LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Margret-Ruth Oelker, Jan Gertheiss, Gerhard Tutz

Regularization and Model Selection with Categorial Predictors and Effect Modifiers in Generalized Linear Models

Technical Report Number 122, 2012
Department of Statistics
University of Munich

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

REGULARIZATION AND MODEL SELECTION WITH CATEGORICAL PREDICTORS AND EFFECT MODIFIERS IN GENERALIZED LINEAR MODELS

By Margret-Ruth Oelker[*,†], Jan Gertheiss[†] and Gerhard Tutz[†]

Ludwig-Maximilians-Universität Munich[†]

Varying-coefficient models with categorical effect modifiers are considered within the framework of generalized linear models. We distinguish between nominal and ordinal effect modifiers, and propose adequate Lasso-type regularization techniques that allow for (1) selection of relevant covariates, and (2) identification of coefficient functions that are actually varying with the level of a potentially effect modifying factor. We investigate large sample properties, and show in simulation studies that the proposed approaches perform very well for finite samples, too. In addition, the presented methods are compared with alternative procedures, and applied to real-world medical data.

1. Introduction. In regression modeling, categorical predictors, also called factors, are a standard case. Nevertheless, variable selection for discrete covariates and the connected problem which categories within one factor are to be distinguished has been somewhat neglected. More concrete, in our application, we model the effects of pregnancy related covariates on the type of delivery, that is, if birth was given vaginally or by means of a Cesarean. Cases were observed over a period of several years. As medical standards typically change over time, modeling the type of delivery requires to consider discrete time-effects, and more importantly, to consider how effects change over years. In general, we are going to address model selection with discrete covariates in a slightly extended version of generalized linear models (GLMs), namely GLMs with varying coefficients.

Varying-coefficient models (Hastie and Tibshirani, 1993) are a quite flexible tool to capture complex model structures and interactions. In the setting of GLMs, regression coefficients β_j are allowed to vary with the value of other variables u_j. Hence the linear predictor has the form

$$\eta = \beta_0(u_0) + x_1\beta_1(u_1) + \ldots + x_p\beta_p(u_p), \tag{1.1}$$

[*] This work was partially supported by DFG project “Regularisierung für diskrete Datenstrukturen”.

Keywords and phrases: Categorical predictors, generalized linear model, varying-coefficients, fused Lasso, variable selection

where $x_1, x_2, \ldots, x_p$ are continuous covariates, and $u_1, \ldots, u_p$ are the so called effect modifiers, which modify the effects of the covariates in an unspecified, typically smooth form $\beta_j(\cdot)$. Thus, the predictor is still linear in the regressors $x_1, \ldots, x_p$, but scalar coefficients β_j turn into functions depending on the effect modifiers u_j, $j = 0, \ldots, p$. As common in GLMs, it is assumed that the predictor η is linked to the conditional mean of response y by a known response function h, that is,$\mu = \mathbb{E}(y|x_1, \ldots, x_p) = h(\eta)$, and y follows a simple exponential family. Throughout the paper we assume that covariates $x_1, \ldots, x_p$ are measured on comparable scales or have been scaled. For continuous effect modifiers, unknown functions $\beta_j(\cdot)$ are typically smooth and have been modeled by splines (Hastie and Tibshirani, 1993; Hoover et al., 1998;Lu, Zhang and Zhu, 2008), using localizing techniques (Wu, Chiang and Hoover, 1998; Fan and Zhang, 1999; Kauermann and Tutz, 2000) or boosting (Hofner, Hothorn and Kneib, 2008). Inference requires to distinguish between varying and non-varying coefficients and between relevant and non-relevant terms. Hastie and Tibshirani (1993) proposed to adopt techniques for additive models. Leng (2009) distinguishes between varying and non-varying coefficients by applying the Cosso (Lin and Zhang, 2006) penalty, while Wang, Li and Huang (2008) obtain selection of spline coefficients by groupwise SCAD-penalization. Wang and Xia (2009) select covariates by local polynomial regression with the grouped Lasso (Yuan and Lin, 2006). However, apart from Hofner, Hothorn and Kneib (2008), selection of predictors and identification of smooth/constant functions is not reached simultaneously.

In contrast to most existing approaches, we consider categorical effect modifiers $u_j \in \{1, \ldots, k_j\}$. In the pregnancy example, for instance, the effect modifier indicates the year considered. Then functions $\beta_j(u_j)$ have the form $\sum_{r=1}^{k_j} \beta_{jr} I(u_j = r)$, where $I(\cdot)$ denotes the indicator function and $\beta_{j1}, \ldots, \beta_{jk_j}$ represent regression parameters. Therefore the linear predictor is given by

$$\eta = \sum_{r=1}^{k_0} \beta_{0r} I(u_0 = r) + \sum_{j=1}^{p} x_j \sum_{r=1}^{k_j} \beta_{jr} I(u_j = r).$$

The total coefficient vector is given by $\beta^T = (\beta_0^T, \ldots, \beta_p^T)$, where sub-vector $\beta_j^T = (\beta_{j1}, \ldots, \beta_{jk_j})$ contains the parameters for the jth predictor. With categorical effect modifiers, the number of parameters $q = \sum_{j=0}^{p} k_j$ can become very large, even for a moderate number of predictors p. Consequently, usual maximum likelihood (ML) estimates may not exist and alternative tools such as regularization techniques are needed. And even if ML estimates exist, it is desirable to reduce the model to the relevant terms. That means, one wants to determine which predictors are influential, and if so, which categories have to be distinguished.

The methods proposed here extend the work of Gertheiss and Tutz (2012), as the latter is restricted to the classical linear model and hence cannot be used for analyzing non-normal response variables such as the Cesarean data described above. Two approaches are presented that allow to model categorical effect modifiers within the GLM framework. In Section 2 we propose a penalized ML criterion. For computation, a penalized iteratively reweighted least squares algorithm is employed. Moreover, large sample properties are derived. As an alternative, a forward selection procedure using information criteria is shortly sketched (Section 3). The proposed methods are shown to be highly competitive in numerical experiments (Section 4). In Sections 5 and 6, the approaches are applied to the Caesarean data and to data on the reduction of mortality after myocardial infarction; the special case of categorical effects is discussed in Section 7.

2. Penalized Estimation.

Our main tool for regularization and model selection is the use of penalties. In GLMs, penalized estimation means to minimize

$$\mathcal{M}_n^{pen}(\beta) = -l_n(\beta) + P_\lambda(\beta) = -l_n(\beta) + \lambda \cdot J_n(\beta), \tag{2.1}$$

where $l_n(\beta)$ denotes the log-likelihood for sample size n, and $P_\lambda(\beta)$ stands for a general penalty depending on tuning parameter λ. The expression $\lambda \cdot J_n(\beta)$ breaks the penalty down to a product, underlining the dependency on one scalar tuning parameter only. With $\lambda = 0$, ordinary ML-estimation is obtained.

The main issue is to choose an adequate penalty $J_n(\beta)$: The Ridge penalty (Hoerl and Kennard, 1970), for instance, shrinks coefficients, while the Lasso (Tibshirani, 1996) combines shrinkage and selection of coefficients, and the fused Lasso (Tibshirani et al., 2005) applies the Lasso to differences of adjacent parameters. Thus, parameters are shrunk towards each other and potentially fused in order to gain a local consistent profile of ordered coefficients. In contrast, the grouped Lasso (Yuan and Lin, 2006) selects whole groups of coefficients simultaneously. Although variable selection is implied, both the Lasso and its grouped version are off target since they do not enforce $\beta_{jr} = \beta_{js}$ for some $r \neq s$. The pure fused Lasso indeed leads to (piecewise) constant functions $\beta_j(u_j)$ but disregards the selection of whole predictors. A combination of both allows not only for shrinkage and selection but also for gradual fusion of related coefficients – such that effects of

the grouped Lasso are embedded.
As nominal and ordinal effect modifiers in (1.1) contain different information, they should be treated differently. Therefore, we consider the general penalty

$$J_n(\beta) = \sum_{j=0}^{p} J_j(\beta_j), \tag{2.2}$$

where $J_j(\beta_j) = 0$ if covariate j is not modified, $J_j(\beta_j) = J_j^{nom}(\beta_j)$ for nominal effect modifiers and $J_j(\beta_j) = J_j^{ord}(\beta_j)$ for ordinal effect modifiers.

For a *nominal* effect modifier u_j we propose

$$J_j^{nom}(\beta_j) = \sum_{r>s} |\beta_{jr} - \beta_{js}| + b_j \sum_{r=1}^{k_j} |\beta_{jr}|, \tag{2.3}$$

where b_j is an indicator that (de-)activates the second sum if wanted. Penalty (2.3) is equivalent to a fused Lasso penalty applied on all pairwise differences of coefficients belonging to $\beta_j(u_j)$. Thus, not only adjacent coefficients but each subset of nominal categories can be collapsed. In the case of strong penalization, effects $\beta_{j1}, \ldots, \beta_{jk_j}$ of covariate j are reduced to one constant coefficient and do not depend on the categories of u_j anymore; one obtains $\hat{\beta}_{j1} = \ldots = \hat{\beta}_{jk_j} = \hat{\beta}_j$. The second sum in (2.3) conforms to a Lasso penalty shrinking all coefficients belonging to $\beta_j(u_j)$ individually toward zero. The effect is selection and exclusion of covariates. For strong penalization $\hat{\beta}_{j1} = \ldots = \hat{\beta}_{jk_j} = 0$ is obtained, and covariate j is excluded. In most cases, a constant intercept shall remain in the model; hence, we typically have $b_0 = 0$.

If u_j is *ordinal*, this additional information should be used. Our proposal is to allow for the fusion of adjacent categories β_{jr} and $\beta_{j,r-1}$. Hence, for ordinal predictors we use

$$J_j^{ord}(\beta_j) = \sum_{r=2}^{k_j} |\beta_{jr} - \beta_{j,r-1}| + b_j \sum_{r=1}^{k_j} |\beta_{jr}|, \tag{2.4}$$

where b_j denotes the same indicator as above. Instead of all pairwise differences now only differences of neighbored coefficients are penalized, which corresponds exactly to a fused Lasso-type penalty (Tibshirani et al., 2005). Again, with setting $b_0 = 0$, the intercept can be treated separately.

Apart from their different amount of information, J_j^{nom} and J_j^{ord} work similarly: one term leads to fusion within the predictor, while a Lasso-type penalty selects coefficients. Thus, overall variable selection as well as distinction of varying and non-varying coefficients is obtained.

If, for example, emphasis should be put on the selection of covariates, it may be advantageous to use weights for the two components of the penalty (compare Tibshirani et al., 2005). With parameter $\psi \in (0,1)$, the weighted penalty for nominal effect modifier j is

$$J_j^{nom}(\beta,\psi) = \psi \sum_{r>s} |\beta_{jr} - \beta_{js}| + (1-\psi) b_j \sum_{r=1}^{k_j} |\beta_{jr}|, \tag{2.5}$$

for ordinal effect modifiers, it is

$$J_j^{ord}(\beta,\psi) = \psi \sum_{r=2}^{k_j} |\beta_{jr} - \beta_{j,r-1}| + (1-\psi) b_j \sum_{r=1}^{k_j} |\beta_{jr}|. \tag{2.6}$$

Parameter ψ is restricted to $(0,1)$ in order to separate it strictly from tuning parameter λ. It allows to place emphasis on the fusion or on the selection part of the penalty, but even so, it is another tuning parameter that has to be chosen.

If effect modifiers u_j have different numbers of categories, additional weighting of penalty terms analogously to Bondell and Reich (2009) could be used to prevent eventual selection bias.

2.1. *Computational Issues.* Since penalty (2.2) contains absolute values, a convex but not continuously differentiable optimization problem has to be solved. However, non-differentiability can be evaded by approximating the penalty at the critical points, i.e., in a neighborhood of $|\xi|$, $\xi = 0$. We employ a slightly adjusted version of the algorithm proposed by Fan and Li (2001) and described in detail by Ulbricht (2010). In general, we assume a penalty that can be written as $P_\lambda(\beta) = \sum_{l=1}^{L} p_{\lambda,l}(|a_l^T\beta|)$, where a_l are known constants. Penalty terms $p_{\lambda,l}(|a_l^T\beta|)$ are supposed to map $|a_l^T\beta|$ onto the positive real numbers, to be continuous and monotone in $|a_l^T\beta|$. In addition, penalty terms $p_{\lambda,l}(|a_l^T\beta|)$ are assumed to be continuously differentiable $\forall\ a_l^T\beta \neq 0$ such that $\mathrm{d}p_{\lambda,l}(|a_l^T\beta|)/\mathrm{d}|a_l^T\beta| \geq 0\ \forall\ a_l^T\beta > 0$ holds. Approximating the absolute values by $|\xi| \approx \sqrt{\xi^2 + c}$, where c is a small positive constant, allows for derivatives of the objective function. Thus, the Fisher scoring algorithm, which is typically used for ordinary GLMs, can be modified to a version that handles the approximated penalty.

Also penalty $J_n(\beta)$ from equation (2.2) can be rewritten this way. Let the vectors a_l denote the columns of a block-diagonal matrix $A = \text{diag}(A_0, \dots, A_p) \subset \mathbb{R}^{q \times L}$ and functions $p_{\lambda,l}(\nu)$ be defined as $\lambda \cdot \nu$. Let block A_j refer to the effect modifier u_j. If u_j is nominal, $A_j^T \beta_j$ shall give the values of the according coefficients $\beta_{j1}, \dots, \beta_{jk_j}$ and their pairwise differences. The former is reached when using the columns of a $(k_j \times k_j)$ identity matrix, the latter by columns containing these combinations of ±1 building the needed differences. Hence, e.g. for $k_j = 4$, we have

$$A_j^{nom} = \begin{pmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 0 & -1 & -1 & 0 \\ 0 & 1 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 0 & 1 & -1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix},$$

which is a $k_j \times (\frac{1}{2}k_j(1+k_j))$ dimensional matrix. If u_j is ordinal, only pairwise differences of coefficients $\beta_{j1}, \dots, \beta_{jk_j}$ are penalized. Thus in $(k_j \times (2k_j - 1))$ matrix A_j^{ord} the last three columns of matrix A_j^{nom} are omitted. If the intercept is modified by any effect modifier, matrix A_0 depends on the concrete form of the penalty. In general, if $b_j = 0$ the "diagonal part" part of A_j^{nom}, A_j^{ord} respectively, is omitted. For a covariate x_j, whose influence on y is not modified by any u_j, matrix A_j^{none} is an empty matrix with zero columns and k_j rows.

The generalized hat matrix of the algorithm's final iteration allows to estimate the model's degrees of freedom. But the LQA-algorithm is only locally convergent. Only if the objective function is strictly convex, a local optimum is ensured to be the global optimum, too. Strict convexity implies that the penalized Fisher information matrix is positive definite. Nevertheless, the penalty applied here leads to a positive semi-definite information matrix. Therefore the quasi-Newton approach will find descent directions in each iteration but it may happen that the solution is not unique (Ulbricht, 2010).

2.2. *Large Sample Properties.* For asymptotics, general assumptions have to hold and the number of observations has to grow in accordance with the requirements of categorical covariates: If sample size n tends to infinity it is assumed that the number of observations n_{jr} on level r of u_j tends to infinity for all j, r at the same rate. Then we have

Theorem 2.1. *Suppose $0 \leq \lambda < \infty$ has been fixed, and all class-wise sample sizes n_r satisfy $n_{jr}/n \to c_{jr}$, where $0 < c_{jr} < 1$. Then the estimate $\hat{\beta}$ that minimizes (2.1) with $J_n(\beta)$ defined by (2.2), (2.3) and (2.4) is consistent, i.e. $lim_{n \to \infty} \mathbb{P}(||\hat{\beta} - \beta^*||^2 > \epsilon) = 0$ for all $\epsilon > 0$.*

The proof is given in Appendix A. Employing the generalized versions (2.5) and (2.6) does not affect the consistency results.

As pointed out in Zou (2006), regularization as used so far does not ensure consistency in terms of variable selection. In order to gain selection consistency of the original Lasso, Zou (2006) proposed an adaptive version that has the so-called oracle properties. A corresponding modification for penalty (2.2) is available: Given effect modifiers u_j, $j = 1, \ldots, p$, penalty $J_n(\beta)$ (2.2) is modified to the adaptive penalty $J_n^{ad}(\beta)$ by employing

$$J_j^{ad,nom}(\beta) = \sum_{r>s} w_{rs(j)} |\beta_{jr} - \beta_{js}| + b_j \sum_{r=1}^{k_j} w_{r(j)} |\beta_{jr}| \text{ and} \tag{2.7}$$

$$J_j^{ad,ord}(\beta) = \sum_{r=2}^{k_j} w_{r,r-1(j)} |\beta_{jr} - \beta_{j,r-1}| + b_j \sum_{r=1}^{k_j} w_{r(j)} |\beta_{jr}|, \tag{2.8}$$

which replace (2.3) and (2.4), and by using adaptive weights

$$w_{rs(j)} = \phi_{rs(j)}(n) |\hat{\beta}_{jr}^{ML} - \hat{\beta}_{js}^{ML}|^{-1} \text{ and} \tag{2.9}$$

$$w_{r(j)} = \phi_{r(j)}(n) |\hat{\beta}_{jr}^{ML}|^{-1}. \tag{2.10}$$

Here $\hat{\beta}_{jr}^{ML}$ denotes the ML-estimate of β_{jr}. For functions $\phi_{rs(j)}(n)$ and $\phi_{r(j)}(n)$, convergence to fixed values is assumed; that is, $\phi_{rs(j)}(n) \rightarrow q_{rs(j)}$ and $\phi_{r(j)}(n) \rightarrow q_{r(j)}$, with $0 < q_{rs(j)}, q_{r(j)} < \infty$. If $\phi_{rs(j)}(n)$ and $\phi_{r(j)}(n)$ are positive constants, that sum up to one, we obtain a generalization as given in equations (2.5) and (2.6); tuning parameter λ and functions $\phi_{rs(j)}(n)$, $\phi_{r(j)}(n)$ are clearly separated.

To ensure consistency, penalty parameter λ has to increase with sample size n; one assumes that $\lambda = \lambda_n$ with $\lambda_n/\sqrt{n} \rightarrow 0$ and $\lambda_n \rightarrow \infty$, and all class-wise sample sizes n_r satisfy $n_r/n \rightarrow c_r$, where $0 < c_r < 1$.

In addition, we define vector $\theta = A^T\beta$. Hence, θ is a vector that contains all terms that penalty $J_n(\beta)$ (2.2) considers. That is, the absolute values of all penalized coefficients β_{ij} and – according to the level of measurement – the absolute values of their differences. $\hat{\theta}^n$ denotes the estimate of θ based on sample size n. Furthermore, there are some sets to be defined: $\mathcal{C}$ denotes the set of indices corresponding to those entries of θ which are truly non-zero. $\mathcal{C}_n$ is the set corresponding to those entries of $\hat{\theta}^n$ which are estimated to be non-zero with sample size n, and based on estimate $\hat{\beta}^n$. $\theta_{\mathcal{C}}^*$ denotes the vector of θ-entries which are truly included in $\mathcal{C}$, $\hat{\theta}_{\mathcal{C}}^n$ is the corresponding estimate.

Previous assumptions concerning ML-estimation are extended: the model must hold, the negative log-likelihood $-l_n(\beta)$ has to be convex. $l_n(\beta)$ has to be at least three times continuously differentiable, the third moments of y have to be finite. The information matrix F_n/n must have a positive definite limit; for score function $s(\beta)$, we suppose $\mathbb{E}(s(\beta)) = 0$. Then one obtains

Theorem 2.2. *Suppose $\lambda = \lambda_n$ with $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n \to \infty$, and all class-wise sample sizes n_{jr} satisfy $n_{jr}/n \to c_{jr}$, where $0 < c_{jr} < 1$. Then penalty $J_n^{ad}(\beta)$ employing terms (2.7) and (2.8) with weights (2.9) and (2.10), where $\hat{\beta}_{jr}^{ML}$, $\phi_{rs(j)}(n)$ and $\phi_{r(j)}(n)$ are defined as above, ensures that*

(a) $\sqrt{n}(\hat{\theta}_{\mathcal{C}}^n - \theta_{\mathcal{C}}^*) \xrightarrow{d} N(0, Cov(\theta_{\mathcal{C}}^*))$
(b) $lim_{n\to\infty}\mathbb{P}(\mathcal{C}_n = \mathcal{C}) = 1$

The proof uses ideas from Zou (2006) and Bondell and Reich (2009), and is given in Appendix B. The concrete form of $\text{Cov}(\theta_{\mathcal{C}}^*)$ results from the asymptotic marginal distribution of a set of non-redundant truly non-zero regression parameters or differences threof. Since all estimated differences are (deterministic) linear functions of estimated parameters, the covariance-matrix $\text{Cov}(\theta_{\mathcal{C}}^*)$ is singular.
$F_n/n \overset{n\to\infty}{\to} F$ with positive definite F is typically assumed in observational studies but it raises problems in experiments. In this case the given proof can be extended to matrix normalization (see for example Fahrmeir and Kaufmann, 1985). For $\lambda = 0$, the unpenalized likelihood is maximized; therefore asymptotic normality and consistency hold as shown by McCullagh (1983). Distributional properties for $n \to \infty$ given a fixed λ are not discussed since the penalty shall not vanish in proportion to $-l_n(\beta)$ for $n \to \infty$. For the normality part of Proposition 2.2, the speed of convergence is $\lambda_n/\sqrt{n} \to 0$. Since $n^{-1/2}s_n(\beta) \sim N(0, F(\beta)) + \mathcal{O}(n^{-1/2})$ and $\mathbb{P}(\sqrt{n}|\hat{\beta}_{lq}^{ML}| \le \lambda_n^{1/2}) \to 1$ like $c/\sqrt{n} \to 0$, the consistency part behaves the same. Thus, the overall speed of convergence is $\mathcal{O}(n^{-1/2})$. Since the penalized model from Proposition 2.2 converges to an ordinary GLM for $n \to \infty$, and since the scale parameter of the exponential family φ and β are orthogonal (see the mixed second derivatives $\frac{\partial l}{\partial\varphi\partial\beta}$ given in Claeskens and Hjort, 2008) it is possible to replace φ by $\hat{\varphi}$. Hence, all used arguments are valid for quasi likelihood models, too. Only the estimates' covariance matrix cannot be reduced to $F(\beta)^{-1}$ anymore but remains $F(\beta)^{-1}V(\beta)F(\beta)^{-1}$, where $V(\beta) = \text{cov}(s(\beta))$ and $F(\beta) = \mathbb{E}\left(-\frac{\partial^2 l_n(\beta)}{\partial\beta\partial\beta^T}\right)$, see McCullagh (1983) for details.
In some cases, in particular for small sample sizes, ML-estimates required for adaptive weighting may not exist. If necessary, ML-estimates can be replaced

by other $\sqrt{n}$-consistent estimates, e.g. Ridge estimates with fixed tuning parameter. However, adaptive estimation is as good as the used weights and hence not recommended by all means.

3. Alternative Selection Strategies.

For the selection of variables, stepwise procedures are typically used. In particular, forward and backward selection methods based on information criteria like the AIC or the BIC are popular. One tries to find the model that performs best with respect to the criterion. By construction, these strategies yield variable selection but no fusion of categories.

Gertheiss and Tutz (2012) obtained fusion of categories by using an enlarged setting. For a nominal effect modifier u_j with three categories having impact on covariate x_j, for example, the varying coefficient $\beta_j(u_j)$ corresponds to $(\beta_{j1}, \beta_{j2}, \beta_{j3})^T$ in coefficient vector β. All possible combination of coefficients belonging to x_j would be: $\{()$, (β_{j1}), (β_{j2}), (β_{j3}), (β_{j1}, β_{j2}), (β_{j1}, β_{j3}), (β_{j2}, β_{j3}), $(\beta_{j1}, \beta_{j2}, \beta_{j3})\}$. Allowing for fusion increases the number of possibilities by $\{(\beta_{j1}, \beta_{j2} = \beta_{j3})$, $(\beta_{j2}, \beta_{j1} = \beta_{j3})$, $(\beta_{j3}, \beta_{j2} = \beta_{j1})$, $(\beta_{j1} = \beta_{j2} = \beta_{j3})\}$. When selecting a model, all possibilities to fuse coefficients must be considered.

Concretely, we start with a model containing an intercept only. In each step, the degrees of freedom of the model are enlarged by one until the chosen criteria (AIC or BIC) is not improved anymore; with the degrees of freedom being defined as the number of non-zero coefficient blocks in $\hat{\beta}$ (Tibshirani et al., 2005). Hence in each step a former zero coefficient can be set to non-zero, or a former zero group of coefficients can become non-zero. Alternatively a group of equal coefficients can be split into two groups of non-zero but identical coefficients.

4. Numerical Experiments.

The proposed methods are compared in simulation studies. For illustration, we start with a simple example.

4.1. *An illustrative example.*

We assume a logistic regression model with two covariates x_1, x_2 and one nominal effect modifier u with categories 1, 2 and 3. u possibly impacts all covariates plus the intercept. Concretely, the predictor is

$$\begin{aligned} \eta_{true} &= \beta_0(u) + x_1\beta_1(u) + x_2\beta_2(u) \\ &= \beta_0 + x_1\left(\ \beta_{11}I(u=1) + \beta_{12}I(u=2) + \beta_{13}I(u=3)\ \right) + x_2\beta_2 \qquad (4.1) \\ &= 0.2 + x_1\left(\ 0.3I(u=1) + 0.7I(u=2) + 0.7I(u=3)\ \right) - x_2 \cdot 0.5 \end{aligned}$$

That means, while the intercept and x_2 do not depend on u, covariate x_1 varies with categories 1 and 2/3 of u. Covariates x_1 and x_2 are independently

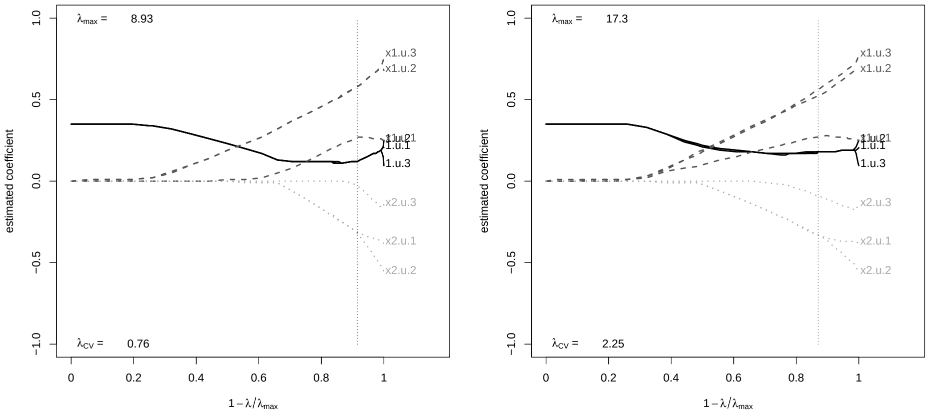

FIG 1. *Coefficient paths for binary model (4.1) assuming predictor (4.2) – with adaptive weights (left) and the standard penalty (right).*

drawn from an uniform distribution $U(0,2)$; the effect modifier u is multinomial with probabilities 0.3, 0.4, 0.3 for categories 1, 2 and 3, respectively. For response y, $y = h(\eta)$ holds, where $h^{-1}(\cdot)$ is the natural link (logit) function. We generate $n = 400$ observations. When fitting the model, all coefficients are allowed to vary with effect modifier u, i.e., we have

$$\eta_{model} = \beta_0(u) + x_1 \cdot \beta_1(u) + x_2 \cdot \beta_2(u). \tag{4.2}$$

Figure 1 shows the resulting coefficient paths for the proposed estimator subject to penalty parameter λ. λ is scaled as $1-\lambda/\lambda_{max}$, where λ_{max} refers to the smallest value of penalty parameter λ that already gives maximal penalization, i.e., the smallest λ that sets all penalized coefficients to zero. Hence, we see ML-estimates at the right end. The left end relates to maximal penalization, here only the intercept remains non-zero. In the left panel, the penalty is adaptive, the weights are fixed (see equation (2.7) with $b_0 = 0$, $\phi_{rs(j)} = \phi_{r(j)} = 0.5$). The paths show how clustering/selection of coefficients works: Even slight penalization discovers the intercept to be non-varying, coefficients of covariate x_1 are fused such that only category 1 makes a difference. Concerning covariate x_2 coefficients should be fused to one non-varying scalar. But stronger penalties are necessary to make this happen. The dotted line marks the optimal model in terms of 5-fold cross-validation with the predictive deviance $Dev(y, \hat{\mu})$ as loss function. It shrinks coefficients slightly – in return all but one relevant structures are identified. Absolute deviation to the true coefficients is small.

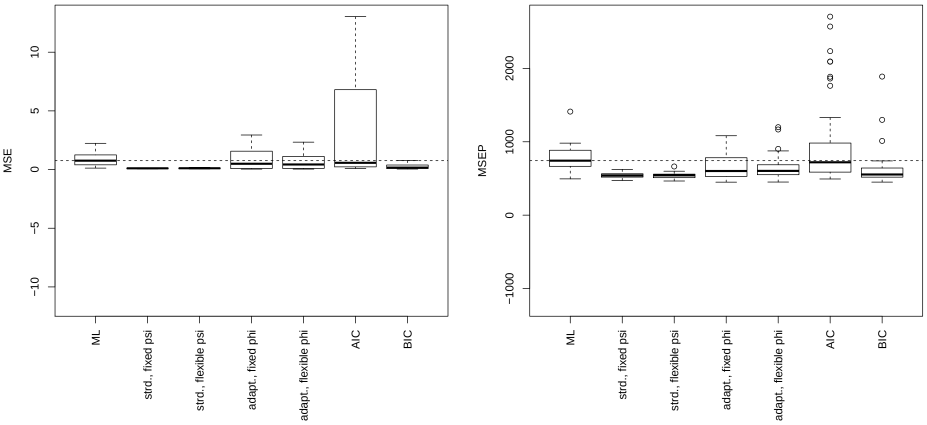

Fig 2. *Boxplots of scaled squared errors (MSE, left panel) and deviances (MSEP) for setting b26.200; in the left panel outliers are omitted.*

When the standard penalty (2.3) is used instead, results change: while coefficient paths remain basically the same in structure, the standard penalty slows down fusion and selection of coefficients (see Figure 1, right panel). To reach the same effects, stronger penalization is needed. Cross-validated λ_{CV} is 2.25 now. However, performance is worse than with adaptive weights: in the model chosen by cross-validation (see dotted line), coefficients of covariate x_1 are not fused.

4.2. *Simulation Settings.* To compare the proposed methods, various model features are systematically varied. Concretely, we consider a binomial response, there are two influential covariates, and we add 6 non-influential noise variables. Training data sets contain $n = 200$ and $n = 600$ observations, test data sets $n = 600$ and $n = 1800$ observations, respectively. That is, we have two settings named *b26.200* and *b26.600*. All covariates are continuous and independently drawn from an uniform distribution $U[-2, 2]$. There is a known effect modifier. It is nominal, has four categories $1, \dots, 4$ and is independently drawn from a multinomial distribution with probability 0.25 per category. The true linear predictor is

$$\begin{aligned}
\eta_{true} &= \beta_0(u) + x_1\beta_1(u) + x_2\beta_2(u)\\
&= (\ 0.7I(u=1) + 0.7I(u=2) + 0I(u=3) + 0I(u=4)\ )\\
&\quad + x_1\,(\ 1I(u=1) - 1.5I(u=2) - 1.5I(u=3) + 0.5I(u=4)\ )\\
&\quad + x_2\,(\ 0I(u=1) + 1I(u=2) + 2I(u=3) - 3I(u=4)\ )\,.
\end{aligned}$$

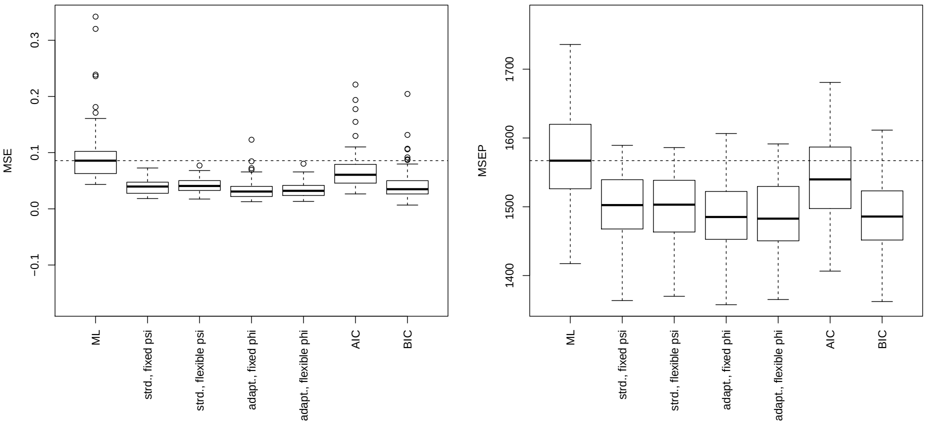

FIG 3. *Boxplots of scaled squared errors (MSE, left panel) and deviances (MSEP) for setting b26.600; medians mark estimates of MSE and MSEP.*

Since the procedure does not know which coefficients are actually varying, all coefficients are allowed to vary with effect modifier u. As six non-influential noise variables $n_3, \ldots, n_8$ are added, the assumed predictor is

$$\eta_{model} \quad = \quad \beta_0(u) + x_1 \cdot \beta_1(u) + x_2 \cdot \beta_2(u) + n_3 \cdot \beta_3(u) + \ldots + n_8 \cdot \beta_8(u).$$

This model is estimated using all the methods discussed. That means, we consider various penalized estimates: with weight ψ fixed at 0.5, with flexible weight ψ, with adaptive weights and fixed $\phi_{rs(j)}$, $\phi_{r(j)}$ ($\phi_{rs(j)} = \phi_{r(j)} = \phi = 0.5$), with adaptive weights and flexible $\phi_{rs(j)}$, $\phi_{r(j)}$ ($\phi_{rs(j)} = \phi$, $\phi_{r(j)} = 1 - \phi$). In addition, we consider forward selection strategies with criteria AIC and BIC, and the usual ML-estimate. For ML-estimates, neither regularization nor model selection is required. They are the benchmark for all the other estimators' performances. Penalty parameter λ is chosen by 5-fold cross-validation. If weights ψ and ϕ are flexible, they are cross-validated, too. For each setting, all models are computed 50 times in order to make the results reliable.

4.3. *Results.* To assess parameter estimation, we compute the coefficients' mean squared error for each simulation run:

$$\hat{\text{MSE}}(\beta, \hat{\beta}) = \frac{1}{q} \sum_{j=1}^{q} \left(\beta_j - \hat{\beta}_j \right)^2 ,$$

		ML	strd., fixed psi	strd., flexible psi	adapt., fixed phi	adapt., flexible phi	AIC	BIC
Setting *b26.200*	$\mathrm{FPR}_{selection}$	1	0.77	0.65	0.34	0.39	0.41	0.16
	$\mathrm{FNR}_{selection}$	0	0.03	0.04	0.08	0.06	0.07	0.11
	$\mathrm{FPR}_{clustering}$	1	0.64	0.69	0.43	0.42	0.40	0.10
	$\mathrm{FNR}_{clustering}$	0	0.05	0.03	0.15	0.15	0.19	0.27
Setting *b26.600*	$\mathrm{FPR}_{selection}$	1	0.81	0.71	0.43	0.39	0.39	0.11
	$\mathrm{FNR}_{selection}$	0	0.00	0.01	0.01	0.01	0.02	0.03
	$\mathrm{FPR}_{clustering}$	1	0.77	0.76	0.45	0.42	0.37	0.05
	$\mathrm{FNR}_{clustering}$	0	0.01	0.00	0.04	0.03	0.08	0.17

TABLE 1

Estimates of false positive and false negative rates for settings b26.200 *and* b26.600.

where $q = \sum_{j=0}^{p} k_j$, β denotes the vector of true coefficients, and $\hat{\beta}$ its estimate. To judge the prediction accuracy, the mean predictive deviance $Dev(y, \hat{\mu})$ is considered, referred to as MSEP. Figures 2 and 3 show the box plots of MSE and MSEP for both settings. Median values of penalized approaches and forward selection strategies are smaller than those of the ML-estimates. However, forward selection strategies suffer from a high variability – especially for $n = 200$ they are very unstable. For $n = 600$ and adaptive weights, interquartile ranges become smaller compared to “standard” penalization. This is due to the construction of the adaptive weights, which are the inverses of the ML-estimates. The more observations we have the better is the ML-estimate and so are the adaptive weights.

In addition, we evaluate the clustering and selection performance. A model selection strategy should exclude non-influential covariates, especially pure noise variables. That is, truly zero coefficients should not be selected. Truly non-varying coefficients should be fused. For evaluation, we consider false negative (FNR) and false positive rates (FPR). False positive means that a truly zero coefficient is fitted as non-zero. False negative means that truly non-zero values are estimated to be zero. With # denoting “the number of coefficients” we have

$$\mathrm{FPR}_{selection} = \frac{\#(\text{truly zero set to non-zero})}{\#(\text{truly zero})} \quad \text{and}$$

$$\mathrm{FNR}_{selection} = \frac{\#(\text{truly non-zero set to zero})}{\#(\text{truly non-zero})}.$$

Variable	Description
cesarean	Type of delivery (0: vaginal, 1: Cesarean), response
term	Term of pregnancy in weeks form the last menstruation
c.height	Height of child at birth in centimeter
c.weight	Weight of child at birth in gram
m.age	Age of mother before pregnancy in years
m.height	Height of mother in centimeter
m.bmi	BMI of mother before pregnancy (mass (kg)/(height (m))^2)
m.gain.w	Gain in weight of mother during pregnancy in kg
m.prev	Number of previous pregnancies
ind	Was the labor induced?
memb	Did the membranes burst before the beginning of the throes?
rest	Was a strict bed rest ordered to the mother for at least one month during the pregnancy?
cephalic	Was the child in cephalic presentation before birth?
t	Year of birth, effect modifier

TABLE 2

Short description of response, covariates and the effect modifier for birth data. The coding of binary covariates is 0 for "no", 1 for "yes".

FPR$_{clustering}$ and FNR$_{clustering}$ are defined analogously, but refer to differences of coefficients. Table 1 shows false positive and negative rates for both settings. Overall it stands out that forward selection strategies perform well. However, having the high variability of forward selection strategies in mind and looking at both clustering and selection, the previous recommendation for adaptive weights still holds.

5. Application: Cesareans among Francophone Mothers. Our data set contains various variables related to the pregnancy and delivery recruited on French-speaking websites. The data was presented by Boulesteix (2006) and is available in `R` add-on package `catdata` (Tutz and Schauberger, 2010). As described in Section 1, we are interested in the type of delivery, in whether birth was given vaginally or by means of a Cesarean. Between 2001 and 2004, 578 deliveries were observed, and modeling the type of delivery requires to allow covariate effects to vary with time, since, e.g., medical standards may have changed over time. As the time is measured discretely and on a rough grid, we consider the time in years as an ordinal effect modifier in a varying-coefficient model. The response is binary indicating the type of delivery; 0 stands for a vaginal birth, 1 for a Cesarean. The model considers all covariates that were available and meaningful for all women. Details on the covariates are found in Table 2. To be on comparable scales, all covariates are standardized. As terms and delivery circumstances differ immensely for multiple births, these cases are excluded.

As we have no prior knowledge about the model's structure, effect modifier t potentially impacts all coefficients. As there is a relatively large number of covariates, we are not only interested in parameter fusion, but also in the selection of coefficients $\beta_j(t)$. Table 4 shows the resulting estimates. The values of ML-estimates are quite extreme. To obtain a stable estimation procedure, that is able to select among predictors, regularization is required. As suggested by the numerical experiments in Section 4, we employ an adaptive penalty with fixed weights $\phi_{rs(j)} = \phi_{r(j)} = 0.5$. Penalty parameter λ is cross-validated and set to 1.33. This is small compared to the minimal value of λ giving maximal penalization $\lambda_{max} = 211.5$. But it stabilizes estimation and shrinks unstable ML-estimates enormously. As an alternative, we consider forward selection strategies as presented in Section 3. Forward selection strategies produce very sparse estimates. Only three (AIC), respectively one (BIC), coefficients are partly varying. ML-estimates, by contrast, argue for a strong dependency on time, see for example the intercept of the year 2001, which is ignored by forward selection strategies. Penalized estimation gives a more differentiated picture. It selects predictors and shows that not all time points have different effects. For example the intercept shows that there is a decrease of vaginal births over time and that in years 2002 and 2003 the preference of vaginal births is the same. In contrast, highly volatile ML-estimates of the intercepts are not strictly decreasing.

6. Application: Reducing Mortality after Myocardial Infarction.

In this second application we consider a 22-center clinical trial of beta-blockers for reducing mortality after myocardial infarction. The dataset is for example described in Aitkin (1999) and available in `R` add-on package `flexmix` (Grün and Leisch, 2008). For each center the number of deceased/successfully treated patients in control/test groups is known. We are going to model the mortality rate depending on the centers and the treatment groups; that means the response y is binomial. The data has been analyzed by different authors: Aitkin (1999) modeled the effect of the study centers by random intercepts. That is, the predictor is defined as

$$\eta_{ij} = \beta_0 + b_{0i} + \beta_T \cdot Treatment_{ij}, \quad i = 1, \ldots, 22 \text{ Centers}, \quad j \in \{\text{control}, \text{test}\},$$

where b_{0i} is normally distributed, b_{0i} $N(0, \sigma^2)$. The corresponding marginal likelihood is numerically approximated by a Gauss-Hermite quadrature with four mass points. One obtains the treatment effect β_T and estimates $\hat{b}_{0i}$. However, centers are not clustered.

Grün and Leisch (2008) try to find similar centers with discrete mixture

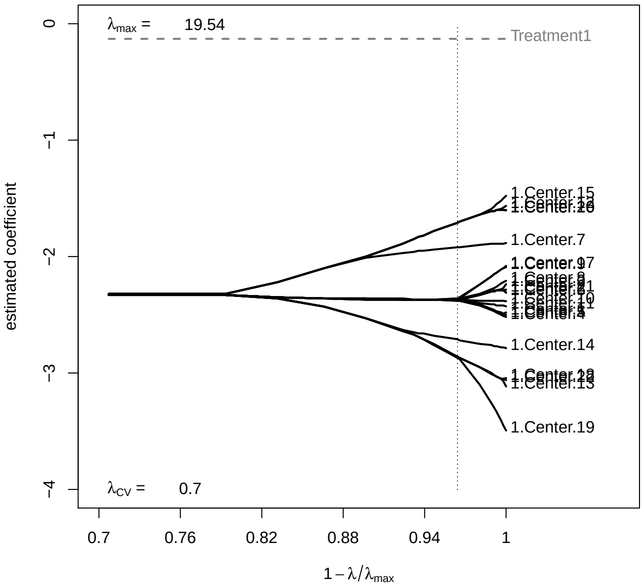

Fig 4. *Coefficient paths beta-blocker data.*

models. They use the predictor

$$\eta_i = \beta_{0m} + \beta_T \cdot Treatment_i,\ i = 1, \ldots, 44 \text{ Cases},$$

where $m \in \{1, \ldots, K\}$ refer to the partition of the 22 centers into K groups. The predictor contributes to the mixture likelihood

$$L(\beta_0, \beta_T, \pi; y) = \prod_{i=1}^{44} \left(\sum_{m=1}^{K} \pi_m f_m(\eta_i, \Psi_m) \right),$$

with $\beta_0 = (\beta_{01}, \ldots, \beta_{0k})^T$ and with $\pi = (\pi_1, \ldots, \pi_K)^T$ denoting the priori probabilities of the components ($\sum_{m=1}^{K} \pi_m = 1, \quad \pi_m > 0\ \forall m$). Functions $f_m(\cdot)$ denote the components' densities; for each component a simple exponential family with parameters Ψ_m is assumed. For estimation an iterative EM-algorithm (Dempster, Laird and Rubin, 1977, Leisch, 2004) with $K = 3$, respectively $K = 5$, components is employed. Hence, the centers are clustered, but the number of clusters has to be specified in advance.

To overcome these problems, we assume a varying intercept model with predictor:

$$\eta_i = \beta_0(Center_i) + \beta_T \cdot Treatment_i,\ i = 1, \ldots, 44 \text{ Cases}. \tag{6.1}$$

Coefficients		ML	Random Intercept Model	Varying Intercept Model	Discrete Mixture Model	
					5 Cluster	3 Cluster
Center-specific Intercept	$\beta_{0,15}$	-1.4782	-1.5519	-1.71	-1.5687	-1.7388
	$\beta_{0,12}$	-1.5644	-1.6052			
	$\beta_{0,16}$	-1.5999	-1.6493			
	$\beta_{0,20}$	-1.6038	-1.6523			
	$\beta_{0,7}$	-1.8832	-1.8917	-1.92	-1.9024	
	$\beta_{0,17}$	-2.0801	-2.1065	-2.36	-2.3224	-2.3793
	$\beta_{0,9}$	-2.0910	-2.1079			
	$\beta_{0,8}$	-2.2083	-2.2132	-2.37		
	$\beta_{0,3}$	-2.2370	-2.2574			
	$\beta_{0,21}$	-2.2832	-2.2859			
	$\beta_{0,2}$	-2.3059	-2.3097			
	$\beta_{0,6}$	-2.3113	-2.3162			
	$\beta_{0,10}$	-2.3840	-2.3832			
	$\beta_{0,11}$	-2.4278	-2.4239			
	$\beta_{0,1}$	-2.4798	-2.4145			
	$\beta_{0,5}$	-2.5015	-2.4881	-2.38	-2.4589	
	$\beta_{0,4}$	-2.5189	-2.5151			
	$\beta_{0,14}$	-2.7862	-2.7670	-2.71	-2.9632	-2.9628
	$\beta_{0,18}$	-3.0433	-2.8805	-2.86		
	$\beta_{0,22}$	-3.0610	-3.0123			
	$\beta_{0,13}$	-3.1155	-3.0022			
	$\beta_{0,19}$	-3.4942	-3.1541	-2.87		
Treatment	β_T	-0.1305	-0.1305	-0.13	-0.1295	-0.1291

TABLE 3

Resulting estimates of all considered methods for the beta-blocker data. Intercept-coefficients are ordered such that their structure becomes obvious. "ML" stands for the ML-estimate of a GLM containing an intercept and effect coded covariates Center, Treatment; to keep things comparable, that linear combination of the coefficients that corresponds to the other models is shown. Presented intercept-coefficients of the mixed model are the sum of the fixed and the random effects. Horizontal lines denote clusters of coefficients.

In order to obtain comparable results and as there is only one covariate, the data is not scaled. The nominal information about the center is the effect modifier. In analogy to Aitkin (1999) and Grün and Leisch (2008), the explanatory covariate "Treatment" is not modified and effect coded. For estimation the penalized likelihood (2.1) with adaptive weights (2.9) and (2.10) is employed. As suggested in Section 4, weighting parameter ψ is fixed at 0.5. Hence, the centers' possible diversity is considered. Due to penalized estimation the intercept-coefficients of several centers can be merged – clusters of similar centers are detected. As penalty parameter λ is cross-validated, quantity and quality of clusters are determined by the data.

Figure 4 gives the resulting coefficient paths for model (6.1). There seem to be three, respectively five different types of basically different study centers. Cross-validation yields $\lambda_{CV} = 0.7$ and is marked by the dotted line in Figure 4. At this point the main clusters are detected, while subtle distinctions between the centers are still apparent. Table 3 gives the resulting coefficients. Results are compared to the random intercept model of Aitkin (1999) and the finite mixture model of Grün and Leisch (2008) with adjusted coding. It is seen that the obtained clusters of the varying intercept model show the same structure as finite mixture models. The random intercepts show the same profile as our results, but no clusters. All estimates have the same scale. The treatment effect is detected in all models and – this is remarkable – of approximately the same size. But only the varying coefficient model combines data driven clustering with stable results. When weighting parameter ϕ and penalty parameter λ are cross-validated, we obtain nearly the same results; order and clusters of coefficients are the same. Note that predictor (6.1) in the varying intercept model corresponds to a GLM with penalized nominal covariates Center and Treatment. However, the representation as varying coefficient model makes interpretation easier. It offers an attractive alternative to finite mixture models.

One may also wonder whether the treatment effect does depend on the according study center, too. For this reason we consider a second model with predictor

$$\eta_i = \beta_0(Center_i) + \beta_T(Center_i) \cdot Treatment_i, \; i = 1, \ldots, 44 \text{ Cases} \tag{6.2}$$

and the same assumptions as above. As there is only one covariate and one effect modifier, which are both categorial, predictor (6.2) corresponds to a GLM with covariates Center, Treatment and their interaction. This is a saturated model. There are as many free parameters as observed Center-Treatment constellations. Hence, observed mortality is perfectly replicated by the model. In this case, only regularization results in a model that can be interpreted. Cross-validation of λ (and ϕ) fuses $\beta_T(Center_i)$ to one constant coefficient. The varying intercept $\beta_0(Center_i)$ shows the same clusters as for predictor (6.1); such that the “fixed” treatment effect assumed in Aitkin (1999) and Grün and Leisch (2008) is supported.

7. Special Case: Categorical Effects. So far, we considered categorical effect modifiers in general. We did not touch categorical effects, which are a special case of categorical effect modifiers. One obtains a coded categorical effect, when the effect modifier u_j is categorical and the modified covariate x_j is a constant vector. We have for example $1 \cdot \beta_j(u_j) = 1 \cdot \sum_{r=1}^{k_j} \beta_{jr} I(u_j = r)$.

Penalization remains the same. Statements made for penalized varying coefficients hold for penalized categorical effects, too. Especially large sample properties can be transferred. However, the devil is in the details: unlike usual coding, the obtained coding does not contain a reference category. This implies at least two things: the design matrix is not of full rank and interpretation changes. As estimation is penalized and the tuning parameter λ will be cross-validated in most cases, the first aspect can be neglected. Concerning interpretation, penalized estimates can be transformed, such that they correspond directly to usual coding of categorical effects. Note, however, the penalty we use here is not designed for a reference category. In contrasts to Gertheiss, Stelz and Tutz (2012), all categories of a categorical effect are penalized in the same way.

8. Concluding Remarks. We investigated categorical effect modifiers within the framework of GLMs. When selecting a model with categorical effect modifiers, one wants to find out which covariates have an effect on the response, and if so, which categories have to be distinguished. In fact, this is a recoding of usual interactions between categorial and metric predictors, but the concept of effect modifiers allows for interpretable model selection strategies. We presented two different approaches: on the one hand we extended the ideas of Tibshirani et al. (2005) to varying-coefficient models with categorical effect modifiers. Thus, we are able to simultaneously identify varying coefficients and select covariates in GLMs. The penalty adjusts for the different amount of information in nominal and ordinal effect modifiers. An adaptive version of the proposed penalty was shown to be asymptotically normal and consistent. These results remain valid when scale parameter ϕ of the exponential family is estimated and plugged-in, which allows for quasi-likelihood approaches. On the other hand, we investigated a modified forward selection strategy: start with a null-model and add one degree of freedom in each iteration until a chosen criterion is not improved anymore. Numerical experiments suggested both methods to be highly competitive. Penalized estimates and forward selection strategies performed distinctly better than un-penalized ML-estimates. Forward selection strategies, however, suffer from immense variability, which makes them less attractive.

Lasso-type penalties imply not continuously differentiable optimization problems, which we solved by adopting an algorithm of Fan and Li (2001). All functions are available in the `R` add-on package `gvcm.cat` (Oelker, 2012).

In practice, varying-coefficient models are highly relevant. We analyzed Cesareans among francophone mothers. We were interested in how the influence of various medical indicators changed over time. The data is quite challeng-

Coefficients	ML-estimation				Penalized estimation				Forward Selection AIC				Forward Selection BIC			
	t				t				t				t			
	2001	2002	2003	2004	2001	2002	2003	2004	2001	2002	2003	2004	2001	2002	2003	2004
$\beta_0(t)$	81.5383	22.5480	30.5718	7.0074	22.81	13.03	13.03	8.26	14.78				11.53			
$\beta_{\text{term}}(t)$	2.1706	2.8295	−.5030	−.3152	1.12	1.12	−.34	−.33	−.26							
$\beta_{\text{c.height}}(t)$	.5654	−.9737	−.2556	.0692	.10	−.08	−.09									
$\beta_{\text{c.weight}}(t)$	4.6391	1.6888	−.1752	.0290	1.83	.68	−.01									
$\beta_{\text{m.age}}(t)$	−2.8575	.9047	.4242	.3424	−.98	.29	.29	.29	.31							
$\beta_{\text{m.height}}(t)$	−5.7050	−2.8659	−.2985	−.1649	−2.14	−1.56	−.12	−.12	−.40				−.42			
$\beta_{\text{m.bmi}}(t)$	−.2629	.9572	.2548	.0586		.26	.04									
$\beta_{\text{m.gain.w}}(t)$	−.9757	.0079	.2575	−.1357												
$\beta_{\text{m.prev}}(t)$	−14.0886	−13.2753	−.9469	−.7127	−2.98	−2.98	−.68	−.68	−.85				−.74			
$\beta_{\text{ind}}(t)$	5.7853	−1.2505	.7874	.2229	2.19	−.22	.63	.14	.65	.65	.65	.25	.48			
$\beta_{\text{memb}}(t)$	−.5437	.6613	−.4523	−.5386			−.38	−.38			−.51	−.51			−.45	−.45
$\beta_{\text{rest}}(t)$	−1.4703	−8.2928	−.2618	−.1436		−.80			−.39	−.39	−.39					
$\beta_{\text{cephalic}}(t)$	−.7559	−22.4778	−10.4814	−.8532		−4.95	−3.23	−.74	−1.51				−1.52			

TABLE 4. *Estimates for all methods fitted to the birth data. Excluded coefficients are omitted. Non-varying coefficients are represented by the remaining scalar only.*

ing, standard approaches fail. However, penalized estimates give a coherent trend.

Furthermore, we applied the proposed methods to a clinical trial on reducing mortality after myocardial infarction. We were interested in how diverse study centers are. Penalized estimation turned out to be a stable alternative to finite mixture models. Quantity and quality of clusters was detected data-driven. We observed the same coefficient profile as for a random intercept model.

So far we employed a single penalty parameter λ only; for a modest number of effect modiffiers, however, one tuning parameter per effect modifier could be advantageous. But computational complexity increases very fast with the number of tuning parameters.

The proposed penalty's potential is apparent: for longitudinal studies its scope can be enlarged to marginal models; and it can be further generalized: varying coefficients may depend on more than one effect modifier. In this paper we assumed continuous covariates $x_1, \dots, x_p$. But of course covariates can be categorical, too. Then there are even more coefficients, and hence, there is an even stronger demand for regularization.

APPENDIX A: PROOF OF THEOREM 2.1

If $\hat{\beta}$ minimizes $\mathcal{M}_n^{pen}(\beta)$ with $J_n(\beta)$ as defined by $J_n(\beta)$, with $J_j^{nom}(\beta_j)$ and $J_j^{ord}(\beta_j)$, then it also minimizes $\mathcal{M}_n^{pen}(\beta)/n$. The ML-estimate $\hat{\beta}^{ML}$ minimizes $\mathcal{M}_n(\beta) = -l_n(\beta)$, respectively $\mathcal{M}_n(\beta)/n$. Since λ is fixed, $\mathcal{M}_n^{pen}(\hat{\beta})/n \xrightarrow{\mathbb{P}} \mathcal{M}_n(\hat{\beta}^{ML})/n$ and $\mathcal{M}_n^{pen}(\hat{\beta})/n \xrightarrow{\mathbb{P}} \mathcal{M}_n(\hat{\beta})/n$, $\mathcal{M}_n(\hat{\beta})/n \xrightarrow{\mathbb{P}} \mathcal{M}_n(\hat{\beta}^{ML})/n$ hold as well. Since $\hat{\beta}^{ML}$ is the unique minimizer of $\mathcal{M}_n(\beta)/n$, and $\mathcal{M}_n(\beta)/n$ is convex, we have $\hat{\beta} \xrightarrow{\mathbb{P}} \hat{\beta}^{ML}$; and consistency follows from consistency of the ML-estimate $\hat{\beta}^{ML}$, under assumptions given for example by Fahrmeir and Kaufmann (1985).

APPENDIX B: PROOF OF THEOREM 2.2

Due to the additivity of arguments, a predictor of the following form can be assumed without loss of generality:

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

i.e., only one effect modifier u is assumed. In addition, let Z denote the design matrix given by $Z = (Z_0, \dots, Z_p)$, where

$$Z_j = \begin{pmatrix} x_{1j}I(u_{1j} = 1) & \cdots & x_{1j}I(u_{1j} = k_j) \\ \vdots & \ddots & \vdots \\ x_{nj}I(u_{nj} = 1) & \cdots & x_{nj}I(u_{nj} = k_j) \end{pmatrix}.$$

B.1. Normality.

B.1.1. *Redefinition of the Objective Function.* Redefine optimization problem $\mathcal{M}_n^{pen}(\beta)$ as $\text{argmin}_\beta\ \Psi_n(\beta)$, where $\Psi_n(\beta) = -l_n(\beta) + \frac{\lambda_n}{\sqrt{n}} J_n(\beta)$. $J_n(\beta)$ denotes the penalty term. Unlike before tuning parameter λ is divided by factor $\sqrt{n}$, in turn the penalty $J_n(\beta)$ is multiplied by the same factor:

$$J_n(\beta) = \sqrt{n}\left(\sum_{j=0}^{p}\sum_{r>s} w_{rs(j)}|\beta_{jr} - \beta_{js}| + \sum_{j=1}^{p}\sum_{r=1}^{k} w_{r(j)}|\beta_{jr}|\right).$$

The log-likelihood is defined as

$$l_n(b) = \sum_{i=1}^{n} \frac{y_i\vartheta_i(\mu_i) - b(\vartheta_i(\mu_i))}{\varphi_i} = \sum_{i=1}^{n} \frac{y_i\vartheta_i(h(z_i^T\beta)) - b(\vartheta_i(h(z_i^T\beta)))}{\varphi_i},$$

that is, $l_n(b)$ is determined by a simple exponential family where $\vartheta_i \in \Theta \subset \mathbb{R}$ is the natural parameter of the family depending on expectation μ_i; φ_i is a scale or dispersion parameter, $b(\cdot)$ and $c(\cdot)$ are specific functions corresponding to the type of the family. For given φ_i, one assumes Θ to be the natural parameter space, i.e., the set of all ϑ_i satisfying $0 < \int \exp(y_i\vartheta_i/\varphi_i + c(y_i, \varphi_i))\mathrm{d}y_i < \infty$. Then, Θ is convex, and in the nonempty interior Θ^0 all derivatives of $b(\vartheta_i)$ and all moments of y_i exist, see Fahrmeir and Tutz, 2001. Hence it is equivalent to solve

$$\text{argmin}_\beta\ V_n(\beta) = \text{argmin}_\beta\ 2\left(\Psi_n(\beta) - \Psi_n(\beta^*)\right)$$

with

$$\begin{aligned} V_n(\beta) &= -2\left(l_n(\beta) - l_n(\beta^*)\right) + 2\frac{\lambda_n}{\sqrt{n}}\left(J_n(\beta) - J_n(\beta^*)\right) \\ &= -2\left(l_n(\beta) - l_n(\beta^*)\right) + 2\frac{\lambda_n}{\sqrt{n}}\tilde{J}_n(\beta). \end{aligned}$$

B.1.2. *Limit Behavior.* Following Bondell and Reich (2009) closely, $\tilde{J}_n(\beta)$ with respect to b is considered; with $b = \sqrt{n}(\beta - \beta^*)$ and $\beta = \beta^* + b/\sqrt{n}$, where β^* denotes the true coefficient vector:

$$\begin{aligned} &\tilde{J}_n(\beta) = J_n(\beta) - J_n(\beta^*) \Rightarrow \\ &\tilde{J}_n(b) = J_n(b) - J_n(0) \\ &\quad = \sum_{j=0}^{p}\sum_{r>s}\sqrt{n}\frac{\phi_{rs(j)}(n)}{|\hat{\beta}_{jr}^{ML} - \hat{\beta}_{js}^{ML}|}\left|\beta_{jr}^* - \beta_{js}^* + \frac{b_{jr} - b_{js}}{\sqrt{n}}\right| \end{aligned}$$

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

Distinction of cases (1) $\beta_{jr}^* \neq \beta_{js}^*$ and $\beta_{jr}^* \neq 0$, i.e., if $\theta_i^* \neq 0$.
As given in Zou (2006), we will consider the limit behavior of $(\lambda_n/\sqrt{n})\tilde{J}_n(b)$. If $\beta_{jr}^* \neq \beta_{js}^*$, then

$$|\hat{\beta}_{jr}^{ML} - \hat{\beta}_{js}^{ML}| \xrightarrow{\mathbb{P}} |\beta_{jr}^* - \beta_{js}^*|$$

and

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

(if n large enough); and similarly, if $\beta_{jr}^* \neq 0$, then

$$|\hat{\beta}_{jr}^{ML}| \xrightarrow{\mathbb{P}} |\beta_{jr}^*|$$

and

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

(if n large enough). 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

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

and

$$\frac{\lambda_n}{\sqrt{n}}\sqrt{n}\frac{\phi_{r(j)}(n)}{|\hat{\beta}_{jr}^{ML}|} \left(\left|\beta_{jr}^* + \frac{b_{jr}}{\sqrt{n}}\right| - \left|\beta_{jr}^*\right| \right) \xrightarrow{\mathbb{P}} 0$$

respectively. That means, if $\theta_i^* \neq 0$, we have $\frac{\lambda_n}{\sqrt{n}}\tilde{J}(b) \xrightarrow{\mathbb{P}} 0$.

Distinction of cases (2) $\beta^*_{jr} = \beta^*_{js}$ or $\beta^*_{jr} = 0$, i.e., if $\theta^*_i = 0$
Here it holds that

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

and

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

Moreover, due to the consistency of the ML-estimates, we have

$$\hat{\beta}^{ML} - \beta^* = F_n^{-1}(\beta^*)s_n(\beta^*) + \mathcal{O}(n^{-1}),$$

where $\mathcal{O}$ denotes the Landau notation, $F_n(\beta^*) = \mathcal{O}(n)$ and $s_n(\beta^*) = \mathcal{O}(n^{1/2})$. Therefore $s_n(\beta^*)/F_n(\beta^*) < c \cdot n^{-1/2}$ (c is some constant), $s_n(\beta^*)/F_n(\beta^*) = \mathcal{O}(n^{-1/2})$ and $\hat{\beta}^{ML} - \beta^* = \mathcal{O}(n^{-1/2})$ (McCullagh, 1983). As a conclusion, it holds that

$$\lim_{n\to\infty}\mathbb{P}\left(\sqrt{n}|\hat{\beta}^{ML}_{jr} - \hat{\beta}^{ML}_{js}| \leq \lambda_n^{1/2}\right) = 1$$

or

$$\lim_{n\to\infty}\mathbb{P}\left(\sqrt{n}|\hat{\beta}^{ML}_{jr}| \leq \lambda_n^{1/2}\right) = 1$$

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

$$\frac{\lambda_n}{\sqrt{n}}\sqrt{n}\frac{\phi_{rs(j)}(n)}{|\hat{\beta}^{ML}_{jr} - \hat{\beta}^{ML}_{js}|}\left(\left|\beta^*_{jr} - \beta^*_{js} + \frac{b_{jr} - b_{js}}{\sqrt{n}}\right| - \left|\beta^*_{jr} - \beta^*_{js}\right|\right) \xrightarrow{\mathbb{P}} \infty$$

or

$$\frac{\lambda_n}{\sqrt{n}}\sqrt{n}\frac{\phi_{r(j)}(n)}{|\hat{\beta}^{ML}_{jr}|}\left(\left|\beta^*_{jr} + \frac{b_{jr}}{\sqrt{n}}\right| - \left|\beta^*_{jr}\right|\right) \xrightarrow{\mathbb{P}} \infty$$

if $b^*_{jr} \neq 0$, respectively $b^*_{jr} \neq b^*_{js}$. That means, if for any r, s, j with $\beta^*_{jr} = 0$ ($j > 0$) or $\beta^*_{jr} = \beta^*_{js}$ ($j \geq 0$), $b_{jr} \neq 0$ or $b_{jr} \neq b_{js}$, respectively, then we have $\frac{\lambda_n}{\sqrt{n}}\tilde{J}(b) \xrightarrow{\mathbb{P}} \infty$.

B.1.3. *Normality.* Before we have a look at $-2\left(l_n(\beta) - l_n(\beta^*)\right)$ remember that an expansion of usual ML-equations $s(\beta) = 0$ about β^* gives

$$s_n(\beta^*) = \frac{\partial^2 l_n(\beta)}{\partial\beta\partial\beta^T}|_{\beta=\beta^*}(\beta - \beta^*).$$

Hence in usual GLMs, it holds that

$$\beta - \beta^* = \frac{\partial^2 l_n(\beta)}{\partial\beta\partial\beta^T}|_{\beta=\beta^*}s_n(\beta^*) = F_n^{-1}(\beta^*)s_n(\beta^*) + \mathcal{O}_p(n^{-1})$$

Multiplying both sides by $n^{1/2}$, using $F_n(\beta^*)/n \overset{n\to\infty}{\to} F(\beta^*)$ and $n^{-1/2}s_n(\beta^*) \xrightarrow{d} N(0, F(\beta^*))$, one obtains

$$n^{1/2}(\hat{\beta}^n - \beta^*) \xrightarrow{d} N(0, F(\beta^*)^{-1})$$

in usual GLMs (McCullagh, 1983). Back to the given varying-coefficient model, consider now $-2\,(l_n(\beta) - l_n(\beta^*))$ instead of $V_n(\beta) = -2\,(l_n(\beta) - l_n(\beta^*)) + 2\frac{\lambda_n}{\sqrt{n}}\tilde{J}_n(\beta)$. An expansion of $l_n(\beta)$ about β^* gives

$$-2\,(l_n(\beta) - l_n(\beta^*)) = (\beta - \beta^*)^T \frac{\partial^2 l_n(\beta)}{\partial\beta\partial\beta^T}|_{\beta=\beta^*}(\beta - \beta^*).$$

Applying $\frac{\partial^2 l_n(\beta)}{\partial\beta\partial\beta^T}|_{\beta=\beta^*}(\beta - \beta^*) = s_n(\beta^*)$ for $-2\,(l_n(\beta) - l_n(\beta^*))$ as well, one obtains

$$-2\,(l_n(\beta) - l_n(\beta^*)) = (\beta - \beta^*)^T \frac{\partial^2 l_n(\beta)}{\partial\beta\partial\beta^T}|_{\beta=\beta^*}(\beta - \beta^*) = s_n^T(\beta^*)F_n^{-1}(\beta^*)s_n(\beta^*).$$

Following Bondell and Reich (2009), let $\theta_{\mathcal{C}}$ denote the vector of θ-entries which are truly non zero, i.e., from $\mathcal{C}$, and let $\beta_{\mathcal{C}}$ be the subset of entries of $\theta_{\mathcal{C}}$ which are part of β. By contrast $\theta_{\mathcal{C}^{\mathrm{C}}}$ denotes the vector of θ-entries which are truly zero and therefore not from $\mathcal{C}$ but from $\mathcal{C}^{\mathrm{C}}$; analogously to $\beta_{\mathcal{C}}$, $\beta_{\mathcal{C}^{\mathrm{C}}}$ is defined as the subset of entries of $\theta_{\mathcal{C}^{\mathrm{C}}}$ which are part of β. Since $n \to \infty$, and applying $F_n(\beta^*)/n \overset{n\to\infty}{\to} F(\beta^*)$ one more time, we have $V_n(\beta) \to V(\beta)$ for every β, where

$$V(\beta) = \begin{cases} \frac{1}{n}s_n^T(\beta_{\mathcal{C}})F^{-1}(\beta_{\mathcal{C}})s_n(\beta_{\mathcal{C}}) & \text{if } \theta_{\mathcal{C}^{\mathrm{C}}} = 0, \\ \infty & \text{otherwise,} \end{cases}$$

and where $s_n(\beta_{\mathcal{C}})$ are regular ML-equations. Therefore it holds that $n^{-1/2}s_n(\beta^*_{\mathcal{C}}) \xrightarrow{d} N(0, F(\beta^*_{\mathcal{C}}))$ and $n^{-1/2}(\beta_{\mathcal{C}} - \beta^*_{\mathcal{C}}) \xrightarrow{d} N(0, F(\beta^*_{\mathcal{C}})^{-1})$ like mentioned above. Since the considered minimization problem is convex, the unique minimum of $V(\beta)$ is $(\beta^{ML}_{\mathcal{C}}, 0)^T$ and we have

$$\hat{\beta}^n_{\mathcal{C}} \to \beta^{ML}_{\mathcal{C}} \text{ and } \hat{\beta}^n_{\mathcal{C}^{\mathrm{C}}} \to 0.$$

Hence, we have as well

$$n^{-1/2}(\hat{\beta}^n_{\mathcal{C}} - \beta^*_{\mathcal{C}}) \xrightarrow{d} N(0, F(\beta^*_{\mathcal{C}})^{-1})$$

Via a reparametrization of β as, for example, $\check{\beta} = (\check{\beta}_0^T, ..., \check{\beta}_p^T)^T$, with $\check{\beta}_j = (\beta_{jr} - \beta_{j1}, ..., \beta_{jr}, ..., \beta_{jr} - \beta_{jk})^T$, i.e., changing the subset of entries of θ which are part of β, asymptotic normality can be proved for all entries of $\theta_{\mathcal{C}}$.

B.2. $\lim_{n\to\infty}\mathbb{P}(\mathcal{C}_n = \mathcal{C}) = 1$. To show consistency it has to be proved that $\lim_{n\to\infty}\mathbb{P}(\mathcal{J} \in \mathcal{C}_n) = 1$ if $\mathcal{J} \in \mathcal{C}$ and that $\lim_{n\to\infty}\mathbb{P}(\mathcal{J} \in \mathcal{C}_n) = 0$ if $\mathcal{J} \notin \mathcal{C}$, where $\mathcal{J}$ denotes a triple of indices (j, s, r) or pair (j, r).

B.2.1. *$lim_{n\to\infty}\mathbb{P}(\mathcal{J} \in \mathcal{C}_n) = 1$ if $\mathcal{J} \in \mathcal{C}$.* follows from part (a).

B.2.2. *$lim_{n\to\infty}\mathbb{P}(\mathcal{J} \in \mathcal{C}_n) = 0$ if $\mathcal{J} \notin \mathcal{C}$.* A similar proof is found in Bondell and Reich (2009). Let $\mathcal{B}_n$ denote the (nonempty) set of indices $\mathcal{J}$ which 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 $\hat{\beta}_{lq} > 0$, $l \geq 0$. 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.
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 $\hat{\beta} = \operatorname{argmin}_\beta \Psi(\beta) = \operatorname{argmin}_\beta \; -l(\beta) + \frac{\lambda_n}{\sqrt{n}}J(\beta)$ like defined in (a) is equivalent to

$$\operatorname{argmin}_{\mathfrak{B}} \; -l_n(\beta) + \lambda_n \sum_j J_j(\beta)$$

with

$$\begin{aligned}
\mathfrak{B} = \{\beta : {} & \beta_{0,1}, \ldots, \beta_{l-1,k}, \beta_{l,1} \leq \ldots \\
& \leq \beta_{l,z} \leq 0 \leq \beta_{l,z+1} \leq \ldots \leq \beta_{l,k}, \beta_{l+1,1}, \ldots, \beta_{p,k}\}, \\
J_j(\beta) = {} & \sum_{r>s} \frac{\phi_{rs(j)}(n)}{|\hat{\beta}^{ML}_{jr} - \hat{\beta}^{ML}_{js}|} \, |\beta_{jr} - \beta_{js}| + I(j \neq 0) \sum_{r=1}^{k} \frac{\phi_{r(j)}(n)}{|\hat{\beta}^{ML}_{jr}|} \, |\beta_{jr}| \, , \; j \neq l, \\
J_l(\beta) = {} & \sum_{r>s} \frac{\phi_{rs(l)}(n)}{|\hat{\beta}^{ML}_{lr} - \hat{\beta}^{ML}_{ls}|} \, (\beta_{lr} - \beta_{ls}) + \sum_{r \geq z+1} \frac{\phi_{r(l)}(n)}{|\hat{\beta}^{ML}_{lr}|} \, (\beta_{lr}) \\
& - \sum_{r \leq z} \frac{\phi_{r(l)}(n)}{|\hat{\beta}^{ML}_{lr}|} \, (\beta_{lr}) \, .
\end{aligned}$$

Since $\hat{\beta}^n_{lq} \neq 0$ is assumed, at the solution $\hat{\beta}^n$ 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 $\mathcal{C}_n$ can be omitted, since they will vanish in $J(\hat{\beta}^n) = \sum_j J_j(\hat{\beta}^n)$. If $x_{(l)q}$ denotes the column of design matrix Z which belongs to β_{lq}, due to differentiability, estimate $\hat{\beta}^n$ must satisfy

$$\frac{s_n(\beta)}{\sqrt{n}} = \frac{x_{(l)q}^T D_n(\beta)\Sigma_n^{-1}(\beta)(y-\mu)}{\sqrt{n}} = A_n + D_n,$$

with

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

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

From part (a) we know that $n^{-1/2}s_n(\beta) \overset{d}{\to} N(0, F(\beta))$. Hence for any $\epsilon > 0$, we have

$$\lim_{n\to\infty}\mathbb{P}(\frac{s_n(\beta)}{\sqrt{n}} \le \lambda_n^{1/4} - \epsilon) = 1$$

Since $\lambda_n/\sqrt{n} \to 0$, we also know $\exists\epsilon > 0$ such that $\lim_{n\to\infty}\mathbb{P}(|A_n| < \epsilon) = 1$. By assumption $\lambda_n \to \infty$; due to consistency of the ordinary ML-estimate $(\mathcal{O}(n^{-1/2}))$, we know that

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

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

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

As a consequence

$$\lim_{n\to\infty}\mathbb{P}(\frac{s_n(\beta)}{\sqrt{n}} = A_n + D_n) = 0.$$

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

$$\lim_{n\to\infty}\mathbb{P}(\mathcal{J} \in \mathcal{C}) = 0.$$

ACKNOWLEDGEMENTS

Special thanks go to Professor Nils Lid Hjort (Department of Mathematics, University of Oslo) for his precise and helpful comments at the very beginning of this project!

REFERENCES

AITKIN, M. (1999). A General Maximum Likelihood Analysis of Variance Components in Generalized Linear Models. *Biometrics* **55** 117–128. MR1705676

BONDELL, H. D. and REICH, B. J. (2009). Simultaneous Factor Selection and Collapsing Levels in ANOVA. *Biometrics* **65** 169–177. MR2665858

BOULESTEIX, A.-L. (2006). Maximally Selected Chi-square Statistics for Ordinal Variables. *Biom. J.* **48** 451–462. MR2240092

CLAESKENS, G. and HJORT, N. L. (2008). Minimizing Average Risk in Regression Models. *Econometric Theory* **24** 493–527. MR2422864

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM-Algorithm. *R. Stat. Soc. Ser. B Stat. Methodol.* **39** 1–38. MR0501537

FAHRMEIR, L. and KAUFMANN, H. (1985). Consistency and Asymptotic Normality of the Maximum Likelihood Estimator in Generalized Liner Models. *Ann. Statist.* **13** 342–368. MR0773172

FAHRMEIR, L. and TUTZ, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer Verlag, New York. MR1832899

FAN, J. and LI, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FAN, J. and ZHANG, W. (1999). Statistical Estimation in Varying Coefficient Models. *Ann. Statist.* **27** 1491–1518. MR1742497

GERTHEISS, J., STELZ, V. and TUTZ, G. (2012). Regularization and Model Selection with Categorical Covariates. In *Proc. of the Joint Conf. of the German Classif. Soc. and the German Assoc. for Pattern Recognit.* Accepted for publication.

GERTHEISS, J. and TUTZ, G. (2012). Regularization and Model Selection with Categorial Effect Modifiers. *Statist. Sinica*. To appear.

GRÜN, B. and LEISCH, F. (2008). FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters. *J. Stat. Software* **28** 1–35.

HASTIE, T. and TIBSHIRANI, R. (1993). Varying-coefficient Models. *R. Stat. Soc. Ser. B Stat. Methodol.* **55** 757–796. MR1229881

HOERL, A. E. and KENNARD, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12** 55–67.

HOFNER, B., HOTHORN, T. and KNEIB, T. (2008). Variable Selection and Model Choice in Structured Survival Models. *Department of Statistics at the University of Munich: Technical Reports* **43**.

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. MR1666699

KAUERMANN, G. and TUTZ, G. (2000). Local Likelihood Estimation in Varying-Coefficient Models including Additive Bias Correction. *Nonparametr. Stat.* **12** 343–371. MR1760712

LEISCH, F. (2004). FlexMix: A general framework for finite mixture models and latent class regression in R. *J. Stat. Software* **11** 1–18.

LENG, C. (2009). A simple Approach for Varying-coefficient Model Selection. *Statist. Plann. Inference* **139** 2138–2146. MR2507976

LIN, Y. and ZHANG, H. H. (2006). Component Selection and Smoothing in Multivariate Nonparametric Regression. *Ann. Statist.* **34** 2272–2297. MR2291500

LU, Y., ZHANG, R. and ZHU, L. (2008). Penalized Spline Estimation for Varying-Coefficient Models. *Comm. Statist. Theory Methods* **37** 2249–2261. MR2446666

McCullagh, P. (1983). Quasilikelihood Functions. *Ann. Statist.* **11** 59–67. MR0684863

Oelker, M.-R. (2012). gvcm.cat: Regularized categorial effects/categorial effect modifiers in GLMs R package version 1.4.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the LASSO. *R. Stat. Soc. Ser. B Stat. Methodol.* **58** 267–288. MR1379242

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and Smoothness via the fused LASSO. *R. Stat. Soc. Ser. B Stat. Methodol.* **67** 91–108. MR2136641

Tutz, G. and Schauberger, G. (2010). catdata: Categorial and Count Data R package version 1.1.

Ulbricht, J. (2010). *Variable Selection in Generalized Linear Models*. Verlag Dr. Hut, Dissertation, Department of Statistics, Ludwig-Maximilians-Universität München.

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** 1556–1569. MR2504204

Wang, H. and Xia, Y. (2009). Shrinkage Estimation of the Varying Coefficient Model. *J. Amer. Statist. Assoc.* **104** 747–757. MR2541592

Wu, C. O., Chiang, C.-T. and Hoover, D. R. (1998). Asymptotic Confidence Regions for Kernel Smoothing of a Varying-Coefficient Model with Longitudinal Data. *Amer. Statist. Assoc.* **93** 1388–1389. MR1666635

Yuan, M. and Lin, Y. (2006). Model Selection and Estimation in Regression with Grouped Variables. *R. Stat. Soc. Ser. B Stat. Methodol.* **68** 49–67. MR2212574

Zou, H. (2006). The Adaptive LASSO and its Oracle Properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

Department of Statistics
Ludwig-Maximilians-Universität Munich
Adademiestr. 1
80799 Munich, Germany
E-mail: margret.oelker@stat.uni-muenchen.de
jan.gertheiss@stat.uni-muenchen.de
gerhard.tutz@stat.uni-muenchen.de