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 Categorial Predictors and Effect Modifiers in Generalized Linear Models

Margret-Ruth Oelker*†, Jan Gertheiss† & Gerhard Tutz†

March 6, 2012

Abstract

We consider varying-coefficient models with categorial effect modifiers in 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 the estimators' large sample properties, and show in simulation studies that the proposed approaches perform very well for finite samples, too. Furthermore, the presented methods are compared with alternative procedures, and applied to real-world medical data.

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

1 Introduction

In regression modeling categorial predictors, also called factors, are a standard case. Nevertheless variable selection for categorial predictors and the connected problem which categories are to be distinguished has been somewhat neglected. We want to address these problems in a slightly extended version of generalized linear models (GLMs), namely GLMs with varying coefficients.

Varying coefficients (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 smoothly with the value of other variables u_j – the so called effect modifiers. The linear predictor has the form

$$\eta = \beta_0(u_0) + x_1\beta_1(u_1) + \ldots + x_p\beta_p(u_p), \tag{1}$$

*Corresponding author: margret.oelker@stat.uni-muenchen.de

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

where $x_1, x_2, ..., x_p$ are continuous covariates, $u_1, ..., u_p$ are effect modifiers and the functions β_j are unknown. As in GLMs the predictor is linear in the regressors, but scalar coefficients β_j turn into functions depending on the effect modifiers u_j, $j = 0, \ldots, p$. The effect modifiers can but do not have to represent the same variable. However, in order to keep notation at the same time as general and as intuitive as possible, we will always write $x_j\beta(u_j)$. As in GLMs we assume that the predictor η is linked to the conditional mean of response vector y by a knwon response function h, that is, $\mu = \mathbb{E}(y|x_1, \ldots, x_p) = h(\eta)$ and y follows a simple exponential family.

For continuous effect modifiers, unknown functions $\beta_j(\cdot)$ are smooth and have been modeled by splines (Hastie and Tibshirani, 1993; Hoover et al., 1998; Lu et al., 2008), local techniques (Wu et al., 1998; Fan and Zhang, 1999; Kauermann and Tutz, 2000) or boosting (Hofner et al., 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. Wang et al. (2008) obtain selection of whole splines by SCAD-penalization, while Wang and Xia (2009) select covariates by local polynomial regression with the grouped Lasso (Yuan and Lin, 2006). However, apart from Hofner et al. (2008) selection of predictors and specification of smooth/constant functions is not reached simultaneously.

For categorial effect modifiers $u_j \in \{1, ..., k_j\}$, which are considered here, "function" $\beta_j(u_j)$ has the form $\sum_{r=1}^{k_j} \beta_{jr} I(u_j = r)$, where $I(.)$ denotes the indicator function and $\beta_{j1}, \ldots, \beta_{jk_j}$ represent 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 $\beta_j^T = (\beta_{j1}, \ldots, \beta_{jk_j})$ contains the parameters for the jth predictor. With categorial predictors the number of parameters $q = \sum_{j=0}^{p} k_j$ can become very large, even for a moderate number of predictors. Consequently maximum likelihood (ML-)estimates may not exist and regularization techniques are needed. Even if estimates exist, one wants to reduce the model to the relevant terms. That means, one wants to determine which predictors are influential, and if they are influential, which categories have to be distinguished.

Regularization methods that enforce selection of predictors and fusion of categories have been considered by Gertheiss and Tutz (2012). However, they treat the case of Gaussian responses only; computational methods and derived asymptotics are limited to Gaussian responses. In this paper, two approaches are presented that allow to fit categorial effect modifiers within the GLM framework. In Section 2 we propose a penalized ML criterion for estimation. For computation a penalized iteratively reweighted least squares algorithm is employed. Moreover, large sample properties are derived. As an alternative a forward selection procedure with information criteria is shortly sketched (Section 3). In Section 4, the proposed methods are shown to be highly competitive in numerical experiments - whereby the penalized approach performs definitely better. Finally, the approaches are applied to real-world data (Section 5).

2 Penalized Estimation

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

$$\mathcal{M}_n^{pen}(\beta) = -l_n(\beta) + P_\lambda(\beta) = -l_n(\beta) + \lambda \cdot J_n(\beta), \tag{2}$$

where $l_n(\beta)$ denotes the log-likelihood for sample size n, and $P_\lambda(\beta)$ stands for a general penalty that depends 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. Without penalty $P_\lambda(\beta)$, that is with $\lambda = 0$, the ordinary ML-estimate is obtained.
The main issue is to chose an adequate penalty $J_n(\beta)$: The Ridge penalty (Hoerl and Kennard, 1970) shrinks coefficients, the Lasso (Tibshirani, 1996) combines shrinkage and selection of coefficients. The fused Lasso (Tibshirani et al., 2005) applies the Lasso to differences. Adjacent parameters are shrunk towards each other and are fused in order to gain a local consistent profile of ordered coefficients. In contrast, the grouped Lasso (Yuan and Lin, 2006) selects whole blocks of coefficients simultaneously. Although variable selection is implied, both the Lasso and its grouped version are off target since it does not enforce $\beta_{jr} = \beta_{js}$ for some $r \neq s$ in coefficient vector $\beta_j = (\beta_{j1}, \ldots, \beta_{jk_j})^T$, which is required for potentially (piecewise) constant functions $\beta_j(u_j)$. 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.
In predictor (1) it is not distinguished between nominal and ordinal effect modifiers. To use the information in the variable adequately, these cases should be distinguished. Therefore, we consider the general penalty

$$J_n(\beta) = \sum_{j=0}^{p} J_j(\beta_j), \tag{3}$$

where $J_j(\beta_j) = 0$ if covariate j is not modified, $J_j(\beta_j)$ is a nominal penalty term $J_j^{nom}(\beta_j)$, if effect modifier j is nominal, and an ordinal penalty term $J_j^{ord}(\beta_j)$, if effect modifier j is ordinal. For a *nominal* effect modifier u_j we propose

$$J_j^{nom}(\beta_j) = \sum_{r>s} |\beta_{jr} - \beta_{js}| + b_j \sum_{r=1}^{k_j} |\beta_{jr}|, \tag{4}$$

where b_j is an indicator that (de-)activates the second sum if wanted. Penalty (4) 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 (4) conforms to a Lasso penalty shrinking all coefficients belonging to $\beta_j(u_j)$ individually towards 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. For the intercept we will use only the first penalty term because in most cases shrinking towards zero is not requested; hence, we typically have $b_0 = 0$.

If u_j is *ordinal*, one wants to use this information in the predictor. Our option is to allow for the fusion of adjacent categories β_{jr} and $\beta_{j,r-1}$. Hence, for ordinal predictors we use

$$J_j^{ord}(\beta_j) = \sum_{r=2}^{k_j} |\beta_{jr} - \beta_{j,r-1}| + b_j \sum_{r=1}^{k_j} |\beta_{jr}|, \tag{5}$$

where b_j denotes the same indicator as above. Instead of all pairwise differences now only neighbored differences of coefficients belonging to covariate j are penalized, which corresponds exactly to a fused Lasso-type penalty (Tibshirani et al., 2005). Again, with setting b_0 to zero, the intercept can be treated separately.
Apart from the different information contained in the categories, J_j^{nom} and J_j^{ord} work in a similar way: 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.
It may be advantageous to use weights for the two components of the penalty (compare Tibshirani et al., 2005). With parameter $\psi \in (0, 1)$ let the weighted penalty for effect modifier j be given by

$$J_j^{nom}(\beta, \psi) = \psi \sum_{r>s} |\beta_{jr} - \beta_{js}| + (1 - \psi) b_j \sum_{r=1}^{k_j} |\beta_{jr}|, \tag{6}$$

for ordinal effect modifiers by

$$J_j^{ord}(\beta, \psi) = \psi \sum_{r=2}^{k_j} |\beta_{jr} - \beta_{j,r-1}| + (1 - \psi) b_j \sum_{r=1}^{k_j} |\beta_{jr}|. \tag{7}$$

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. In simulation studies in Section 4, we will compare the performance of a penalty with flexible parameter ψ to a "fixed" version with $\psi = 0.5$.

2.1 Computational Issues

Since penalty (3) contains absolute values, a convex but not continuously differentiable optimization problem has to be solved. Convenient optimization methods like Newton-type algorithms (for example Fisher scoring) or Nelder-Mead methods using derivatives cannot be applied. However, non-differentiability can be evaded by approximating the penalty at the critical points, i.e. in a neighborhood of $|\xi|$, $\xi = 0$: We will use the augmented local quadratic approximation (LQA)-algorithm (Fan and Li, 2001, Ulbricht, 2010), which employs a quadratic function for approximating the absolute value. Hence, the approximate optimization problem is differentiable again and a modified version of an iteratively reweighted least squares algorithm can be derived. An alternative are convenient optimization algorithms like function `nlm` (R Development Core Team, 2009), which could be applied directly to the approximated optimization problem. Since results are more stable we focus on the LQA-algorithm.
The LQA-algorithm is an iteratively reweighted least squares algorithm. It assumes penalties

that can be written in the form

$$P_\lambda(\beta) = \sum_{l=1}^{L} p_{\lambda,l}(|a_l^T\beta|), \tag{8}$$

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 $\frac{\mathrm{d}_{p_{\lambda,l}(|a_l^T\beta|)}}{\mathrm{d}_{|a_l^T\beta|}} \geq 0\ \forall\ a_l^T\beta > 0$ holds. Approximating absolute values in penalty (8) by $|\xi| \approx \sqrt{\xi^2 + c}$, where c is a small positive real integer, 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.
Penalty $J_n(\beta)$ from equation (3) can be rewritten such that it fulfills the demands of the LQA-algorithm. Let the vectors a_l denote the columns of a block-diagonal matrix $A = \mathrm{diag}(A_0, \ldots, A_p) \subset \mathbb{R}^{q\times L}$ and functions $p_{\lambda,l}(\nu)$ be defined as $\lambda \cdot \nu$. Let the block A_j refer to the effect modifier u_j. If u_j is nominal, expressions $a_l^T\beta$ contain the absolute values of all coefficients $\beta_{j1}, \ldots, \beta_{jk_j}$ and all possible differences. The former is reached when employing the columns of a $(k_j \times k_j)$ identity matrix, the latter by columns containing these combinations of one and minus one building the needed differences. Hence, e.g. for $k_j = 4$, we have

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

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 $\beta_{j1}, \ldots, \beta_{jk_j}$ are penalized. Thus matrix A_j^{nom} is reduced to the $(k_j \times (2k_j - 1))$ matrix

$$A_j^{ord} = \begin{pmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}.$$

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 j, whose influence on y is not modified by any u_j, matrix A_j^{none} is an empty matrix with zero columns and as many rows as coefficients belonging to covariate j.
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 objective function is strictly convex, a local optimum will be also the global optimum. 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 can happen that the solution is not unique (Ulbricht, 2010).

2.2 Large Sample Properties

In this section, a modified version of the proposed estimate is investigated. It is shown to be consistent in terms of variable selection and identification of relevant differences $\beta_{jr} - \beta_{js}$. For

asymptotics, general assumptions have to hold and the number of observations has to grow in accordance with the requirements of categorial 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. This is necessary to ensure that the ML-estimate is consistent – which we assume, too. Then, estimate $\hat{\beta}$ solving equation (2) with penalty (3) and fixed tuning parameter λ is consistent in terms of $\lim_{n\to\infty}\mathbb{P}(||\hat{\beta} - \beta^*||^2 > \epsilon) = 0$ for all $\epsilon > 0$, where β^* stands for the vector of true coefficients. This behavior is formally described by

Proposition 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) with $J_n(\beta)$ defined by (3), (4) and (5) 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 the Appendix. Employing the generalized versions (6) and (7) 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 (3) is available: Given effect modifiers u_j, $j = 1, \ldots, p$, penalty $J_n(\beta)$ (3) is modified to the adaptive penalty $J_n^{ad}(\beta)$ by employing

$$J_j^{ad,nom}(\beta) = \sum_{r>s} w_{rs(j)}|\beta_{jr} - \beta_{js}| + b_j \sum_{r=1}^{k_j} w_{r(j)}|\beta_{jr}| \tag{9}$$

and

$$J_j^{ad,ord}(\beta) = \sum_{r=2}^{k_j} w_{rs(j)}|\beta_{jr} - \beta_{j,r-1}| + b_j \sum_{r=1}^{k_j} w_{r(j)}|\beta_{jr}|, \tag{10}$$

which replace (4) and (5), and by using adaptive weights

$$w_{rs(j)} = \phi_{rs(j)}(n)|\hat{\beta}_{jr}^{ML} - \hat{\beta}_{js}^{ML}|^{-1} \tag{11}$$

and

$$w_{r(j)} = \phi_{r(j)}(n)|\hat{\beta}_{jr}^{ML}|^{-1}. \tag{12}$$

Let $\hat{\beta}_{jr}^{ML}$ denote the ML-estimate of β_{jr}. For the functions $\phi_{rs(j)}(n)$ and $\phi_{r(j)}(n)$ only convergence to fixed values is assumed, that is, $\phi_{rs(j)}(n) \to q_{rs(j)}$, $\phi_{r(j)}(n) \to q_{r(j)}$, respectively, with $0 < q_{rs(j)}, q_{r(j)} < \infty$. With $\phi_{rs(j)}(n)$ and $\phi_{r(j)}(n)$ being positive constants that sum up to one, we obtain a generalization as given in equations (6) and (7); tuning parameter λ and functions $\phi_{rs(j)}(n)$, $\phi_{r(j)}(n)$ are clearly separated.
In contrast to Proposition 1, the penalty parameter λ is not fixed, but increases with sample size n, that is, one assumes that $\lambda = \lambda_n$ 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 $0 < c_r < 1$.
In addition, we define vector $\theta = A^T\beta$. Hence, θ is a vector that contains all terms that penalty $J_n(\beta)$ (3) 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 indexes corresponding to

those entries of θ^T 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 true vector of θ-entries included in $\mathcal{C}$, $\hat{\theta}^n_{\mathcal{C}}$ is the corresponding estimate based on $\hat{\beta}^n$. $\mathcal{B}_n$ defines the (nonempty) set of indices $\mathcal{J}$, which are in $\mathcal{C}_n$ but not in $\mathcal{C}$.
Previous assumptions concerning ML-estimation are extended: the model holds and the negative log-likelihood $-l_n(\beta)$ is convex. For the sake of asymptotic normality and consistency $l_n(\beta)$ 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, and for score function $s(\beta)$ we suppose $\mathbb{E}(s(\beta)) = 0$. Then one obtains:

Proposition 2 *Suppose $\lambda = \lambda_n$ with $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n \to \infty$, and all class-wise sample sizes n_{jr} satisfy $n_{jr}/n \to c_{jr}$, where $0 < c_{jr} < 1$. Then penalty $J_n^{ad}(\beta)$ employing terms (9) and (10) with weights (11) and (12), where $\hat{\beta}^{ML}_{jr}$, $\phi_{rs(j)}(n)$ and $\phi_{r(j)}(n)$ are defined as above, ensures that*

(a) $\sqrt{n}(\hat{\theta}^n_{\mathcal{C}} - \theta^*_{\mathcal{C}}) \xrightarrow{d} N(0, Cov(\theta^*_{\mathcal{C}}))$

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

The proof is given in the Appendix and uses ideas of Zou (2006) and Bondell and Reich (2009). Its argumentation follows closely Gertheiss and Tutz (2012). 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 of parameters. Since all estimated differences are (deterministic) linear functions of estimated parameters, the covariance-matrix $\text{Cov}(\theta^*_{\mathcal{C}})$ is singular.
The assumptions $F_n/n \overset{n\to\infty}{\to} F$ 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 as, for example, in Fahrmeir and Kaufmann (1985).
For $\lambda = 0$ the unpenalized likelihood is maximized and therefore for $n \to \infty$ 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 $\lambda J_n(\beta)$ shall not vanish in proportion to $-l_n(\beta)$ for $n \to \infty$. Therefore $\lambda = \lambda_n$ with $\lambda_n \to \infty$ is requested. $\lambda_n/\sqrt{n} \to 0$ ensures an appropriate proportion of likelihood and penalty.
The speed of convergence of the normality part of Proposition 2 is determined by $\lambda_n/\sqrt{n} \to 0$. Since $n^{-1/2}s_n(\beta) \sim N(0, F(\beta)) + \mathcal{O}(n^{-1/2})$ and $\mathbb{P}(\sqrt{n}|\hat{\beta}^{ML}_{lq}| \leq \lambda_n^{1/2}) \to 1$ like $c/\sqrt{n} \to 0$, part (b) of Proposition 2 behaves the same. Thus the overall speed of convergence is $\mathcal{O}(n^{-1/2})$.
Since the penalized model employed in Proposition 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, 2008a) it is possible to replace φ by $\hat{\varphi}$. This leads to quasi likelihood functions:
The used arguments remain valid when expanding the considered model-class to quasi likelihood models. Only the estimates' covariance matrix can not 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.
Even for misspecified models estimate $\hat{\beta}^n$ remains asymptotic normal and consistent but converges towards "the last false parameter value" that minimizes the Kullback-Leibler distance

$$\text{KL}(g, f(\cdot, \beta)) = \int g(y)\log\frac{g(y)}{f(y,\beta)}\text{d}y,$$

where $g(y)$ denotes the "true", data-generating process and $f(y, \beta)$ presents the assumed model. Corresponding arguments can be found in Claeskens and Hjort (2008b, p. 26/27). In this case the estimate's covariance matrix is $J^{-1}KJ^{-1}$, with $J = -\mathbb{E}_g \frac{\partial^2 \log f(y,\beta)}{\partial\beta\partial\beta^T}$ and $K = \text{cov}_g \frac{\partial \log f(y,\beta)}{\partial\beta}$. If adaptive weights are used and refitting is applied after the identification of clusters and relevant variables, asymptotic behavior is obtained which is comparable to Proposition 2. Since clustering and variable selection are directly based on the penalty with adaptive weights, part (b) of this proposition is still valid. Asymptotic normality results from asymptotic normality of the ML-refit.

3 Alternative Selection Strategies

A more traditional way of model choice is based on information criteria like the AIC or the BIC. Several forward/backward selection strategies for a wide range of models have been proposed. The basic idea is to select that model that performs the best with respect to a specified criterion. However, by construction these strategies result in variable selection only – a coefficient is either selected or excluded. If coefficients are to be fused, strategies need to be modified.

Fusion of categories is enabled when the set of coefficients regarded is enlarged: Assuming a nominal effect modifier u_j with three categories having impact on covariate j, varying coefficient $\beta_j(u_j)$ corresponds to $(\beta_{j1}, \beta_{j2}, \beta_{j3})^T$ in coefficient vector β. All possible combinations of coefficients belonging to 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. In settings like this, one can use a forward selection strategy employing information criteria AIC and BIC.

Starting 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. Thereby one degree of freedom is defined as the number of non-zero coefficient blocks in $\hat{\beta}$ (Tibshirani et al., 2005). In other words, clusters of one or more non-zero and equal coefficients belonging to the same covariate j are counted as one degree of freedom. In each step a former zero coefficient can be set to non-zero, 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, identical coefficients. Picking up $\beta_j(u_j) = (\beta_{j1}, \beta_{j2}, \beta_{j3})^T$ from above and assuming $\beta_j(u_j)$ to be $\beta_{j1} = \beta_{j2} = \beta_{j3}$ in the actual iteration, in the next iteration $\beta_j(u_j)$ is either split into one of $\{(\beta_{j1}, \beta_{j2} = \beta_{j3})$, $(\beta_{j2}, \beta_{j1} = \beta_{j3})$, $(\beta_{j3}, \beta_{j2} = \beta_{j1})\}$ or one of the further covariates is changed. Concretely, the selection strategy is:

1. Start with a null model containing only a non-varying intercept, and all non-selectable covariates respectively.

2. In the following steps: Increase the degrees of freedom by one and check all possible models which are based on the model from the previous step. That means, for $j = 1, \ldots, p$, a group of former zero coefficients from $\beta_{j1}, ..., \beta_{jk}$ may become nonzero, alternatively a former cluster of nonzero coefficients may be split. Select the model with minimum AIC/BIC.

3. Stop if the (minimum) AIC/BIC does not decrease anymore.

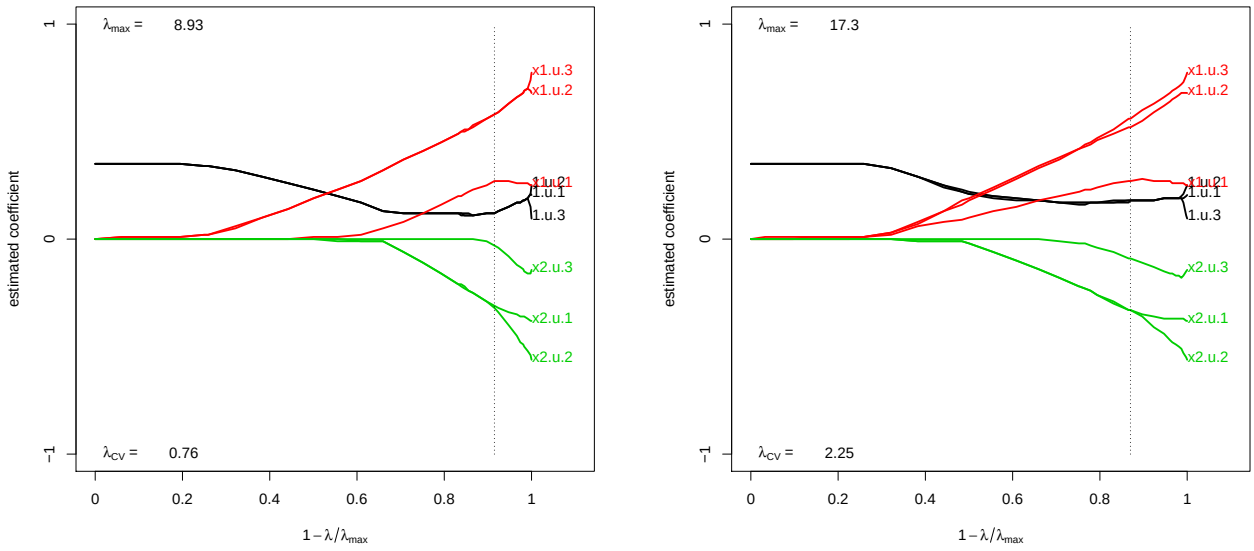

Figure 1: Coefficient paths for binary model (13) assuming predictor (14) – with adaptive weights (left) and the standard penalty (right).

4 Numerical Experiments

The proposed methods are compared in simulation studies in terms of prediction accuracy, selection and clustering performance. For illustration we start with a simple example.

4.1 An illustrative example

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

$$\begin{array}{rcccccccccccc} \eta_{true} & = & \beta_0 & + & x_1\beta_1(u) & & & & & & & + & x_2\beta_2 \\ & = & \beta_0 & + & x_1\,(\,\beta_{11}I(u=1) & + & \beta_{12}I(u=2) & + & \beta_{13}I(u=3)\,) & + & x_2\beta_2 \\ & = & 0.2 & + & x_1\,(\,0.3I(u=1) & + & 0.7I(u=2) & + & 0.7I(u=3)\,) & + & x_2\cdot -0.5 \end{array} \tag{13}$$

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. We generate $n = 400$ observations. Covariates x_1 and x_2 are independently drawn from an uniform distribution $U(0, 2)$, the effect modifier u from a multinomial distribution with probabilities 0.3, 0.4, 0.3 for categories 1, 2 and 3. The response is binary and we assume a logistic regression model with natural link function. When modeling, all coefficients are allowed to vary with effect modifier u, i.e., we have

$$\eta_{model} = \beta_0(u) + x_1 \cdot \beta_1(u) + x_2 \cdot \beta_2(u). \tag{14}$$

Figure 1 shows the resulting coefficient paths for penalized estimation and subject to penalty parameter λ. In the left panel, the penalty is adaptive, weights are fixed (see equation (9) with $b_0 = 0$, $\phi_{rs(j)} = \phi_{r(j)} = 0.5$). λ is scaled as $1 - \lambda/\lambda_{max}$, where λ_{max} refers to the smallest

Response	Number of Covariates	Number of Noise Variables	Number of Observations $n = 200$	$n = 600$
Binomial	2	2	b22.200	b22.600
		6	b26.200	b26.600
	6	2	b62.200	b62.600
		6	b66.200	b66.600
	10	2	b102.200	b102.600
		6	b106.200	b106.600
Poisson	2	2	p22.200	p22.600
		6	p26.200	p26.600
	6	2	p62.200	p62.600
		6	p66.200	p66.600
	10	2	p102.200	p102.600
		6	p106.200	p106.600

Table 1: Overview on simulation settings: without prior knowledge on the coefficients' structure the type of response and the number of covariates/added noise variables/observations is systematically varied.

value of penalty parameter λ that already gives maximal penalization, i.e., the smallest λ that sets all penalized coefficients to zero (up to a certain accuracy – here, to two digits). Hence, in Figure 1 the ML-estimate is seen at the right end. The left end relates to maximal penalization, here only the intercept remains non-zero. Black curves correspond to the intercept's coefficients, red ones to truly varying covariate x_1, green ones to covariate x_2. The paths show how clustering/selection of coefficients is done subject to penalty parameter λ: 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 doted line marks the optimal model in terms of 5-fold-cross-validation with the predictive deviance $Dev(y, \hat{\mu})$ as loss function ($\lambda_{CV} = 0.76$). It shrinks coefficients slightly – in return all but one relevant structures are identified. Absolute deviation to the true coefficients is small.
When the standard penalty (4) is used instead, results change: while coefficients 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 – the value of λ yielding maximal penalization is roughly doubled: $\lambda_{max} = 17.3$ (was 8.93 before). Consequently cross-validated λ_{CV} is 2.25 now. More importantly, 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; coefficients for categories 1, 2 and 3 of effect modifier u remain autonomous.

4.2 Simulation Settings

For further investigations we extend the illustrative example. Various model features are systematically varied – such that the proposed penalty with all possible weights can be compared to the ML-estimate and model selection via AIC/BIC in different situations. Concretely we consider binomial and Poisson response, the number of influential covariates is either 2, 6 or 10, we add either 2 or 6 non-influential noise variables. Training data sets contain $n = 200$ and

Method	Two Noise Variables			
	$n = 200$ (b26.200)		$n = 600$ (b26.600)	
	MSE	MSEP	MSE	MSEP
ML	0.76 (2.23)	742.5 (195.4)	0.09 (0.06)	1567 (72.4)
standard, ψ fixed	0.11 (0.03)	541.4 (33.8)	0.04 (0.01)	1502.4 (49.9)
standard, ψ flexible	0.12 (0.04)	541.5 (33.6)	0.05 (0.02)	1507.3 (49.3)
adaptive, ϕ fixed	0.49 (2.31)	601.6 (160.1)	0.03 (0.02)	1485.2 (56.4)
adaptive, ϕ flexible	0.59 (4.35)	596.4 (167.1)	0.03 (0.01)	1472.6 (53.2)
AIC	0.57 (21201.08)	720.6 (573.6)	0.06 (0.04)	1539.8 (64.3)
BIC	0.16 (1848.4)	554.2 (229.5)	0.04 (0.04)	1485.8 (55.3)

Table 2: Observed errors of parameter estimates (MSE) and predictions accuracy in terms of the deviance (MSEP) for settings b26.200 and b26.600. Estimated standard deviations of MSE and MSEP are given in parentheses.

$n = 600$ observations, test data sets $n = 600$, respectively $n = 1800$ observations (see Table 1). All covariates are continuous and independently drawn from an uniform distribution $U[-2, 2]$. All scenarios are characterized by a quite realistic assumption – namely that there is a known effect modifier. It is nominal, has four categories $1, \ldots, 4$ and is independently drawn from a multinomial distribution with probability 0.25 per category. However, we do not know which coefficients are varying. As in the illustrative, example we assume to have no prior knowledge about the coefficients' structure. In settings b26.200 and b26.600, for example, the true linear predictor is

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

Some coefficients are varying across all levels of u, some only partly, some are partly zero – that is, true coefficients are diversified. When fitting the model, however, all coefficients are allowed to vary with effect modifier u. Furthermore, we add six, non-influential noise variables that shall be detected, such that the assumed predictor is

$$\begin{aligned}
\eta_{model} &= \beta_0(u) + x_1 \cdot \beta_1(u) + x_2 \cdot \beta_2(u) + n_3 \cdot \beta_3(u) + n_4 \cdot \beta_4(u) + \\
&\quad n_5 \cdot \beta_5(u) + n_6 \cdot \beta_6(u) + n_7 \cdot \beta_7(u) + n_7 \cdot \beta_7(u).
\end{aligned}$$

This model is estimated with the different strategies for model selection that we 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 ML-estimate. For ML-estimates neither regularization nor model selection is required. They are the benchmark for all other estimates' performance. Penalty parameter λ is chosen by 5-fold cross-validation. If weights ψ and ϕ are flexible, they are cross-validated, too. For computation, the LQA-algorithm is employed. For each setting, all models are computed 50 times in order to make result reliable.

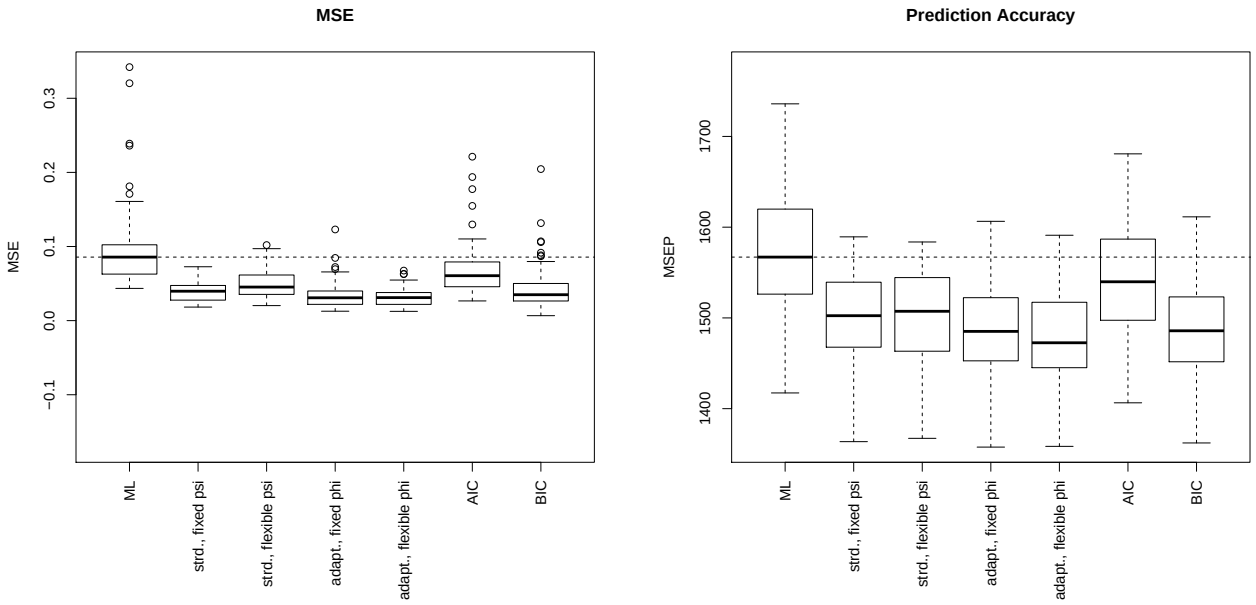

Figure 2: Boxplots of scaled squared errors and deviances for setting b26.600 (binomial response, 2 covariates, 6 noise variables, $n = 600$); medians mark estimates of MSE and MSEP.

4.3 Results

To evaluate the results we analyze estimation and prediction accuracies, and check whether the right coefficients are fused and/or selected, that is, the clustering and selection performance. To assess parameter estimation, we compute the coefficients' mean squared error for each simulation run:

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

where $q = \sum_{j=0}^{p} k_j$, β denotes the vector of true coefficients, $\beta = (\beta_0^T, \ldots, \beta_p^T)^T$, and $\hat{\beta}$ its estimate. In order to obtain stable estimates of the MSE we compute median values over all 50 simulations. To judge the prediction accuracy, the mean predictive deviance $Dev(y, \hat{\mu})$ is considered, referred to as MSEP. Again quantities of models $1, \ldots, 50$ are averaged by the median. To keep things simple, here, results are represented for settings b26.200 and b26.600 only. Table 2 lists MSE, MSEP and their standard deviations for all estimates of these two settings. It is seen that penalized approaches perform better than the ML-estimates. Considering their standard deviations points out how penalties stabilize estimation, while forward selection strategies suffer from immense variability. In addition, Table 2 shows how the standard and the adaptive penalty differ: for settings b26.200 and setting b26.600, standard deviations of the MSE of the standard approaches (standard, ψ fixed and flexible) are relatively small; while standard deviations of the MSE of the adaptive approaches (adaptive, ϕ fixed and flexible) are large when the number of observations is small ($n = 200$, left side in Table 2); in contrast, standard deviations are small for $n = 600$. This is due to the construction of the adaptive weights, which are the inverse of the ML-estimates. For few observations the ML-estimate is relatively bad and so are the adaptive weights. AIC/BIC based forward selection strategies without shrinkage effect perform also better than the ML-estimate, but for $n = 200$ they are very unstable

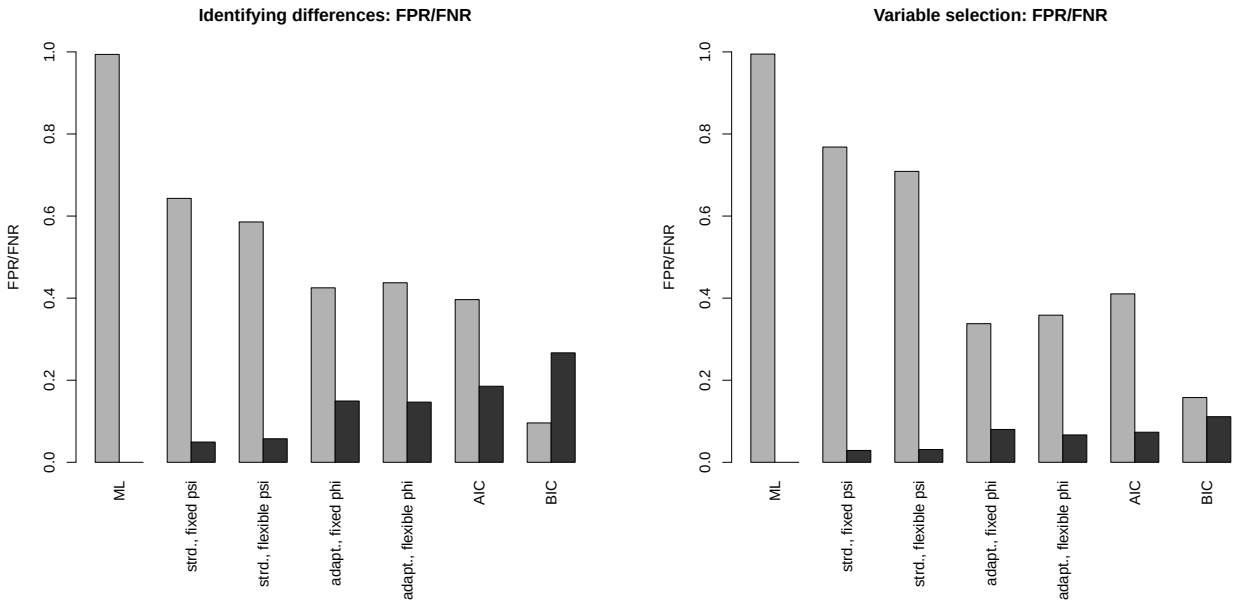

Figure 3: False positive rates (light gray, FPR) and false negative rates (dark gray, FNR) for setting b26.600 (binomial response, 2 covariates, 6 noise variables, $n = 600$); the left figure relates to clustering, the right side to selection performance.

(impressively documented by the observed standard deviations of the MSE, method "AIC"). Settings with more/less variables give approximately equivalent results: Adaptive penalization fails when ML-estimates are bad. On average, forward selection strategies produce similar results as penalized approaches but their variation is big, especially in prediction. If so, flexible weights (methods "standard, ψ flexible" and "adaptive, ϕ flexible") seem to increase performance only little and require two-dimensional cross-validation. Hence, for binary response, adaptive penalization with fixed weights is recommended when the sample size is large.
For count data results are very similar. The MSE and the MSEP of penalized models are smaller than those of the ML-estimate. Forward selection strategies suffer from a higher variability but the effect seems to be smaller. In contrast to binary response, adaptive approaches perform already better than standard penalties when $n = 200$. Hence, for Poisson-distributed response, adaptive weights can be recommended for smaller sample sizes. We advise against flexible 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. That is their differences should be set to zero.
To judge clustering and selection performance we consider false negative (FNR) and false positive rates (FPR). False positive means that a truly zero difference of coefficients belonging to the same predictor is fitted as non-zero, or that a truly zero coefficient is set to non-zero, respectively. False negative means that truly non-zero values are estimated to be zero. With # denoting "the number of" we have

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

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

$\text{FPR}_{clustering}$ and $\text{FNR}_{clustering}$ are defined analogously. If the denominator of one of these fractions equals zero the value is understood as missing. For ML-estimates false positive rates will always be one, false negative rates always zero.
Previous results favored penalized estimation with adaptive weights for exemplary setting b26.600. For less observations the standard penalty performed better. Figure 3 shows false positive and negative rates for setting b26.600. On the left rates for clustering, on the right rates for selection are shown. FNR are naturally quite low. Overall it stands out that forward selection strategies perform well. Apart from AIC and BIC, the approach with the adaptive penalty and fixed weights scores the best when considering clustering performance. Having the high variability of forward selection strategies in mind and looking at both clustering and selection, the previous recommendation (adaptive penalty, fixed weights) still holds. In this case stable estimation and a good clustering/selection performance are balanced. For count data former results are strongly supported, too: Employing an adaptive penalty with fixed weights decreases FPR by nearly 50%, for both $n = 200$ and $n = 600$. Rates are in the range of forward selection strategies while being stable. Overall for count data truly zero/non-zero coefficients and their differences are fairly well detected.

5 Application to Real-World Data

5.1 Reducing Mortality after Myocardial Infarction

In this first example 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 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;

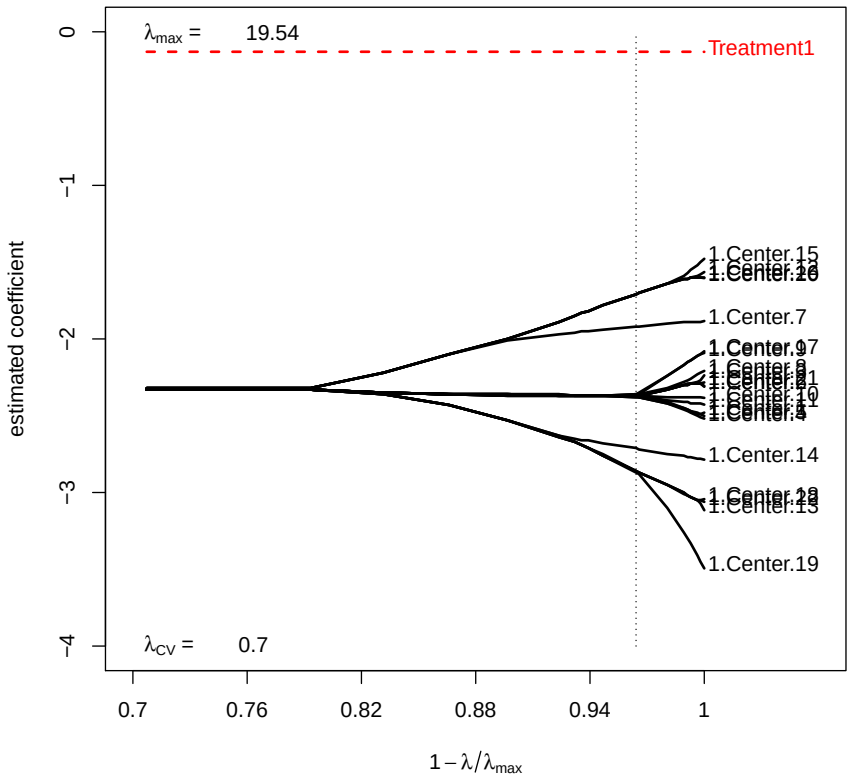

Figure 4: Coefficient paths beta-blocker data.

for each component a simple exponential family with parameters Ψ_m is assumed. For estimation an iterative EM-algorithm (Dempster et al., 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:

$$\eta_i = \beta_0(Center_i) + \beta_T \cdot Treatment_i,\ i = 1, \ldots, 44 \text{ Cases}. \tag{15}$$

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) with adaptive weights (11) and (12) 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 (15). 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

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.

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 (15) 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

$$\eta_i = \beta_0(Center_i) + \beta_T(Center_i) \cdot Treatment_i,\ i = 1, \ldots, 44 \text{ Cases} \tag{16}$$

and the same assumptions as above. As there is only one covariate and one effect modifier, which are both categorial, predictor (16) 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

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? (0: no, 1: yes)
memb	Did the membranes burst before the beginning of the throes? (0: no, 1: yes)
rest	Was a strict bed rest ordered to the mother for at least one month during the pregnancy? (0: no, 1: yes)
cephalic	Was the child in cephalic presentation before birth? (0: no, 1: yes)
t	Year of birth, effect modifier

Table 4: Short description of response, considered covariates and the effect modifier for birth data.

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 (15); such that the "fixed" treatment effect assumed in Aitkin (1999) and Grün and Leisch (2008) is supported.

5.2 Cesareans among Francophone Mothers

In a second example we analyze a data set presented by Boulesteix (2006). It contains various variables related to the pregnancy and delivery of 775 women recruited on French-speaking websites and is available in `R` add-on package `catdata` (Tutz and Schauberger, 2010). We are interested in the type of delivery, in whether birth was given vaginally or by means of a Cesarean. Cases were observed between 1983 and 2004, i.e., in a period of more than 20 years. In this period medical standards changed. Modeling the type of delivery requires to consider time, and even more important, it requires to consider how various aspects eventually developed over time. Due to the number of observations per year, we focus on the period from 2001 on and investigate a total of 603 deliveries by a varying coefficient model with ordinal effect modifier time t. The response is binary indicating the type of delivery; 0 stands for a vaginal birth, 1 for a Cesarean. The model considers in principal all covariates which were available and meaningful for all women. Covariates which are postnatal (e.g. "Days that the child spent in intensive care"), that refer to a subset of women only (e.g. "Was the realized Cesarean planned?") or that are observed barely (e.g. "Head circumference of child at birth"), are omitted. As their terms and delivery circumstances differ immensely, multiple births are excluded, too. For better interpretation we consider the womens' height and their body mass index (BMI) instead of their height and their weight. Details on all employed covariates are found in Table 4.

So far, we assumed continuous covariates $x_1, \ldots, x_p$. Here some covariates are binary and effect coded. However, as effect coded binary covariates result in one-dimensional covariates, the proposed penalty can be applied. As we have no prior knowledge about the model's structure, effect modifier t potentially impacts all coefficients. Compared to the first example, there are

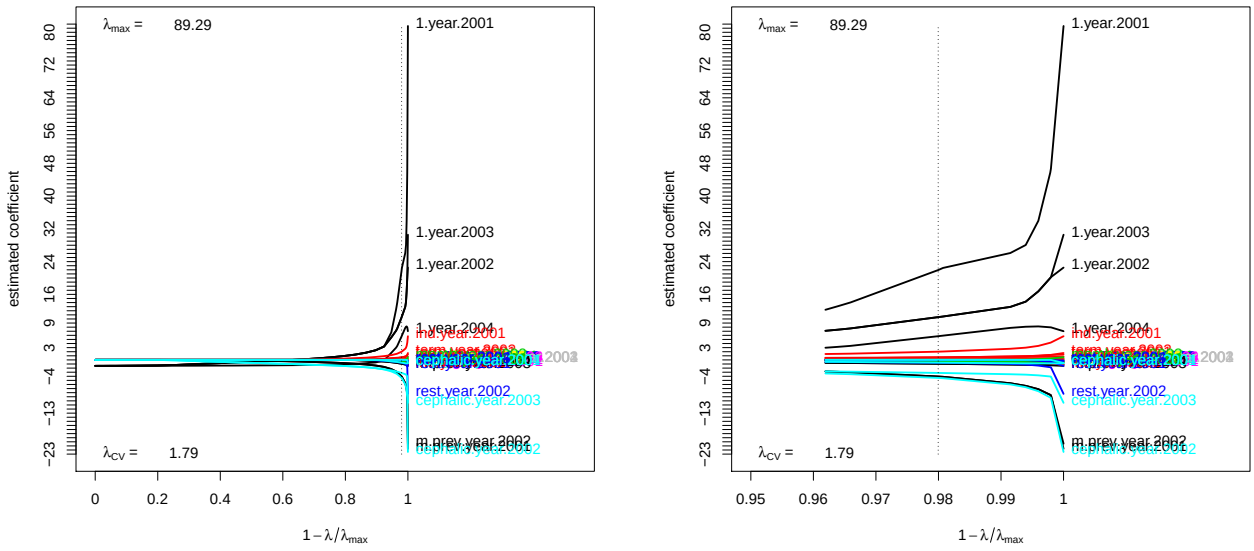

Figure 5: Coefficient path for the birth data when weights are adaptive and parameter ψ is fixed. All covariates listed in Table 4 are taken into account. All coefficients including the intercept are assumed to vary over time t. The dotted line marks the cross-validated choice of penalty parameter λ.

much more covariates. We are not only interested in the fusion of some parameters belonging to one varying coefficients $\beta_j(t)$, but as well in the selection of coefficients $\beta_j(t)$. Hence, we do not compare penalized estimation and cluster methods. Instead we consider the forward selection strategies presented in Section 3 as alternative options. Figure 5 shows the resulting coefficient path when weights are adaptive and parameter ψ is fixed. In the left panel we see the whole path, while the right panel focuses on that part where coefficients are fused and selected. The paths' very right end stands for $\lambda = 0$. In this case the range of coefficients is pretty large. Pure ML-estimation seems to be unstable. For instance, in 2001 we observe an intercept larger than 80. To obtain a stable model, regularization is required. Penalty parameter λ is cross-validated and set to 1.79. That is small compared to the minimal value of λ giving maximal penalization. But it stabilizes estimation and shrinks the unstable coefficients enormously. Table 5 gives the exact results for all considered methods. To keep things simple, excluded coefficients are omitted. Coefficients, that are found to be non-varying, that is to be only one constant, are represented by that constant coefficient only. We see that forward selection strategies give very sparse estimates. Only three (AIC), respectively one (BIC) coefficient are partly varying. ML-estimates argue for a strong dependency on time, see for example the intercept for the year 2001, but forward selection strategies ignore it. In contrast, penalized estimation selects more coefficients. All selected coefficients vary over time. For most coefficients we see a clear trend (for example $\beta_{cephalic}$). The time-varying intercept reflects the effect in year 2001 and how it shrinks over time. At the same time estimation is stable and trends of the other coefficients are clearly separated from this effect in 2001. Penalized estimation stabilizes estimation and takes the data's structure into account.

Coefficients	Penalized estimation				Forward Selection AIC				Forward Selection BIC			
	t				t				t			
	2001	2002	2003	2004	2001	2002	2003	2004	2001	2002	2003	2004
$\beta_0(t)$	21.94	10.45	10.45	5.79			14.78				11.53	
$\beta_{\text{term}}(t)$	0.58	0.58	-0.15	-0.15			-0.16					
$\beta_{\text{c.height}}(t)$												
$\beta_{\text{c.weight}}(t)$												
$\beta_{\text{m.age}}(t)$	-0.21	0.06	0.06	0.06			0.08					
$\beta_{\text{m.height}}(t)$	-0.33	-0.23	-0.02	-0.02			-0.07				-0.07	
$\beta_{\text{m.bmi}}(t)$		0.04	0.01									
$\beta_{\text{m.gain.w}}(t)$												
$\beta_{\text{m.prev}}(t)$	-4.06	-4.06	-0.94	-0.94			-1.31				-1.15	
$\beta_{\text{ind}}(t)$	2.01	-0.09	0.62	0.10	0.65	0.65	0.65	0.25			0.48	
$\beta_{\text{memb}}(t)$			-0.32	-0.33			-0.51	-0.51			-0.45	-0.45
$\beta_{\text{rest}}(t)$		-0.71		-0.71	-0.39	-0.39	-0.39					
$\beta_{\text{cephalic}}(t)$		-4.40	-3.18	-0.67			-1.51				-1.52	

Table 5: Estimates for all methods fitted to the birth data. Coefficients that are excluded are omitted. Coefficients, that are found to be non-varying, that is to be only one constant, are represented by that constant coefficient only.

6 Special Case: Categorial Effects

So far we considered categorial effect modifiers in general. We did not touch categorial effects, which are a special case of categorial effect modifiers. One obtains a coded categorial effect, when the effect modifier u_j is categorial 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 categorial 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 categorial effects. Note, however, the penalty we use here is not designed for a reference category. All categories of a categorial effect are penalized in the same way. For sufficiently strong penalization, all coefficients are set to zero. Hence, transformed coefficients are shrunken. In contrasts to Gertheiss et al. (2012), parts of the transformed intercept are based upon penalized coefficients. Apart from these details, there are no restrictions. Thus, large sample theory for penalized categorial effects is generalized to GLMs.

7 Summary and Discussion

We considered categorial effect modifiers in GLMs. By nature, categorial effect modifiers result in much more coefficients than in usual models. When selecting a model, one wants to know which covariates impact the response, and if so, how. To answer these questions, we propose two different approaches: One 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. The penalty adjusts for the different amount of information in nominal and ordinal effect modifiers. In accordance with Zou (2006), an adaptive version of the proposed penalty was shown to be asymptotically normal and consistent, with the speed of convergence being $\mathcal{O}(n^{-1/2})$. These results remain valid when scale parameter ϕ of the exponential family is estimated and plugged-in, which allows for quasi-likelihood approaches. Similar results can be derived for refitting procedures. On the other hand, we investigate 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. The degrees of freedom are defined as in Tibshirani et al. (2005). As in many other best subset selection strategies, we considered AIC and BIC as criteria. Numerical experiments suggested both proposed methods to be highly competitive. In systematically altered settings, we compared penalized approaches, forward selections strategies and the ordinary ML-estimate. Penalized estimates performed distinctly better than unpenalized ML-estimates. Forward selection strategies employing information criteria AIC or BIC challenge the proposed penalty. With the former approaches, estimation accuracy is partly better, and mostly more truly zero (differences) of coefficients were detected. However, forward selection strategies suffer from immense variability; such that they do not seem to be a real alternative.

Lasso-type penalties as employed here require to solve not continuously differentiable optimization problems. Fan and Li (2001) proposed a local quadratic approximation for such problems. Ulbricht (2010) specified a more concrete form for quite general penalties, the so called LQA-algorithm. We adopt this algorithm for the proposed penalty effectively. All functions will be available in `R` add-on package `gvcm.cat`.

In practice, varying coefficient models are highly relevant. 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. In addition Cesareans among francophone mothers were analyzed. We were interested in how the influence of various medical indicators changed over time. The data is quite challenging, standard approaches fail. However, penalized estimates give a coherent trend.

Especially in medicine time plays an important role as effect modifier – typically in longitudinal studies, where each case is monitored repeatedly. Hence, observations are no longer independent. To consider individual dependencies, the method's scope can be enlarged to marginal models (Liang and Zeger, 1986). Moreover, concepts to combine the proposed penalty with working correlation matrices can be developed. The proposed penalty can be further generalized, too: Varying coefficients can depend on more than one effect modifier. Already with two effect modifiers, the scope of functions $\beta_j(\cdot)$ widens enormously. Penalties and selection strategies have to be modified. In this paper we assumed continuous covariates $x_1, \ldots, x_p$. But of course covariates can be categorial, too. Then there are even more coefficients, there is an even stronger demand for regularization. In addition, this case corresponds to two categorial covariates and their interaction in a usual GLM. Hence, it is even more important to find strategies to fuse and/or select coefficients.

Acknowledgements

This work was partially supported by DFG project "Regularisierung für diskrete Datenstrukturen". Special thanks go to Professor Nils Lid Hjort (Department of Mathematics, University of Oslo) for his precise and helpful comments in the very beginning of this project!

Appendix

Proof of Proposition 1

If $\hat{\beta}$ minimizes $\mathcal{M}_n^{pen}(\beta)$ (2) with $J_n(\beta)$ as defined by (3), (4) and (5), 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 $\mathcal{M}_n^{pen}(\hat{\beta})/n \xrightarrow{\mathbb{P}} \mathcal{M}_n(\hat{\beta})/n$, $\mathcal{M}_n(\hat{\beta})/n \xrightarrow{\mathbb{P}} \mathcal{M}_n(\hat{\beta}^{ML})/n$ holds 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).

Proof of Proposition 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 $Z = (Z_0, \ldots, Z_p)$, where

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

(a) Normality

(i)

Redefine optimization problem (2) 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

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

with

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

(ii)
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:

$$\tilde{J}_n(\beta) = J_n(\beta) - J_n(\beta^*)$$

$$\Rightarrow \tilde{J}_n(b) = J_n(b) - J_n(0) =$$

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

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

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

$$+ \sum_{j=1}^{p}\sum_{r=1}^{k}\sqrt{n}\frac{\phi_{r(j)}(n)}{|\hat{\beta}_{jr}^{ML}|}\left(\left|\beta_{jr}^* + \frac{b_{jr}}{\sqrt{n}}\right| - \left|\beta_{jr}^*\right|\right)$$

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

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

or

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

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

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

or

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

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 $\frac{\lambda_n}{\sqrt{n}}\tilde{J}(b)\xrightarrow{\mathbb{P}}\infty$.

(iii)
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 $n^{-1/2}s_n(\beta^*) \overset{d}{\to} N(0, F(\beta^*))$, one obtains

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

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

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

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

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

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 $n^{-1/2}s_n(\beta_{\mathcal{C}}^*) \overset{d}{\to} N(0, F(\beta_{\mathcal{C}}^*))$ and $n^{-1/2}(\beta_{\mathcal{C}} - \beta_{\mathcal{C}}^*) \overset{d}{\to} N(0, F(\beta_{\mathcal{C}}^*)^{-1})$ like mentioned above.
Since the considered minimization problem is convex, the unique minimum of $V(\beta)$ is $(\beta_{\mathcal{C}}^{ML}, 0)^T$ and we have

$$\hat{\beta}_{\mathcal{C}}^n \to \beta_{\mathcal{C}}^{ML}$$

and

$$\hat{\beta}_{\mathcal{C}^{\mathrm{C}}}^n \to 0.$$

Hence, we have as well

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

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

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).

(i)

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

(ii)

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 $\hat{\beta} = \text{argmin}_\beta\ \Psi(\beta) = \text{argmin}_\beta\ -l(\beta) + \frac{\lambda_n}{\sqrt{n}}J(\beta)$ like defined in (a) is equivalent to

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

with

$$\mathfrak{B} = \{\beta : \beta_{0,1}, \ldots, \beta_{l-1,k}, \beta_{l,1} \leq \ldots \leq \beta_{l,z} \leq 0 \leq \beta_{l,z+1} \leq \ldots \leq \beta_{l,k}, \beta_{l+1,1}, \ldots, \beta_{p,k}\}$$

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

and

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

Since $\hat{\beta}_{lq}^n \neq 0$ 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

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

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 $n^{-1/2} s_n(\beta) \xrightarrow{d} N(0, F(\beta))$. Hence for any $\epsilon > 0$, we have

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

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

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

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

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

As a consequence

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

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

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

References

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics 55*(1), 117–128.

Bondell, H. D. and B. J. Reich (2009). Simultaneous factor selection and collapsing levels in ANOVA. *Biometrics 65*(1), 169–177.

Boulesteix, A.-L. (2006). Maximally selected chi-square statistics for ordinal variables. *Biometrical Journal 48*(3), 451–462.

Claeskens, G. and N. L. Hjort (2008a). Minimising average risk in regression models. *Econometric Theory 24*, 493–527.

Claeskens, G. and N. L. Hjort (2008b). *Model Selction and Model Averaging*. Cambridge University Press.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM-algorithm. *Journal of the Royal Statistical Society, Series B 39*, 1–38.

Fahrmeir, L. and H. Kaufmann (1985). Consistency and asymptotic normality of the maximum likelihood estimation in generalized liner models. *The Annals of Statistics 13*(1), 342–368.

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

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Fan, J. and W. Zhang (1999). Statistical estimation in varying coefficient models. *The Annals of Statistics 27*(5), 1491–1518.

Gertheiss, J., V. Stelz, and G. Tutz (2012). Regularization and model selection with categorical covariates. In *Proceedings of the Joint Conference of the German Classification Society and the German Association for Pattern Recognition*. Accepted for publication.

Gertheiss, J. and G. Tutz (2012). Regularization and model selection with categorial effect modifiers. *Statistica Sinica*. To appear.

Grün, B. and F. Leisch (2008). Flexmix version 2: Finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software 28*(4), 1–35.

Hastie, T. and R. Tibshirani (1993). Varying-coefficient models. *Journal of the Royal Statistical Society, Series B 55*(4), 757–796.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*(1), 55–67.

Hofner, B., T. Hothorn, and T. Kneib (2008). Variable selection and model choice in structured survival models. *Department of Statistics at the University of Munich: Technical Reports 43*.

Hoover, D. R., J. A. Rice, C. O. Wu, and L.-P. Yang (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika 85*(4), 809–822.

Kauermann, G. and G. Tutz (2000). Local likelihood estimation in varying-coefficient models including additive bias correction. *Journal of Nonparametric Statistics 12*(3), 343–371.

Leisch, F. (2004). Flexmix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software 11*(8), 1–18.

Leng, C. (2009). A simple approach for varying-coefficient model selection. *Journal of Statistical Planning and Inference 139*(7), 2138–2146.

Liang, K.-Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*(1), 13.

Lin, Y. and H. H. Zhang (2006). Component selection and smoothing in smoothing spline analysis of variance models. *The Annals of Statistics 34*(5), 2272–2297.

Lu, Y., R. Zhang, and L. Zhu (2008). Penalized spline estimation for varying-coefficient models. *Communications in Statistics, Theory and Methods 37*(14), 2249–2261.

McCullagh, P. (1983). Quasi likelihood functions. *The Annals of Statistics 11*(1), 59–67.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B 58*(1), 267–288.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005). Sparsity and smoothness via the fused LASSO. *Journal of the Royal Statistical Society, Series B 67*(1), 91–108.

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

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

Wang, H. and Y. Xia (2009). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association 104*(486), 747–757.

Wang, L., H. Li, and J. Z. Huang (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association 103*(484), 1556–1569.

Wu, C. O., C.-T. Chiang, and D. R. Hoover (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *Journal of the American Statistical Association 93*(444), 1388–1389.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B 68*(1), 49–67.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.