beta)^T (y - Z\beta) + \lambda J(\beta), \tag{2.2}$$

where $y = (y_1, \ldots, y_n)^T$ 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 the penalty term $J(\beta)$ one has ordinary least squares estimation. With increasing λ the influence of $J(\beta)$ is increased. The crucial point is to chose an adequate penalty. Classical penalties are the *Ridge* $J(\beta) = \sum_{j,r} \beta_{jr}^2$ (Hoerl and Kennard (1970)) and the *Lasso* $J(\beta) = \sum_{j,r} |\beta_{jr}|$ (Tibshirani (1996)). While the Ridge only shrinks estimates $\hat{\beta}_{jr}$ toward zero, the Lasso additionally allows some $\hat{\beta}_{jr}$ to be set to zero. 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 $r \neq s$, and this is needed to obtain potentially (piecewise) constant functions $\hat{\beta}_j(u)$. In the following we present an approach that 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 the penalty

$$J(\beta) = \sum_{j=0}^{p} \sum_{r>s} |\beta_{jr} - \beta_{js}| + \sum_{j=1}^{p} \sum_{r=1}^{k} |\beta_{jr}|. \tag{2.3}$$

The first term enforces the collapsing of categories of the effect modifier. In the case of very strong penalization, the effects of covariates do not depend on the category of u and one obtains $\hat{\beta}_{j1} = \hat{\beta}_{j2} = \cdots = \hat{\beta}_{jk} = \hat{\beta}_{j.}$. The second term in (2.3) steers selection/exclusion of covariates. In the extreme case, $\hat{\beta}_{j1} = \hat{\beta}_{j2} = \cdots = \hat{\beta}_{jk} = 0$ is obtained and covariate x_j is omitted.

If u is ordinal, levels can be reasonably ordered. Here we use

$$J(\beta) = \sum_{j=0}^{p} \sum_{r=2}^{k} |\beta_{jr} - \beta_{j,r-1}| + \sum_{j=1}^{p} \sum_{r=1}^{k} |\beta_{jr}|. \tag{2.4}$$

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.

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

$$J(\beta; \psi) = \psi \sum_{j=0}^{p} \sum_{r>s} |\beta_{jr} - \beta_{js}| + (1 - \psi) \sum_{j=1}^{p} \sum_{r=1}^{k} |\beta_{jr}|, \tag{2.5}$$

or, depending on the scale level of u,

$$J(\beta; \psi) = \psi \sum_{j=0}^{p} \sum_{r=2}^{k} |\beta_{jr} - \beta_{j,r-1}| + (1 - \psi) \sum_{j=1}^{p} \sum_{r=1}^{k} |\beta_{jr}|. \tag{2.6}$$

The use of a flexible ψ means, however, that another tuning parameter is introduced, and it is not clear if this modification really has better performance than the use of (2.3) or (2.4), where $\psi = 0.5$ is fixed. This issue is investigated further in simulation studies in Section 4. In practice, tuning parameters like λ (or ψ) are typically selected using K-fold cross-validation, where $K = 5$ or $K = 10$ ist commonly chosen (cf., Hastie, Tibshirani and Friedman (2009)). Since we are in the context of the classical linear model, we select the tuning parameter minimizing the (cross-validated) squared error of prediction.

2.2. Computational issues

When computing estimates it is useful to consider (2.2) as the constrained optimization problem

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

where the tuning parameter s plays a role comparable to that of λ; see, for example, Hastie, Tibshirani and Friedman (2009). In matrix notation we have

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

with y and Z chosen as in (2.2). Furthermore, if $\nu = (\beta^T, \delta^T)^T$ with $\delta^T = (\delta_0^T, \ldots, \delta_p^T)$, and $\delta_j = (\beta_{j2} - \beta_{j1}, \beta_{j3} - \beta_{j1}, \ldots, \beta_{jk} - \beta_{j,k-1})^T$, we can write $\delta = D\beta$ with adequately chosen D. Then, for (2.3), $\hat{\beta}$ can be computed via

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

where $\|\nu\|_1$ denotes the L_1-norm of ν, a possible choice of U is $U = (Z|0)$, and $A = (D| - I)$ since $A\nu = D\beta - \delta = 0$. If every entry of ν is split into positive and negative parts, 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 (2010)). For (2.4), computation can be done in a completely analogous way. With a 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 in 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 that can be computed using methods for the usual Lasso estimation, as e.g., the 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. Let

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

Here the first penalty term, which is weighted by γ, penalizes violations of $A\nu = 0$. Since $\min_\nu h(\nu, \gamma)$ is a monotone function of γ bounded 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 $\hat{\nu}$ 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 large γ an acceptable approximation of $\hat{\nu}$ can be obtained as $\hat{\nu}_\gamma$. 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. When constructing the extended design matrix to compute the Lasso/Elastic Net solution (see, e.g., Zou and Hastie (2005)), A is multiplied by $\sqrt{\gamma}$. We usually take $\sqrt{\gamma} = 10^5$ or $\sqrt{\gamma} = 10^6$. To find an adequate γ-value and to judge precision in general, $\Delta_\gamma = (A\hat{\nu}_\gamma)^T A\hat{\nu}_\gamma$ can be used (with Δ_γ also depending on the chosen λ, resp. s). An upper bound for Δ_γ can be computed as was done in Gertheiss and Tutz (2010). In our analyses we mostly obtained Δ_γ-values of about 10^{-20} or better, comparable to results obtained by using the `kernlab` package. (Note also that if quadratic programming is used to compute "exact" solutions, constraints are just "numerically" met.) It has been our experience that the proposed algorithm works well as long as the number of levels k and the number of predictors p is modest ($k < 10$, $p < 30$). If k is small ($k = 2, 3$), p may be larger.

3. Large Sample Properties and Modifications

In the following we 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 $\beta_{jr} - \beta_{js}$.

Proposition 1. *Suppose $0 \leq \lambda < \infty$ is fixed, and all class-wise sample sizes n_r satisfy $n_r/n \to c_r$, where $0 < c_r < 1$. Then $\hat{\beta}$ from (2.1) with penalty (2.3) is consistent in that $\lim_{n\to\infty} P(||\hat{\beta} - \beta^*||^2 > \epsilon) = 0$ for all $\epsilon > 0$, with β^* denoting the vector of true coefficients.*

The proof is given in the Appendix. If u is ordinal, (2.4) is employed and consistency is proved in a completely analogous way. Employing the generalized version (2.5) or (2.6) 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 (and fusion) – the probability that the correct model is identified does not necessarily tend to one if the number of observations tends to infinity.

To counter 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 here. Given nominal u, we employ the adaptive penalty

$$J(\beta) = \sum_{j=0}^{p} \sum_{r>s} w_{rs(j)} |\beta_{jr} - \beta_{js}| + \sum_{j=1}^{p} \sum_{r=1}^{k} w_{r(j)} |\beta_{jr}| \tag{3.1}$$

with weights

$$w_{rs(j)} = \phi_{rs(j)}(n)\left|\hat{\beta}_{jr}^{(LS)} - \hat{\beta}_{js}^{(LS)}\right|^{-1} \text{ and } w_{r(j)} = \phi_{r(j)}(n)\left|\hat{\beta}_{jr}^{(LS)}\right|^{-1}, \qquad (3.2)$$

$\hat{\beta}_{jr}^{(LS)}$ 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 (2.5). In contrast to Proposition 1, λ at (2.2) is not fixed, but increases 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 (2.1) with penalty (3.1). 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. *If $\lambda = \lambda_n$ 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 for the penalty at* (3.1) *with weights* (3.2),

(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 (2010), 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 (see the proof). Since all estimated differences are (deterministic) linear functions of estimated parameters, Σ is singular. As seen from (b), the probability that the correct model is identified tends to one if the number of observations tends to infinity.

If the effect modifier u is ordinal, the weighting scheme and the asymptotic behavior of the corresponding estimator are completely analogous.

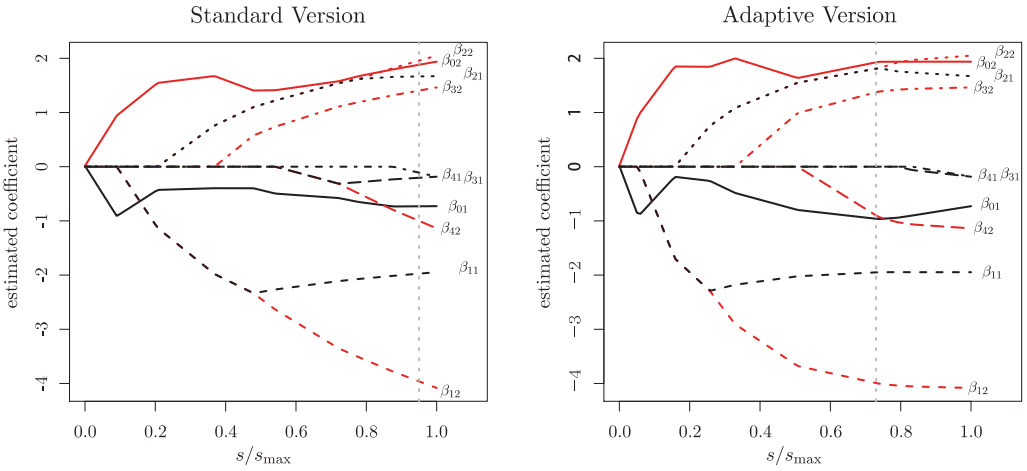

Figure 2. Fitted β-coefficients as functions of tuning parameter s for the standard penalty approach (left) and the adaptive version (right), s values chosen by cross-validation are marked by the vertical lines; true coefficient values are $\beta_{01} = -1$, $\beta_{02} = 1$, $\beta_{11} = -2$, $\beta_{12} = -4$, $\beta_{21} = \beta_{22} = 2$, $\beta_{31} = 0$, $\beta_{32} = 2$, $\beta_{41} = \beta_{42} = 0$.

4. Numerical Experiments

Before the presented approach is applied to data, we evaluate the method in simulation studies where the underlying model is known.

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 take model $y = -1-2x_1+2x_2+\epsilon$, and $y = 1-4x_1+2x_2+2x_3+\epsilon$ on level $u = 2$. Here the coefficients of x_2 do not vary with u, and x_3 is relevant only if $u = 2$. In addition, a pure noise variable x_4 is considered as a potential regressor in both models, i.e. $\beta_{41} = \beta_{42} = 0$. We generated $n = 200$ data points with x_{ij} independently $U[0, 1]$, class levels $u_1 = \cdots = u_{100} = 1$ and $u_{101} = \cdots = u_{200} = 2$, and standard normal error ϵ.

Figure 2 (left) shows fitted coefficient paths for all β_{jr} as functions of the tuning parameter s, when the standard (non-adaptive) approach with $\psi = 0.5$ was applied. At $s/s_{\max} = 1$ ordinary least squares (ols) estimates are obtained. With decreasing s (increasing penalty λ) coefficients are successively fused and shrunken toward zero. 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}; later, β_{41} and β_{42} are simultaneously set to zero, as desired. Later still, β_{11} and β_{12} are (wrongly) fused and non-zero coefficients are set to zero. Intercepts β_{01} and β_{02} are not fused until the minimal s is chosen. Since only their difference is penalized, at $s = 0$ they equal

$\bar{y}$ – the empirical mean of y. However, if cross-validation is applied to determine an adequate value of s (marked by the dotted vertical line), a model quite close to the ols estimate is chosen. As seen on the right hand side of Figure 2, the performance in terms of model selection was much better, when adaptive weights were used. In the latter case, cross-validation results in a model where just β_{42} is wrongly fitted as non-zero.

4.2. Comparison of methods

In order to further investigate the potential impact of modifications proposed in Sections 2 and 3, and to compare the proposed techniques to some alternative methods, we extend the simulation setting from above.

Simulation setting

Consider two additional levels of u and two additional predictors, with

$$\begin{aligned}
y &= -1 - 2x_1 + 2x_2 + 0x_3 + 0x_4 + 0x_5 + 0x_6 + \epsilon \quad \text{on level } u = 1,\\
y &= +1 - 4x_1 + 2x_2 + 2x_3 + 0x_4 + 0x_5 + 0x_6 + \epsilon \quad \text{on level } u = 2,\\
y &= +1 + 2x_1 + 2x_2 + 2x_3 - 4x_4 + 0x_5 + 0x_6 + \epsilon \quad \text{on level } u = 3,\\
y &= -1 + 1x_1 + 2x_2 + 3x_3 - 4x_4 - 2x_5 + 0x_6 + \epsilon \quad \text{on level } u = 4, \qquad (4.1)
\end{aligned}$$

where error ϵ is normal with mean zero and variance two. We independently generated 400 training and 1,200 test data data points, both with balanced u, and compared the standard and the adaptive version, both with fixed $\psi = 0.5$ as well as with ψ treated as another tuning parameter, that was chosen via 10-fold cross-validation. Parameter s was chosen via 10-fold cross-validation as well. This procedure of data generation, tuning parameter determination, model estimation, and evaluation was (independently) repeated 100 times. Results in terms of the (empiric) MSE of parameter estimates and prediction accuracies are found in Table 1 (left). Estimated standard errors are given in parentheses. Prediction accuracy is measured by the Mean Squared Error of Prediction (MSEP) on the test set. As a second scenario we introduced two additional pure noise input variables x_7 and x_8, and repeated the analysis. Results are also shown in Table 1.

Reference methods

The proposed regularization techniques are compared to the ordinary least squares estimate without model selection and a forward selection strategy based on the AIC or BIC (assuming normality of the response y). Checking all possible models is too computationally intensive (e.g., $M(8,4) \approx 10^{15}$). A standard forward selection applied to the design matrix of the ols model is not adequate

Table 1. Observed errors of parameter estimates (MSE) and predictions accuracy (MSEP), estimated standard errors are given parentheses; simulation scenario (4.1) (left), and with two additional pure noise input variables (right).

	scenario (4.1)				+ two noise covariates			
method	MSE		MSEP		MSE		MSEP	
ols	7.728	(0.264)	2.149	(0.010)	11.380	(0.380)	2.219	(0.011)
stdrd, fixed ψ	6.126	(0.218)	2.126	(0.010)	7.500	(0.240)	2.163	(0.010)
stdrd, flex. ψ	6.412	(0.229)	2.131	(0.010)	8.183	(0.455)	2.173	(0.010)
adapt, fixed ψ	5.855	(0.273)	2.117	(0.010)	6.920	(0.334)	2.149	(0.010)
adapt, flex. ψ	6.104	(0.293)	2.121	(0.010)	7.091	(0.334)	2.151	(0.010)
forward select, AIC	6.599	(0.302)	2.133	(0.010)	9.755	(0.414)	2.191	(0.011)
forward select, BIC	9.313	(0.489)	2.172	(0.013)	10.856	(0.698)	2.215	(0.016)

because it does not cover the fusion aspect. Therefore the procedure has to be extended. The forward selection strategy we used was as follows:

1. Start with the non-varying intercept model with one degree of freedom (df).
2. At each step increase df by one and check all models that are based on the model from the previous step in which, for $j = 1, \ldots, p$, a group of zero coefficients from $\{\beta_{j1}, \ldots, \beta_{jk}\}$ are changed, or a cluster of nonzero coefficients is split. Select the model with minimum AIC/BIC.
3. Stop if the (minimum) AIC/BIC does not decrease.

The degrees of freedom are defined (Tibshirani et al. (2005)) as the number of non-zero coefficient blocks in $\hat{\beta}$.

Results

From Table 1 it is seen that regularized approaches performed better than the ordinary least squares (ols) estimate in terms of accuracy of parameter estimation (i.e. MSE) and prediction (i.e. MSEP). They also seem to be superior to an AIC/BIC-based forward selection strategy without shrinkage. In addition, results of the stepwise procedures had higher variability, as seen from the MSE standard errors.

If a regularized approach is applied, using adaptive weights as at (3.2) seems to increase accuracy of parameter estimates and prediction. Allowing flexible ψ, by contrast, did not lead to better results.

In the second simulation we introduced two additional pure noise input variables. Then one might think that emphasis should be placed on the penalty's selection part, say $\psi < 0.5$ should be chosen in (2.5). Surprisingly, choosing ψ via cross-validation was not superior to using $\psi = 0.5$. When looking at Table 1, only differences between regularized and ordinary least squares estimates

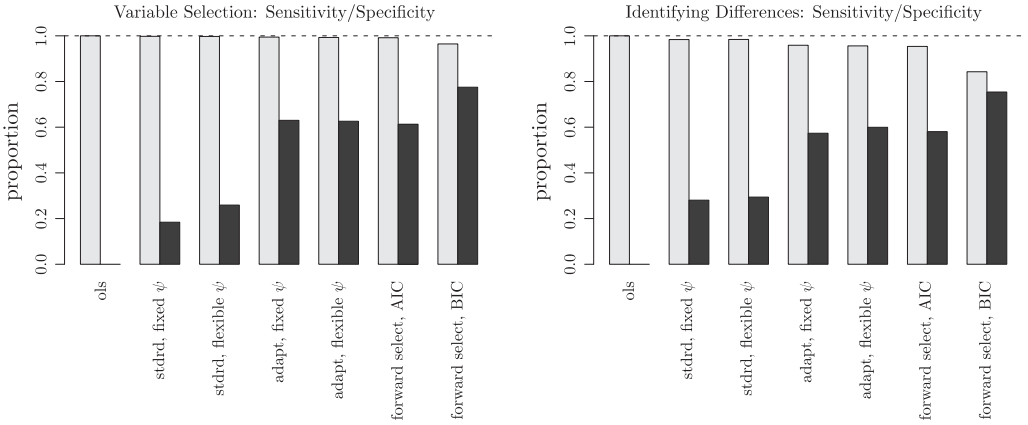

Figure 3. Sensitivity (light-colored) and specificity (dark) concerning variable selection and the identification of differences between regression coefficients belonging to the same predictor, simulation scenario (4.1).

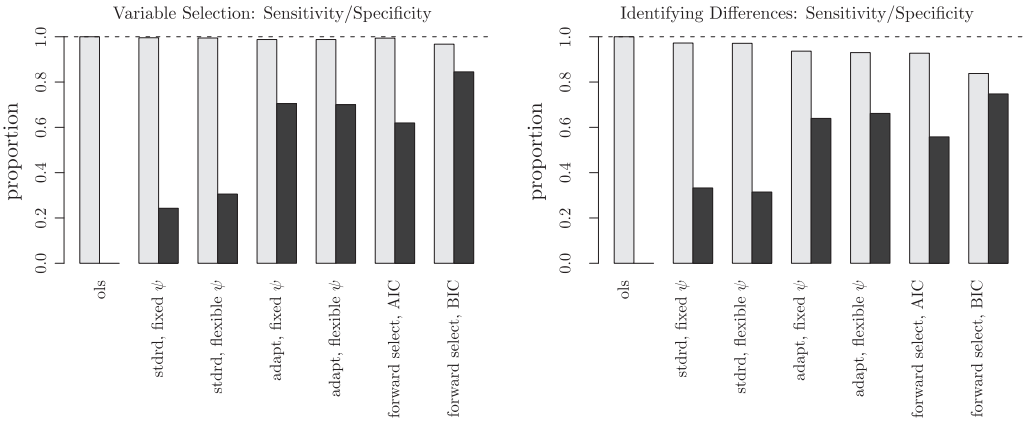

Figure 4. Sensitivity (light-colored) and specificity (dark) concerning variable selection and the identification of differences between regression coefficients belonging to the same predictor, simulation scenario (4.1) with two additional pure noise covariates.

seem larger than before. Using the regularized version with adaptive weights, for example, the MSE of the ols model was reduced by about 40%.

Beside accuracy of prediction and parameter estimation we examined selection and clustering performance. In Figure 3 and 4 we report sensitivities (light colored) and specificities (dark) of variable selection and identification of relevant differences between (potentially) varying coefficients. In our case, sensitivity is the proportion of truly non-zero coefficients from (4.1), resp. differences thereof, that are fitted as non-zero, while specificity is the proportion of truly zero coeffi-

Table 2. Available data for the analysis of the relationship between income and several explanatory 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

cients/differences that are set to zero. These are averages over all 100 simulation runs. If the standard regularized approach is applied, specificity is rather low, while using adaptive weights substantially increases specificity for both variable selection and clustering. Differences between fixed $\psi = 0.5$ and ψ chosen via cross-validation were rather negligible. However, cross-validation to determine the strength of regularization seems to work well. The method that had the highest specificity but the lowest sensitivity was the BIC-based forward selection strategy. Compared to the AIC, the BIC tends to penalize complex models more heavily, giving preferences to simpler/sparser models (cf., Hastie, Tibshirani and Friedman (2009)). This apparently leads to comparably high errors in Table 1.

5. Data Evaluation

In the Introduction, results of an analysis of Whiteside's data were shown. In the following we give analyses of income data from Germany and data collected in Austria during a study on the functioning of lungs of schoolchildren. As simulations have suggested, we fix $\psi = 0.5$. 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 analyzed the relationship of monthly income and several (potentially) explanatory variables. The data were 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 used data from 2002 in a cross-sectional analysis. Table 2 shows the response and predictors we considered 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 a person's gender. From former studies it is known

that the influence of age is rather quadratic than linear. Therefore Age^2 was also included. Thus

$$\begin{aligned}\log(\text{Income}) = {} & \beta_0(\text{Gender}) + \beta_1(\text{Gender})\text{Age} + \beta_2(\text{Gender})\text{Age}^2 \\ & + \beta_3(\text{Gender})\text{Tenure} + \beta_4(\text{Gender})\text{Height} \\ & + \beta_5(\text{Gender})\text{Married} + \beta_6(\text{Gender})\text{Abitur} \\ & + \beta_7(\text{Gender})\text{Blue-collar} + \epsilon. \end{aligned} \tag{5.1}$$

Figure 5 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 the reverse for women. Qualitatively speaking, effects of job tenure, Abitur 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 shorter 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 differences between men and women can be regarded as substantial, an adequate s-value is chosen via cross-validation. Since data were plentiful, we used 5-fold cross-validation, thus reducing variance (and without suffering from a large bias, cf., Hastie, Tibshirani and Friedman (2009)). The vertical dotted line in each path plot in Figure 5 indicates the s with minimum (quadratic) 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 were fit 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.

5.2. Lung capacities of schoolchildren

The data analyzed were 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 of whether the functioning of lungs is affected by industry-induced air pollution. The data came from a cross-sectional study in the district of Brixlegg (Austria). A summary of the data used is found in Table 3. We analyzed the relationship between the capacity of the lungs (in liters) and the provided covariates. The degree of environmental pollution at the place of residence was given as a categorial predictor with three

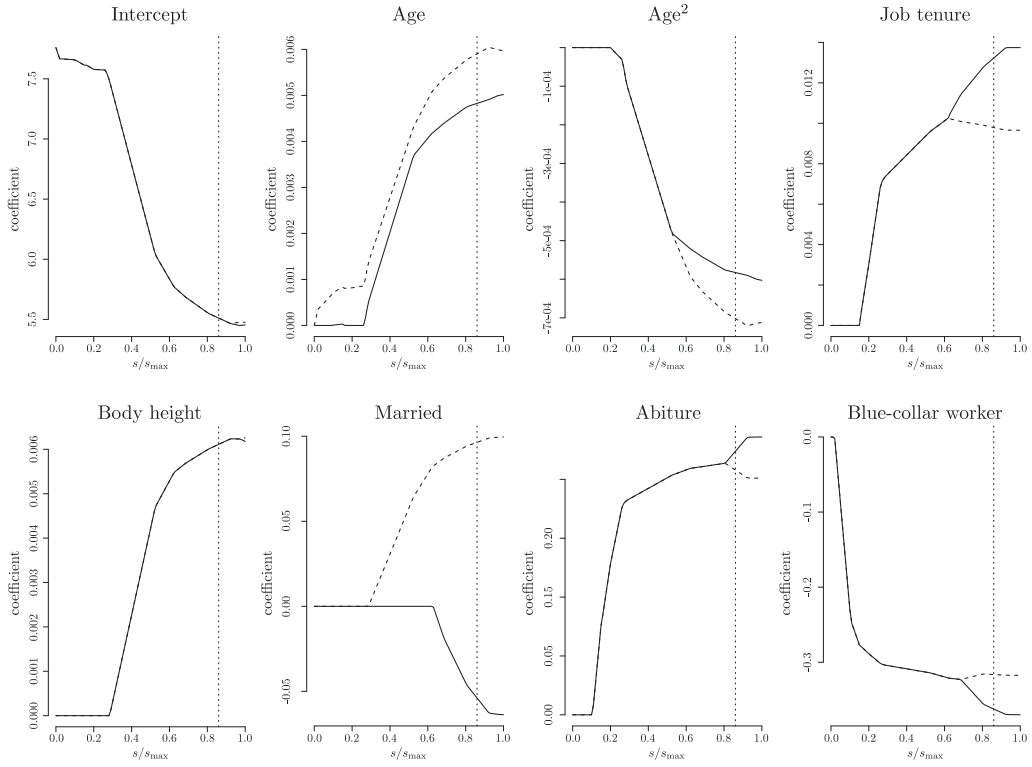

Figure 5. Paths of coefficients (possibly) varying with gender given model (5.1); dashed lines refer to males, solid ones to females, the vertical dotted line indicates coefficients at cv score minimizing $s/s_{\max} = 0.86$.

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 3.

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. When we fit main effect models – except, and then including pollution – the model was not significantly improved if the degree of pollution was taken into account (F-test based p-value 0.13). By contrast, if pollution was included as an effect modifying factor, the initial model was significantly improved (p-value 0.02). However, most regression parameters were far from being 'significantly non-zero'. So we used the proposed regularization technique to obtain a sparser representation.

Table 3. Available data for the analysis of the functioning of lungs of schoolchildren.

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 the mother smoke?	no/yes
	Does the 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 4. 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.

	highly polluted	slightly polluted	ozone exposure
Intercept	<u>-3.35632</u>	-3.31679	-3.31679
Age	0.00017	0.00017	0.00017
Body Weight	0.01689	0.01689	0.01689
Body Height	0.03703	<u>0.03706</u>	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

Via (5-fold) cross-validation a rather small s was chosen ($s/s_{\max} = 0.13$). The resulting regression coefficients on the different levels of pollution are shown in Table 4 (values are back-transformed to the original scale for better interpretation). Actually varying coefficients 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 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 small 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 (not shown), results can be supposed to be reliable. Moreover, the fit was quite good, with the ratio of residual and total sum of squares being 11.4%.

To obtain some measure of uncertainty, the asymptotic normal distribution from Proposition 2 was used. Resulting 90% confidence intervals for estimated nonzero regression coefficients are shown in Figure 6 (dashed). Computing intervals this way, however, is not recommended for three reasons: no intervals are available for regression coefficients (or differences thereof) that are set to zero; variability induced by the selection process and the determination of tuning parameters is ignored; the distribution of adaptive Lasso estimates may be quite far from normal (see Pötscher and Schneider (2009)). In Figure 6 (solid) we additionally give 90% bootstrap percentile intervals using the R add-on package `boot` (Canty and Ripley (2010); Davison and Hinkley (1997)). These intervals tend to be asymmetric and larger than those using the asymptotic normal. The influence of predictors with estimated nonzero coefficients is confirmed (with borderline significance of 'age'). Intervals of intercept values are clearly overlapping (top left), but intervals for intercept differences between zones (bottom left) indicate that a child living in a highly polluted zone has lower lung capacity. Since the difference between zones 2 and 3 is estimated as zero, no interval can be computed using the normal approximation. Still it can be noted that in the case of truly nonzero regression parameters being almost zero bootstrap confidence intervals for adaptive Lasso estimates are also unreliable (see Pötscher and Schneider (2009)).

As the influence of the level of pollution on the coefficients of the other covariates seems rather marginal, and for other reasons, we consider sex as an additional (potentially) effect modifying factor in the next section.

6. Generalizations to Multiple Effect Modifiers

Because in many applications there is not only one potential effect modifying factor, we show how a model with multiple categorial effect modifiers can be specified and regularized.

Suppose two predictors x_1 and x_2 are given, with potential effect modifiers $u_1 \in \{1, \dots, k_1\}$ and $u_2 \in \{1, \dots, k_2\}$. We assume

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

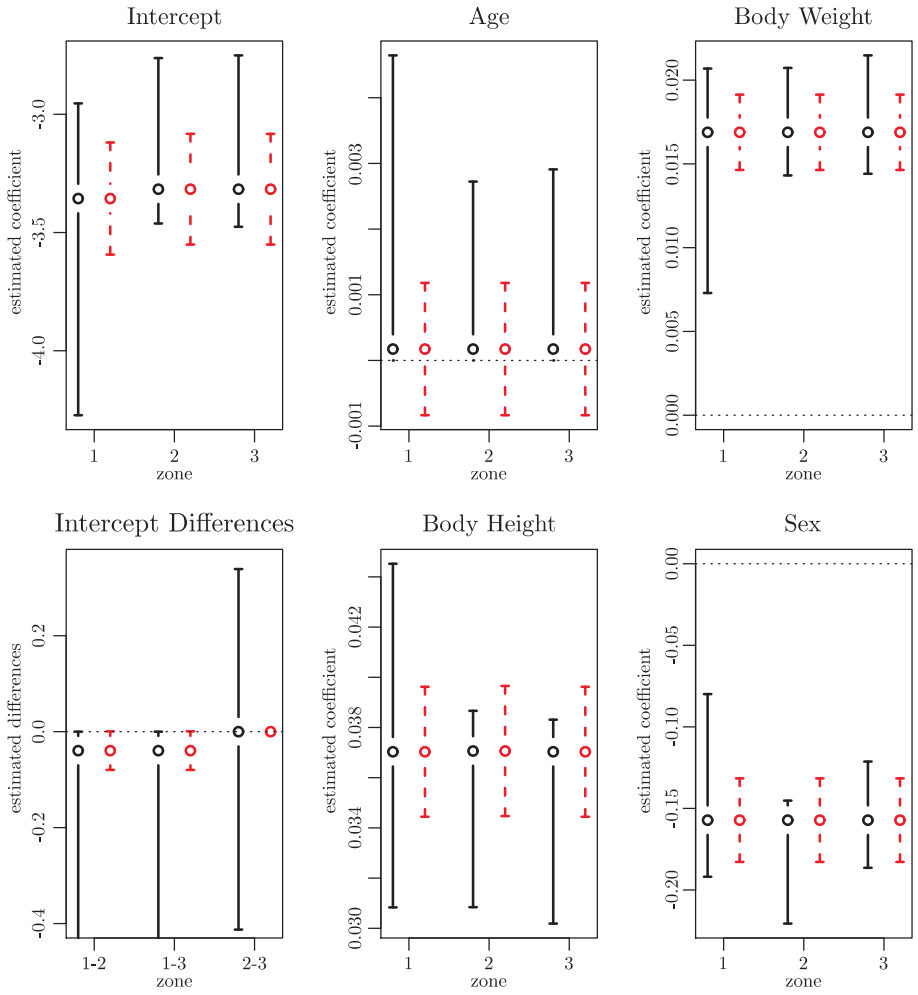

Figure 6. 90% bootstrap percentile confidence intervals (solid) and confidence intervals using the asymptotic normal distribution from Proposition 2 (dashed).

Thus, for the varying functions $\beta_j(u_1, u_2)$ an additive structure is imposed:

$$\beta_j(u_1, u_2) = \beta_{j1}(u_1) + \beta_{j2}(u_2), \tag{6.1}$$

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 identifiability, functions $\beta_{j2}(\cdot)$ need to be restricted, for example by $\beta_{j21} =$

Table 5. Fitted coefficients as defined in (6.1) 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 6.

		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 6. 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.

	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

$0,\ j = 0, \ldots, p$. The penalty term is

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

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, since $\beta_{j21} = 0$ for all j. For the same reason, terms $|\beta_{j2r}|$ are not explicitly penalized since they are implicitly included in the first part of the penalty.

To illustrate the approach, we use the data from Table 3 but only consider covariates that showed relevant effects in Table 4. In Table 5 fitted coefficients are given if the degree of pollution at the place of residence and the child's sex are taken as (potentially) effect modifying factors. The value $s/s_{\max} = 0.56$ was chosen via (5-fold) cross-validation. 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 and differences between other coefficients are the same for both males and females.

In general, the influence of 'sex' on the coefficients of the remaining covariates seems to be higher than that for 'zone'. For a better understanding, $\hat{\beta}_{j1}(u_1)$ and $\hat{\beta}_{j2}(u_2)$ are given in Table 6 where u_1 is 'zone' and u_2 is 'sex'. Since $\beta_{j21} = 0\ \forall j$, for each covariate there is only an u_2-coefficient given for females – the difference between females and males (implicitly) already seen in Table 5. Since coefficients given in the last row of Table 6 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 5): the (absolute) difference in lung capacity between boys and girls increases with age. 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. The ratio of residual and total sum of squares is 10.8%, and the fit of our first model is improved.

7. Summary and Discussion

We showed how regularization can be used to obtain sparser representations of varying-coefficient models with categorial effect modifiers. The proposed regularization technique can lead to stabilization and higher accuracy of estimates, and interpretability of the fitted models is increased. When weights of penalty terms are included and adaptively chosen, selection consistency is obtained, as already shown by Zou (2006) for the Lasso. Also for finite sample sizes, the adaptive version has the potential to outperform the non-adaptive version in both estimation/prediction accuracy and model selection, as shown in simulation studies.

The analysis of data sets showed that the proposed method can be successfully applied in practice. And, indeed, can be applied when there are multiple (categorial) effect modifiers. As for all selection procedures, however, it is difficult to give unbiased measures of uncertainty. We suggest bootstrap confidence intervals as gauges of variability in most situations.

When regression coefficients are allowed to vary with time, time-dependent effects of covariates are commonly captured by fitting them by smooth functions. In some cases, however, observation points are fixed, for example if a certain quantity is measured every day at the same time. 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 estimates, 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 coefficients are identified. So it might be said, for example, that a relevant change occurs between the second and the third day. The difference here is the need to specify a correlation structure for the errors, autocorrelation for instance.

Acknowledgements

This work was partially supported by DFG project TU62/4-1. We thank an associate editor and two anonymous referees for helpful comments and suggestions.

Appendix

Derivation of (1.1). We take covariates $x_1, \ldots, x_p$ and a nominally scaled (possibly) effect modifying factor u with levels $1, \ldots, k$. Then for each $j = 1, \ldots, p$, there are $\binom{k}{s}$ possibilities in selecting s nonzero coefficients from $\{\beta_{j1}, \ldots, \beta_{jk}\}$, with $s = 0, \ldots, k$. Among the selected s coefficients there are $C(s)$ possible partitions with constant coefficients within each cluster. If intercepts are never set to zero, but can be fused one obtains the number of potential models as

$$M(p,k) = C(k)\Big(1 + \sum_{s=1}^{k} \binom{k}{s} C(s)\Big)^p,$$

where

$$C(s) = \sum_{v=1}^{s} S(s,v).$$

Here $S(s,v)$ is the number of possible assignments of s objects to v clusters, given by Jain and Dubes (1988),

$$S(s,v) = \frac{1}{v!} \sum_{l=1}^{v} (-1)^{v-l} \binom{v}{l} l^s.$$

Proof of Proposition 1. If $\hat{\beta}$ minimizes $Q_p(\beta)$ at (2.2), 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 $\hat{\beta}^{(LS)}$, ensured by $n_r/n \to c_r$, with $0 < c_r < 1 \; \forall r$.

Proof of Proposition 2. We first show asymptotic normality, following Zou (2006), Bondell and Reich (2009), and Gertheiss and Tutz (2010). At (2.2), let $\beta = \beta^* + b/\sqrt{n}$, where β^* denotes the true coefficient vector. Then we 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

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

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

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

with

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

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

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

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

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

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

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 $\lambda_n/\sqrt{n} \to 0$, we have (by Slutsky's theorem)

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

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

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

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

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

Moreover, if $\beta^*_{jr}=0$ or $\beta^*_{jr}=\beta^*_{js}$, due to $\sqrt{n}$-consistency of the ordinary least squares estimate,

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

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 $\beta^*_{jr}=\beta^*_{js}$ $(j\geq 0)$ 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 (a) is similar to that of 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 $\forall r\ n_r/n\to c_r$ $(0<c_r<1)$, $n^{-1}Z^{*T}Z^*\to C>0$ and $n^{-1/2}\epsilon^TZ^*\to_d w$, with $w\sim N(0,\sigma^2C)$. Let $\theta_{\mathcal{C}^c}$ denote the vector of θ-entries which are truly zero, and $b_{\mathcal{C}^c}$ be the subset of entries of $\theta_{\mathcal{C}^c}$ which are part of b; analogously, $b_{\mathcal{C}}$ denotes the subset of $\theta_{\mathcal{C}}$-entries which are part of b. As 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 reordering of entries), we have (cf., Zou (2006); Bondell and Reich (2009)) $\hat{b}_{\mathcal{C}}\to_d C^{-1}w$, $\hat{b}_{\mathcal{C}^c}\to_d 0$, and $\hat{b}_{\mathcal{C}}\to_d N(0,\sigma^2C^{-1})$. Writing $\tilde{\beta}=(\tilde{\beta}_0^T,\ldots,\tilde{\beta}_p^T)^T$ with $\tilde{\beta}_j=(\beta_{jr}-\beta_{j1},\ldots,\beta_{jr},\ldots,\beta_{jr}-\beta_{jk})^T$, asymptotic normality can be proven for all entries of $\hat{\theta}_{\mathcal{C}}$.

To show consistency, we first note that $\lim_{n\to\infty}P(\mathfrak{I}\in\mathcal{C}_n)=1$, if $\mathfrak{I}\in\mathcal{C}$, follows from (a), where $\mathfrak{I}$ denotes a triple of indices (j,r,s) or a pair (j,r). We now show that if $\mathfrak{I}\notin\mathcal{C}$, $\lim_{n\to\infty}P(\mathfrak{I}\in\mathcal{C}_n)=0$. A similar idea is found in Bondell and Reich (2009). Let $\mathcal{B}_n$ denote the (nonempty) set of indices $\mathfrak{I}$ that are in $\mathcal{C}_n$ but not in $\mathcal{C}$. Without 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 reparameterize β_l by $\tilde{\beta}_l$ as above, since

all coefficients and differences thereof are penalized in the same way. If $l = 0$, the reparametrization means choosing a reference category whose intercept is not penalized. In this case the proof is analogous (with small modifications) to that of Bondell and Reich (2009).

If we order categories so that $\hat{\beta}_{l1} \leq \ldots \leq \hat{\beta}_{lz} \leq 0 \leq \hat{\beta}_{l,z+1} \leq \ldots \leq \hat{\beta}_{lk}$, estimate $\hat{\beta}$ at (2.2) with penalty (3.1) is equivalent to

$$\hat{\beta} = \underset{\mathfrak{B}}{\operatorname{argmin}} \Big\{ (y - Z\beta)^T (y - Z\beta) + \lambda_n \sum_j J_j(\beta) \Big\}, \text{ with}$$

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

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

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

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

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

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

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

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

From (a) and applying Slutsky's theorem, $2 z_{(l)q}^T Z \sqrt{n} (\beta - \hat{\beta})/n$ is asymptotically normal with mean zero; $2 z_{(l)q}^T \epsilon / \sqrt{n}$ is 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, $\lim_{n\to\infty} P(\sqrt{n}|\hat{\beta}_{lq}^{(LS)}| \leq \lambda_n^{1/2}) = 1$, if $(l,q) \in \mathcal{B}_n$. Hence $\lim_{n\to\infty} P(D_n > \lambda_n^{1/4}) = 1$. As a consequence $\lim_{n\to\infty} P(Q_q'(\hat{\beta})/\sqrt{n} = A_n + D_n) = 0$, so if $\mathfrak{I} \notin \mathcal{C}$, $\lim_{n\to\infty} P(\mathfrak{I} \in \mathcal{C}_n) = 0$.

References

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

Canty, A. and Ripley, B. (2010). *boot: Bootstrap R (S-Plus) Functions*. R package version 1.2-42.

Cardot, H. and Sarda, B. (2008). Varying-coefficient functional linear regression models. *Comm. Statist. Theory Methods* **37**, 3186-3203.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32**, 407-499.

Fan, J., Yao, Q. and Cai, Z. (2003). Adaptive varying-coefficient linear models. *J. Roy. Statist. Soc. Ser. B* **65**, 57-80.

Gertheiss, J. and Tutz, G. (2010). Sparse modeling of categorial explanatory variables. *Ann. Appl. Statist.* **4**, 2150-2180.

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

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* **55**, 757-796.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning* (2nd ed.). Springer, New York.

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

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

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, NJ.

Karatzoglou, A., Smola, A., Hornik, K. and Zeileis, A. (2004). kernlab – an S4 package for kernel methods in R. *J. Statist. Software* **11**, 1-20.

Kauermann, G. and Tutz, G. (2000). Local likelihood estimation in varying coefficient models including additive bias correction. *J. Nonparametr. Stat.* **12**, 343-371.

Kim, M.-O. (2007). Quantile regression with varying coefficients. *Ann. Statist.* **35**, 92-108.

Leng, C. (2009). A simple approach for varying-coefficient model selection. *J. Statist. Plann. Inference* **139**, 2138-2146.

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

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

Pötscher, B. M. and Schneider, U. (2009). On the distribution of the adaptive lasso estimator. *J. Statist. Plann. Inference* **139**, 2775-2790.

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

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

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.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. 4th edition. Springer, New York.

Wang, H. and Xia, Y. (2009). Shrinkage estimation of the varying coefficient model. *J. Amer. Statist. Assoc.* **104**, 747-757.

Wang, L., Li, H. and Huang, J. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *J. Amer. Statist. Assoc.* **103**, 1556-1568.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B* **68**, 49-67.

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

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B* **67**, 301-320.

Department of Statistics, Ludwig-Maximilians-Universität Munich, Germany.

E-mail: jan.gertheiss@stat.uni-muenchen.de

Department of Statistics, Ludwig-Maximilians-Universität Munich, Germany.

E-mail: gerhard.tutz@stat.uni-muenchen.de

(Received August 2010; accepted October 2011)