Statistical Modelling 2014; **14**(**2**): 157–177

Regularization and model selection with categorical predictors and effect modifiers in generalized linear models

Margret-Ruth Oelker[1], **Jan Gertheiss**[2] **and Gerhard Tutz**[1]
[1]Department of Statistics, Ludwig-Maximilians-Universität Munich, Germany
[2]Department of Animal Sciences, Georg-August-Universität Göttingen, Germany

Abstract: 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. For computation, a penalized iteratively reweighted least squares algorithm is presented. We investigate large sample properties of the penalized estimates; in simulation studies, we show 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 data.

Key words: Categorical predictors; fused Lasso; generalized linear model; variable selection; varying-coefficients

Received July 2012; **revised** June 2013; **accepted** July 2013

1 Introduction

In regression modelling, the researcher is often faced with categorical predictors, also called factors. Nevertheless, variable selection for discrete covariates and the connected problem which categories within one factor are to be distinguished has been somewhat neglected in the literature.

We analyze data from a consumer study on the acceptance of boar meat. As surgical castration of male piglets, as typically done, shall be banned by 2018 (European Declaration on alternatives to surgical castration of pigs, 2010), the production of so-called entire male pigs may become an alternative. To investigate whether this is indeed a suitable alternative (Meier-Dinkel *et al.*, 2013), we consider meat from four different product groups: (1) castrate or gilt meat (hereafter referred to as 'control') with label 'pork', (2) control with label 'young boar meat', (3) boar meat with label 'pork' and (4) boar meat with label 'young boar meat'. The response is binary saying weather consumers liked the taste of the meat (see Meier-Dinkel *et al.*, 2013).

Address for correspondence: Margret-Ruth Oelker, Department of Statistics, Ludwig-Maximilians-Universität Munich, Germany. E-mail: margret.oelker@stat.uni-muenchen.de

 10.1177/1471082X13503452

We investigate whether the probability of liking depends on the type of meat and/or the labelling, and furthermore, if the influence of other variables like gender, age or health status (sick: yes/no) on liking depends on the product type. Therefore, the label is considered as an effect modifying factor. That is, we address model selection with discrete covariates in a slightly extended version of generalized linear models (GLMs), namely, GLMs with varying coefficients and categorical effect modifiers.

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

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

where $x_1, x_2, \ldots, x_p$ are continuous covariates, and $u_0, \ldots, u_p$ 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 article, 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 assumed as smooth and have been modelled by splines (Hastie and Tibshirani, 1993; Hoover *et al.*, 1998; Lu *et al.*, 2008), using localizing techniques (Wu *et al.*, 1998; Fan and Zhang, 1999; Kauermann and Tutz, 2000) or boosting (Hofner *et al.*, 2012). 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 penalty (Lin and Zhang, 2006), while Wang *et al.* (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 *et al.* (2012), 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 boar data, for instance, the effect modifier label has four categories indicating the product group. 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, $\beta_{j1}, \ldots, \beta_{jkj}$ 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 $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_0^T, \ldots, \boldsymbol{\beta}_p^T)$, where sub-vector $\boldsymbol{\beta}_j^T = (\beta_{j1}, \ldots, \beta_{jk_j})$ 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. Usual maximum likelihood (ML) estimates may not exist; alternative tools such as regularization techniques are needed. Moreover, it is desirable to reduce the model to the relevant terms. 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 data with non-normal response variables such as the boar data. Hence, we present approaches that allow to model categorical effect modifiers within the GLM framework. In Section 2, we propose a penalized ML criterion. For computation of estimates, a different approach other than the classical linear model is needed; a penalized iteratively reweighted least squares algorithm is employed. Moreover, large sample properties of the penalized estimators are derived. As an alternative, we consider a forward selection procedure using information criteria (Section 3). The proposed methods are shown to be highly competitive in numerical experiments (Section 4). In Section 5, the new approaches are applied to the boar data, and the special case of categorical effects is discussed in Section 6.

2 L_1-penalized estimation in GLMs

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

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

where $l_n(\boldsymbol{\beta})$ denotes the log-likelihood for sample size n, and $P_\lambda(\boldsymbol{\beta})$ stands for a general penalty depending on tuning parameter λ. The expression $\lambda \cdot J_n(\boldsymbol{\beta})$ 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 $J_n(\boldsymbol{\beta})$: 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 $\beta_j(\boldsymbol{u}_j)$ 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. We consider the general penalty

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

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

For a *nominal* effect modifier $\boldsymbol{u}_j$ we propose

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

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 $\boldsymbol{\beta}_j(\boldsymbol{u}_j)$. Thus, not only adjacent coefficients but each subset of nominal categories can be collapsed. In the case of strong penalization, effects $\beta_{j1}, \ldots, \beta_{jkj}$ of covariate j are reduced to one constant coefficient and do not depend on the categories of $\boldsymbol{u}_j$ anymore; one obtains $\hat{\beta}_{j1} = \ldots = \hat{\beta}_{jkj} = \hat{\beta}_j$. The second sum in (2.3) conforms to a Lasso penalty shrinking all coefficients belonging to $\boldsymbol{\beta}_j(\boldsymbol{u}_j)$ individually towards zero. The effect is selection and exclusion of covariates. With strong penalization, $\hat{\beta}_{j1} = \ldots = \hat{\beta}_{jkj} = 0$ 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 $\boldsymbol{u}_j$ is *ordinal*, there is additional information. Our proposal is to allow for the fusion of adjacent categories β_{jr} and $\beta_{j,r-1}$. Hence, for ordinal predictors we use

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

where b_j denotes the same indicator as above. Instead of all pairwise differences, now only differences of neighboured 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

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

and for ordinal effect modifiers, it is

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

Parameter ψ is restricted to (0, 1) in order to separate it strictly from tuning parameter λ.

If effect modifiers $\boldsymbol{u}_j$ have different numbers of categories, additional weighting of penalty terms analogously to Bondell and Reich (2009) could be used to prevent an 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. In the classical linear model, quadratic programming can be used, or the solution can be approximated by employing the lars algorithm (Efron *et al.*, 2004); see Gertheiss and Tutz (2012) for details. In a GLM, however, a more general approach is needed. Non-differentiability can be evaded by approximating the penalty at the critical points, i.e., in a neighbourhood of $|\xi|$, $\xi = 0$. As for example in Koch (1996), the absolute values $|\xi|$ in the penalty are approximated by the differentiable function $\sqrt{\xi^2 + c}$, where c denotes a small positive constant. Combining this approximation with a local trick of Fan and Li (2001) and a proposal to complete the square of Ulbricht (2010) allow to derive a penalized iteratively reweighted least squares (PIRLS) algorithm like the one described in Oelker and Tutz (2013).

We assume a penalty that can be written as $P_\lambda(\boldsymbol{\beta}) = \sum_{l=1}^{L} p_{\lambda,l}(|\boldsymbol{\alpha}_l^T \boldsymbol{\beta}|)$, where a_l are known vectors. Like in Ulbricht (2010), penalty terms $p_{\lambda,l}(|\boldsymbol{a}_l^T \boldsymbol{\beta}|)$ are supposed to map $|\boldsymbol{a}_l^T \boldsymbol{\beta}|$ onto the positive real numbers, to be continuous and monotone in $|\boldsymbol{a}_l^T \boldsymbol{\beta}|$. In addition, penalty terms $p_{\lambda,l}(|\boldsymbol{a}_l^T \boldsymbol{\beta}|)$ are assumed to be continuously differentiable $\forall\, \boldsymbol{a}_l^T \boldsymbol{\beta} \neq 0$ such that $\mathrm{d}p_{\lambda,l}(|\boldsymbol{a}_l^T \boldsymbol{\beta}|)/\mathrm{d}|\boldsymbol{a}_l^T \boldsymbol{\beta}| \geq 0 \;\forall\, |\boldsymbol{a}_l^T \boldsymbol{\beta}| > 0$ holds. Penalty $J_n(\boldsymbol{\beta})$ from

equation (2.2) fits in this framework: let the vectors a_l denote the columns of a block-diagonal matrix $\boldsymbol{A} = \text{diag}(\boldsymbol{A}_0, \ldots, \boldsymbol{A}_p) \subset \mathbb{R}^{q \times L}$ and functions $p_{\lambda,l}(\nu)$ be defined as $\lambda \cdot \nu$. Let block $\boldsymbol{A}_j$ refer to the effect modifier $\boldsymbol{u}_j$. If $\boldsymbol{u}_j$ is nominal, $\boldsymbol{A}_j^T \boldsymbol{\beta}_j$ shall give the values of the according coefficients $\beta_{j1}, \ldots, \beta_{jkj}$ 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

$$\boldsymbol{A}_j^{nom} = \begin{pmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 0 & -1 & -1 & 0 \\ 0 & 1 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 0 & 1 & -1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix},$$

which is a $k_j \times \left(\frac{1}{2} k_j (1 + k_j) \right)$ dimensional matrix. If $\boldsymbol{u}_j$ is ordinal, only pairwise differences of coefficients $\beta_{j1}, \ldots, \beta_{jkj}$ are penalized. Thus, in $(k_j \times (2k_j - 1))$ matrix $\boldsymbol{A}_j^{ord}$ the last three columns of matrix $\boldsymbol{A}_j^{nom}$ 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 $\boldsymbol{A}_j^{nom}, \boldsymbol{A}_j^{ord}$ respectively, is omitted. For a covariate x_j whose influence on $\boldsymbol{y}$ is not modified by any $\boldsymbol{u}_j$, matrix $\boldsymbol{A}_j^{none}$ is an empty matrix with zero columns and k_j rows.

With this representation and starting with an initial value $\hat{\boldsymbol{\beta}}_{(0)}$, we obtain

$$\hat{\boldsymbol{\beta}}_{(k+1)} = (1 - \nu) \cdot \hat{\boldsymbol{\beta}}_{(k)} + \nu \cdot (\boldsymbol{X}^T \boldsymbol{W}_{(k)} \boldsymbol{X} + \boldsymbol{A}_\lambda)^{-1} \boldsymbol{X}^T W_{(k)} \tilde{\boldsymbol{y}}_{(k)},$$

where $\hat{\boldsymbol{\beta}}_{(k)}$ is the estimate of the current iteration; the matrix $\boldsymbol{W}_{(k)}$ denotes weights and $\tilde{\boldsymbol{y}}_{(k)}$ denotes pseudo-observations like in usual GLMs. We have $\boldsymbol{W}_{(k)} = \boldsymbol{D}_{(k)} \boldsymbol{\Sigma}_{(k)}^{-1} \boldsymbol{D}_{(k)}$, $\boldsymbol{D}_{(k)} = \text{diag}(\partial h(\boldsymbol{\eta}_i(\hat{\boldsymbol{\beta}}_{(k)}))/\partial \boldsymbol{\eta})$, $\boldsymbol{\Sigma}_{(k)} = \text{diag}(\sigma_i^2(\hat{\boldsymbol{\beta}}_{(k)}))$, $\tilde{\boldsymbol{y}} = \boldsymbol{D}_{(k)}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_{(k)}) + \boldsymbol{X}\hat{\boldsymbol{\beta}}_{(k)}$ and $\boldsymbol{\mu}_{(k)} = h(\boldsymbol{\eta})$. ν is a step length parameter that usually equals 1; it allows to control the algorithm's convergence whenever necessary and avoids back stepping. Matrix $\boldsymbol{A}_\lambda \in \mathbb{R}^{q \times q}$ implements the approximated penalty:

$$\boldsymbol{A}_\lambda = \sum_{l=1}^{L} \frac{\lambda \cdot 1_{\{a_l^T \hat{\beta}_{(k)} \neq 0\}}}{\sqrt{(a_l^T \hat{\boldsymbol{\beta}}_{(k)})^2 + c}} \cdot \boldsymbol{a}_l \boldsymbol{a}_l^T.$$

It is updated in each iteration; the approximation of the absolute values is enhanced continually. In general, function $\sqrt{\xi^2 + c}$ deviates only slightly from the absolute $|\xi|$; for $\xi = 0$ the deviation is $\sqrt{c}$, for all other values of ξ, the deviation is smaller than $\sqrt{c}$. The algorithm stops when $|\hat{\boldsymbol{\beta}}_{(k+1)} - \hat{\boldsymbol{\beta}}_{(k)}| / |\hat{\boldsymbol{\beta}}_{(k)}| \leq \varepsilon$, for a fixed small $\varepsilon > 0$. The generalized hat matrix of the algorithm's final iteration allows to estimate the model's degrees of freedom. The presented algorithm, however, 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. The penalty applied here leads to a positive semi-definite penalty matrix. Therefore, for $\boldsymbol{X}^T \boldsymbol{W}_{(k)} \boldsymbol{X}$ positive definite, the quasi-Newton approach will find descent directions in each iteration; but for the $q > n$ case, it may happen that the solution is not unique (Ulbricht, 2010). In this case, we recommend to use several starting values and to check the likelihood scores of the according solutions; however, in our experience, this is a minor problem.

Moreover, the proposed computational approach has substantial advantages: it keeps within the established framework of Fisher scoring algorithms. It allows not only for categorical effect modifiers but handles a more general penalized likelihood problem and can hence be extended easily to other penalties. Besides the update of matrix $\boldsymbol{A}_\lambda$, the computational burden is the same as for Fisher scoring algorithms. Even though each set of coefficients has to be computed separately, the method gives coefficient paths.

The proposed algorithm is implemented in the R package gvcm.cat (R Development Core Team, 2012; Oelker, 2013). Besides the algorithm itself, the package provides a formula environment and different options for cross-validation; it is possible to plot coefficient paths, cross-validation scores and coefficient profiles.

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 the sample size n tends to infinity, it is assumed that the number of observations n_{jr} on level r of $\boldsymbol{u}_j$ tends to infinity for all j, r at the same rate. Practically, that means, that asymptotically the probability for an observation on level r of $\boldsymbol{u}_j$ must be positive and tend to a constant c_{jr} for all j, r. Then we have

Theorem 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* $\hat{\boldsymbol{\beta}}$ *that minimizes (2.1) with* $J_n(\beta)$ *defined by (2.2), (2.3) and (2.4) is consistent, i.e.,* $\lim_{n\to\infty} \mathbb{P}(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|^2 > \varepsilon) = 0$ *for all* $\varepsilon > 0$.

The proof is given in Supplement 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. To gain selection consistency, Zou (2006) proposed an adaptive version of the original Lasso that has the so-called oracle properties. A corresponding modification for penalty (2.2) is available: Given effect modifiers $\boldsymbol{u}_j$, $j = 0, \ldots, p$, penalty $J_n(\boldsymbol{\beta})$ (2.2) is modified to the adaptive penalty $J_n^{ad}(\boldsymbol{\beta})$ by employing

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

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

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}; functions $\phi_{rs(j)}(n)$ and $\phi_{r(j)}(n)$ are additional weights for the penalty terms that are assumed to converge to fixed values: $\phi_{rs(j)}(n) \to q_{rs(j)}$ and $\phi_{r(j)}(n) \to q_{r(j)}$, with $0 < q_{rs(j)}, q_{r(j)} < \infty$. If $\phi_{rs(j)}(n) = \phi$ and $\phi_{r(j)}(n) = 1 - \phi$, $0 < \phi < 1$, are global constants, we obtain a generalization with the same structure as given in equations (2.5) and (2.6); $0 < \phi < 1$ or similar constraints for functions $\phi_{rs(j)}(n)$, $\phi_{r(j)}(n)$, guarantee that the effect of the weights and the effect of the global tuning parameter λ are separated. Hence, the adaptive weight of a penalty term becomes huge when the ML-estimate of the penalty term is close to zero. The adaptive weight becomes smaller as the ML-estimate of the penalty term gets bigger. Thus, adaptive weights favour to set coefficients with small ML-estimates to zero, to fuse coefficients with close ML-estimates, respectively. Technically, with some additional assumptions, this ensures selection consistency: First of all, the penalty parameter λ has to increase with sample size n; 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 some vectors: $\hat{\boldsymbol{\beta}}^n$ denotes the estimate of $\boldsymbol{\beta}$; we emphasize that it is based on the sample size n. Then, the vector $\hat{\boldsymbol{\theta}}^n = \boldsymbol{A}^T\hat{\boldsymbol{\beta}}^n$ with block-diagonal

matrix $\boldsymbol{A} \in \mathbb{R}^{q \times L}$ contains the estimates of all the terms in penalty (2.2), that is, the estimated values of all penalized coefficients $\hat{\beta}_{ij}$ and—according to the level of measurement—the estimated values of their differences. Furthermore, define $\mathcal{C}$ and $\mathcal{C}_n$. $\mathcal{C}$ denotes the set of indices corresponding to those entries of $\hat{\boldsymbol{\theta}}^n$ which are truly non-zero; whereas $\mathcal{C}_n$ denotes the estimate of $\mathcal{C}$ based on n observations. $\boldsymbol{\theta}^*_{\mathcal{C}}$ is the vector with the true values of the entries in $\mathcal{C}$; $\hat{\boldsymbol{\theta}}^n_{\mathcal{C}}$ denotes its estimate.

Previous assumptions concerning ML-estimation are extended: the model must hold, the negative log-likelihood $-l_n(\boldsymbol{\beta})$ has to be convex. $l_n(\boldsymbol{\beta})$ has to be at least three times continuously differentiable, the third moments of $\boldsymbol{y}$ have to be finite. Let $\boldsymbol{F}_n = \mathbb{E}\left(-\frac{\partial^2 l_n(\beta)}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}^T}\right)$ denote the expected information matrix, then $\boldsymbol{F}_n/n$ must have a positive definite limit $\boldsymbol{F}$; for the score function $s_n(\boldsymbol{\beta}) = \frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$, we suppose $\mathbb{E}(s_n(\boldsymbol{\beta})) = \mathbf{0}$. Then one obtains

Theorem 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* $J_n^{ad}(\boldsymbol{\beta})$ *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) $\sqrt{n}(\hat{\boldsymbol{\theta}}^n_{\mathcal{C}} - \boldsymbol{\theta}^*_{\mathcal{C}}) \xrightarrow{d} N(\mathbf{0}, \mathrm{Cov}(\boldsymbol{\theta}^*_{\mathcal{C}}))$

(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 Supplement A. The concrete form of $\mathrm{Cov}(\boldsymbol{\theta}^*_{\mathcal{C}})$ results from the asymptotic marginal distribution of a set of non-redundant truly non-zero regression parameters or differences thereof. Since all estimated differences are (deterministic) linear functions of estimated parameters, the covariance-matrix $\mathrm{Cov}(\boldsymbol{\theta}^*_{\mathcal{C}})$ is singular.

$\boldsymbol{F}_n/n \xrightarrow[n\to\infty]{} F$ with positive definite $\boldsymbol{F}$ 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 $-l_n(\boldsymbol{\beta})$ for $n \to \infty$.

For the normality part of Theorem 2, the speed of convergence is $\lambda_n/\sqrt{n} \to 0$. Since $n^{-1/2} s_n(\boldsymbol{\beta}) \sim N(\mathbf{0}, \boldsymbol{F}_n(\boldsymbol{\beta})/n) + \mathcal{O}_p(n^{-1/2})$ and $\mathbb{P}(\sqrt{n}\,|\hat{\beta}_{jr}^{ML}| \le \lambda_n^{1/2}) \to 1$ like $c/\sqrt{n} \to 0$,

the consistency part behaves the same. Thus, the overall speed of convergence is $\mathcal{O}_p(n^{-1/2})$.

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 often used. In particular, forward and backward selection methods based on information criteria like the Akaikes Information Criterion (AIC) or the Bayesian Information Criterion (BIC) are popular. One tries to find the model that performs best with respect to the chosen criterion. By construction, these strategies yield variable selection but no fusion of categories. Gertheiss and Tutz (2012) obtain the 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 sub-vector $(\beta_{j1}, \beta_{j2}, \beta_{j3})^T$ in coefficient vector $\boldsymbol{\beta}$. All possible selections of coefficients belonging to x_j would be: $\{(), (\beta_{j1}), (\beta_{j2}), (\beta_{j3}), (\beta_{j1}, \beta_{j2}), (\beta_{j1}, \beta_{j3}), (\beta_{j2}, \beta_{j3}), (\beta_{j1}, \beta_{j2}, \beta_{j3})\}$. 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 select and/or 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 $\hat{\boldsymbol{\beta}}$ (Tibshirani *et al.*, 2005). Hence, in each step, a former zero coefficient can be set to non-zero, or an entire group of zero coefficients can become non-zero, but with all coefficients within this group being equal. Alternatively, a group of non-zero but identical coefficients can be split into two groups of non-zero coefficients, with coefficients now being identical within each of both groups but different between groups.

4 Simulation studies

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

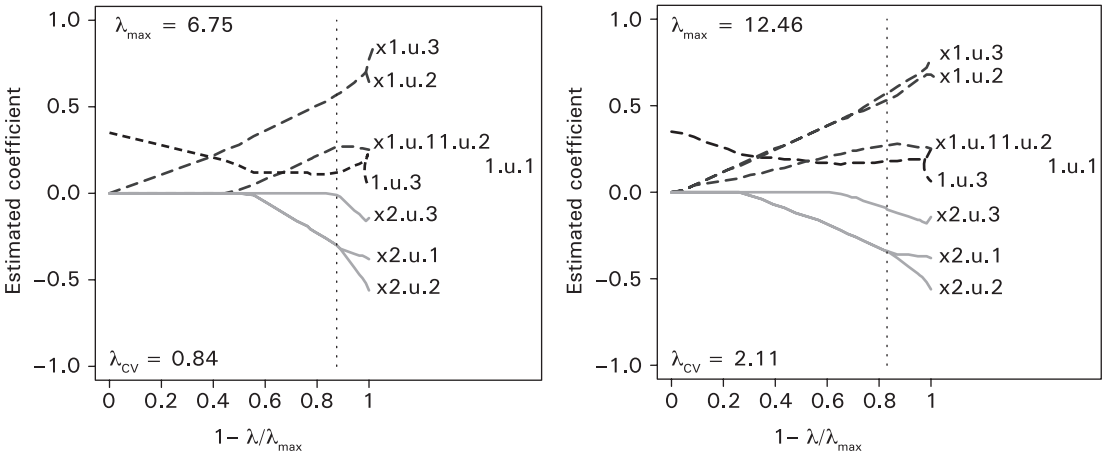

Figure 1 Coefficient paths for binary model (4.1) assuming predictor (4.2)—with adaptive weights (left) and the standard penalty (right).

Source: Authors' own.

4.1 Illustrative example

We assume a logistic regression model with two covariates $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and one nominal effect modifier u with categories 1, 2 and 3. u possibly impacts all covariates plus the intercept. Concretely, the predictor is

$$\begin{aligned}\boldsymbol{\eta}_{true} &= \beta_0(\boldsymbol{u}) + \boldsymbol{x}_1\beta_1(\boldsymbol{u}) + x_2\beta_2(\boldsymbol{u}) \\ &= \beta_0 + \boldsymbol{x}_1\,(\beta_{11}I(\boldsymbol{u}=1) + \beta_{12}I(\boldsymbol{u}=2) + \beta_{13}\boldsymbol{I}(\boldsymbol{u}=3)) + \boldsymbol{x}_2\,\beta_2 \\ &= 0.2 + \boldsymbol{x}_1\,(0.3I(\boldsymbol{u}=1) + 0.7I(\boldsymbol{u}=2) + 0.7I(\boldsymbol{u}=3)) - \boldsymbol{x}_2\cdot 0.5. \qquad (4.1)\end{aligned}$$

That means, while the intercept and $\boldsymbol{x}_2$ do not depend on $\boldsymbol{u}$, covariate $\boldsymbol{x}_1$varies with categories 1 and 2/3 of $\boldsymbol{u}$. Covariates $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are independently drawn from an uniform distribution $U(0, 2)$; the effect modifier $\boldsymbol{u}$ is multinomial with probabilities 0.3, 0.4 and 0.3 for categories 1, 2 and 3, respectively. For response $\boldsymbol{y}$, $\boldsymbol{y} = h(\boldsymbol{\eta})$ 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, that is, we have

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

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 $\boldsymbol{x}_1$ are fused such that only category 1 makes a difference. Concerning covariate $\boldsymbol{x}_2$, 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 $Dev(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ as loss function. It shrinks coefficients slightly—in return all but one relevant structures are identified. Absolute deviation to the true coefficients is small.

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 2.11 now. However, the performance is worse than with adaptive weights: in the model chosen by cross-validation (see dotted line), coefficients of covariate $\boldsymbol{x}_1$ are not fused.

4.2 Comparison of methods

4.2.1 Simulation settings

To compare the proposed methods, various model features are systematically varied. Concretely, we consider a model with binomial response, two influential covariates and six 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 *S200* and *S600*. 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, …, 4 and is independently drawn from a multinomial distribution with probability 0.25 per category. The true linear predictor is

$$\begin{aligned}\boldsymbol{\eta}_{true} &= \beta_0(\boldsymbol{u}) + \boldsymbol{x}_1\beta_1(\boldsymbol{u}) + \boldsymbol{x}_2\beta_2(\boldsymbol{u}) \\ &= (0.7I(\boldsymbol{u}=1) + 0.7I(\boldsymbol{u}=2) + 0I(\boldsymbol{u}=3) + 0I(\boldsymbol{u}=4)) \\ &\quad + \boldsymbol{x}_1\,(1I(\boldsymbol{u}=1) - 1.5I(\boldsymbol{u}=2) - 1.5I(\boldsymbol{u}=3) + 0.5I(\boldsymbol{u}=4)) \\ &\quad + \boldsymbol{x}_2\,(0I(\boldsymbol{u}=1) + 1I(\boldsymbol{u}=2) + 2I(\boldsymbol{u}=3) - 3I(\boldsymbol{u}=4)).\end{aligned}$$

Since the truly varying coefficients are to be detected by the procedure, 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

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

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, we generate 50 data sets, that is, for each setting each method is evaluated 50 times.

4.2.2 Results

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

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

where $q = \sum_{j=0}^{p} k_j$, $\boldsymbol{\beta}$ denotes the vector of true coefficients and $\hat{\boldsymbol{\beta}}$ its estimate. To judge the prediction accuracy, the mean predictive deviance $Dev(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ is considered, referred to as MSEP. Figures 2 and 3 show the boxplots of MSE and MSEP for both settings. Error values of penalized approaches and forward selection strategies tend to be smaller than those of the ML-estimates. However, in particular forward selection based on AIC suffers from a high variability—especially for $n = 200$, several extreme values are observed in Figure 2. Also for penalization with adaptive weights, we see that the variability becomes smaller compared to 'standard' penalization only for increasing sample size. This is due to the construction of the adaptive weights, which are the inverses of the ML-estimates. The more observations we have, the more stable is the ML-estimate and so is the corresponding weight. We clearly see that for small sample sizes oracle properties from Theorem 2 are not given at all, but with n becoming larger adaptive estimates become better and better. Of course, the point at which the estimator with superior asymptotic properties becomes superior for finite samples depends on the concrete setting.

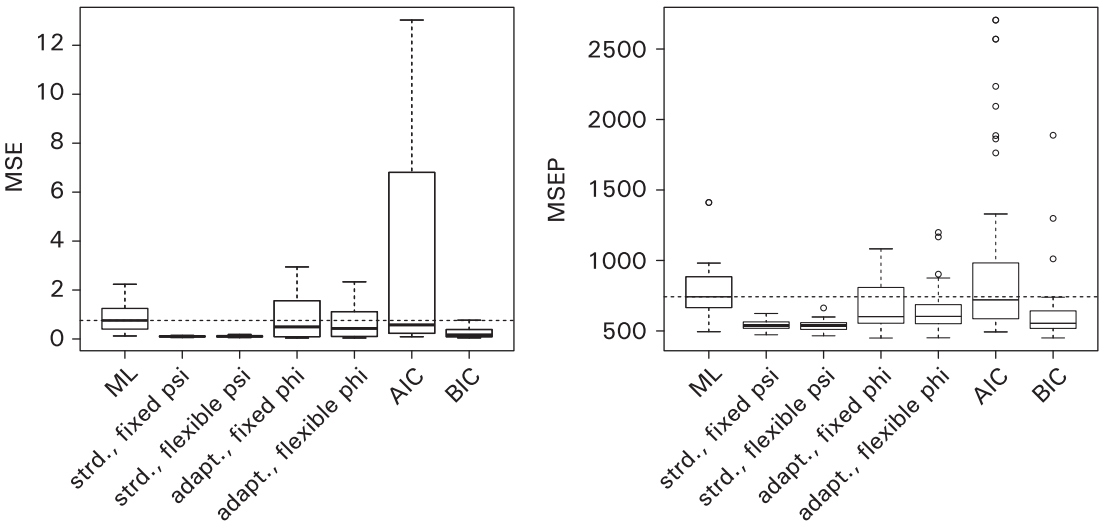

Figure 2 Boxplots of scaled squared errors (MSE, left panel) and deviances (MSEP, right panel) for setting *S200*; in the left panel outliers are omitted.

Source: Authors' own.

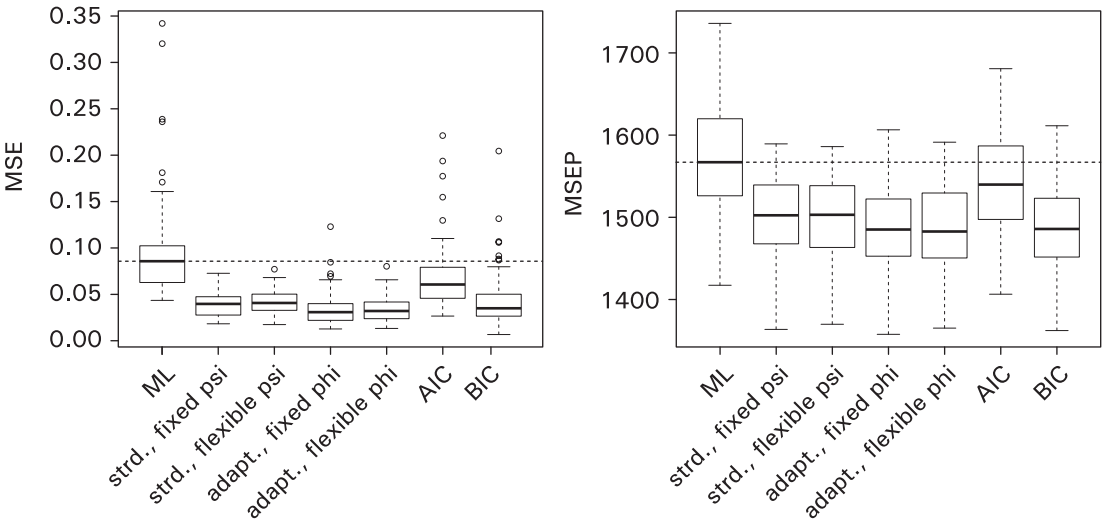

Figure 3 Boxplots of scaled squared errors (MSE, left panel) and deviances (MSEP, right panel) for setting *S600*; medians mark estimates of MSE and MSEP.

Source: Authors' own.

In addition, we evaluate the clustering and selection performance. According to Theorem 2, using the adaptive estimator (with selection consistency) should yield better models in terms of selection and clustering. 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

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

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

$\text{FPR}_{clustering}$ and $\text{FNR}_{clustering}$ are defined analogously, but refer to differences of coefficients. Table 1 shows false positive and negative rates for both settings.

As expected, the adaptive penalty tends to perform better than the standard version. False positive rates are much smaller for the first one. For small n, however, false negative rates are substantially larger when using adaptive weights. This illustrates that selection consistency is an asymptotic property that may not necessarily yield best results for small sample sizes. With increasing n, false negative rates are quite small for the adaptive version, too. The reason why false positive rates are still rather high (for both adaptive and non-adaptive weights) is that the penalty parameters are chosen by cross-validation, and cross-validation tends to select accurate estimates but a somewhat too large model. AIC based forward selection performs similar here. However, having the high variability from above in mind, the

Table 1 Estimates of false positive and false negative rates for settings *S200* and *S600*

		ML	strd., fixed psi	strd., flexible psi	adapt., fixed phi	adapt., flexible phi	AIC	BIC
Setting *S200*	$\text{FPR}_{selection}$	1	0.77	0.65	0.34	0.39	0.41	0.16
	$\text{FNR}_{selection}$	0	0.03	0.04	0.08	0.06	0.07	0.11
	$\text{FPR}_{clustering}$	1	0.64	0.69	0.43	0.42	0.40	0.10
	$\text{FNR}_{clustering}$	0	0.05	0.03	0.15	0.15	0.19	0.27
Setting *S600*	$\text{FPR}_{selection}$	1	0.81	0.71	0.43	0.39	0.39	0.11
	$\text{FNR}_{selection}$	0	0.00	0.01	0.01	0.01	0.02	0.03
	$\text{FPR}_{clustering}$	1	0.77	0.76	0.45	0.42	0.37	0.05
	$\text{FNR}_{clustering}$	0	0.01	0.00	0.04	0.03	0.08	0.17

Source: Authors' own.

previous recommendation for adaptive weights still holds. With BIC, by contrast, typically a much smaller model is selected, leading to smaller false negative but substantially larger false positive rates. So if the primary goal is a sparse model, and the analyst is willing to risk that a number of truly relevant variables/differences are disregarded, the BIC based forward selection may be an alternative. Otherwise, sparseness and relatively low false negative rates are obtained by the proposed penalty with adaptive weights.

5 Application: acceptance of boar meat and the effect of labelling

A known sensory problem with respect to boars is the occurrence of so-called boar taint, which may affect consumer acceptance of boar meat; see, e.g., Mörlein *et al.* (2012), Meier-Dinkel *et al.* (2013), and references therein. However, liking or disliking a food product does not only depend on the product's physicochemical properties but also on the consumers' expectations. Therefore we are interested in whether the consumer acceptance is affected solely by labelling meat as 'boar meat'. In addition, various consumer characteristics may influence individual liking or disliking of boar meat, such as age or gender of the consumer. The data considered here is a subset from Meier-Dinkel *et al.* (2013). Consumers tasted meat from four different product groups: (1) control (castrate or gilt meat) with label 'pork' (2) control with label 'young boar meat', (3) real boar meat with label 'pork' and (4) boar meat with label 'young boar meat'. We are interested in whether the probability of liking the taste of the product (binary response $y \in \{0, 1\}$) is affected by the product type, and in particular whether acceptance of pork/boar meat even differs between labels 'boar' and 'pork'. Hence, we include the product type as a categorical effect modifier in a logistic regression model and allow the influence of various covariates to change with the product type: consumer's age, gender, smoker (no/yes), sick (olfactory disability caused by sickness, in particular cold and allergy: no/yes) and a factor indicating whether the consumer knows what 'boar meat' means (self-reported knowledge: no/yes). In addition, we correct for the effect of contact to animal husbandry (contact: no/yes). Table 2 shows the coefficients estimated by pure maximum likelihood (block 1) and using the proposed penalty approach, for both model selection and estimation of coefficients (block 2). The sample size is 133, which is small for a binary model with 28 parameters. This may explain the quite extreme ML-estimates. Regularized estimates can be expected to be more reliable here.

When using our approach, we see that on average the probability of acceptance is estimated as equal for control and boar meat that is labelled as 'pork', as in the intercept it is not distinguished between these three groups. But when we have a closer look at the consumers, this picture changes. In particular, if the consumer knows what 'boar meat' means, the chance of accepting boar with label 'pork' (product group (3)) decreases quite drastically. At first glance, this seems to contradict the hypothesis that consumers' expectations influence liking of products, but it may be explained by some sort of disappointment effect, as the consumer expected to taste

Table 2 Estimates for all methods fitted to the boar data. Excluded coefficients are omitted. Non-varying coefficients are represented by the remaining scalar only

	ML-estimation				Penalized estimation				Forward selection AIC				Forward selection BIC			
	Product group				Product group				Product group				Product group			
Coefficients	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
$\beta_0(t)$	9.2875	13.6364	10.9538	2.3709	2.28	2.28	2.28	1.49			1.54				1.72	
$\beta_{\text{know.lab}}(t)$	9.5848	9.2609	−9.9293	0.0791	1.60	1.60	−1.60	0.00	0.80	0.80	0.00	0.80	0.73	0.73	0.00	0.73
$\beta_{\text{gender}}(t)$	0.9793	8.8767	−0.3770	0.6693	0.56	2.18	−.01	0.56	0.65	5.41	0.00	0.65	0.78	0.78	0.00	0.78
$\beta_{\text{contact}}(t)$	−.8248	0.0709	9.3493	−.1013	0.00	0.00	0.98	0.00								
$\beta_{\text{age}}(t)$	0.0252	0.2073	−.0360	−.0356	0.00	0.07	−.02	−.02	0.00	0.19	0.00	0.00				
$\beta_{\text{smoker}}(t)$	0.0222	0.4495	−.2935	0.3936												
$\beta_{\text{sick}}(t)$	0.8016	0.3018	9.0947	0.1464	0.08	0.00	1.01	0.00								

Source: Supplement B.

pork, as labelled, and not boar. For control meat (group (1) and (2)), by contrast, there is a positive effect of knowledge which is constant over both labels. Only when boar meat is labelled correctly (group (4)), there is no effect of knowledge as the coefficient is set to zero. The latter two findings, as well as the effect of gender, are rather difficult to explain, but there is another interesting effect of labelling: if boar is labelled as 'pork' (product group (3)), being sick increases the chance that the consumer likes the taste of the product. A possible explanation is that sickness affects the sense of taste and the sense of smell, and sick people hence rather rely on the label saying that it's pork and not boar. Though smoking might affect the sense of taste, the consumer's smoking status is fitted as completely irrelevant for acceptance of the meat, as all coefficients are set to zero. If we look at the effect of age, we see that for older people the chance of liking control meat labelled as boar increases, but the chance of liking boar—no matter which label is attached—decreases. Possibly, the expectation of a certain taste increases with age.

Using the AIC/BIC-based forward selection (block 3/4), the model is much sparser. This may indicate that at least some of the effects found above are false positives. But, as seen in the simulations, forward selection strategies may have rather large false negative rates. To obtain more insight, further studies are necessary and indeed currently conducted.

6 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 $\boldsymbol{u}_j$ is categorical and the modified covariate $\boldsymbol{x}_j$ is a constant vector. We have for example $1 \cdot \boldsymbol{\beta}_j(\boldsymbol{u}_j) = 1 \cdot \sum_{r=1}^{k_j} \beta_{jr} I(\boldsymbol{u}_j = r)$. 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 contrast to Gertheiss and Tutz (2010), all categories of a categorical effect are penalized in the same way.

7 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 categorical 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 the scale parameter 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, particularly when based on the AIC, which makes them less attractive. With BIC, typically a smaller model than with L_1-regularization is selected, which leads to smaller false positive but higher false negative rates.

Lasso-type penalties imply not continuously differentiable optimization problems, and in GLMs different algorithms other than in the linear model have to be used. For categorical effect modifiers, we adopted algorithms of Fan and Li (2001) and of Ulbricht (2010). All functions are available in the R add-on package gvcm.cat (Oelker, 2013).

In practice, varying-coefficient models are highly relevant. We analyzed data from a consumer study on boar meat. We could confirm that the chance of consumer acceptance is smaller for boar meat than for regular pork (castrate or gilt meat). In addition, we could find some evidence of labelling effects. For instance, if wrong labelling causes too high expectations, disappointment substantially reduces the chance of acceptance. If the sense of taste is affected by sickness consumers seem to rely on the labelling.

So far we employed a single penalty parameter λ only; for a modest number of effect modifiers, however, one tuning parameter per effect modifier could be advantageous.

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 article we assumed continuous covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$. But, of course, covariates can be categorical, too. Then, there are even more coefficients, and hence, there is an even stronger demand for regularization. Often there will be additional covariates whose influence is not modified by a categorical variable, but by a continuous one, or not at all. In the latter case the covariate can simply be included in the model without penalization, or with the standard Lasso. Technically, this is just a special case of the model

considered here—the corresponding effect modifier has just one level. If the effect modifier is continuous, smooth functions can be specified. If smoothing is done by using a penalty approach, e.g., penalized splines, similar penalization strategies to fuse and/or select functions can be applied.

Acknowledgements

This work was partially supported by DFG project 'Regularisierung für diskrete Datenstrukturen'. We thank Lisa Meier-Dinkel and Daniel Mörlein (Department of Animal Sciences, University of Göttingen) for providing the boar data and related information. 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!

Supplement A – Proofs
Proofs of Theorems 1 and 2 in Section 2.2.

Supplement B – Example R Code
Data and R-code to reproduce the data analysis from Section 5.

References

Bondell HD and Reich BJ (2009) Simultaneous factor selection and collapsing levels in ANOVA. *Biometrics*, **65**, 169–77.

Efron B, Hastie T, Johnstone I and Tibshirani R (2004) Least angle regression. *The Annals of Statistics*, **32**, 407–99.

Fahrmeir L and Kaufmann H (1985) Consistency and asymptotic normality of the maximum likelihood estimator in generalized liner models. *The Annals of Statistics*, **13**, 342–68.

Fan J and Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348–60.

Fan J and Zhang W (1999) Statistical estimation in varying coefficient models. *The Annals of Statistics*, **27**, 1491–518.

Gertheiss J and Tutz G (2010) Sparse modelling of categorial explanatory variables. *Annals of Applied Statistics*, **4**, 2150–80.

Gertheiss J and Tutz G (2012) Regularization and model selection with categorial effect modifiers. *Statistica Sinica*, **22**, 957–82.

Hastie T and Tibshirani R (1993) Varying-coefficient models. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, **55**, 757–96.

Hoerl AE and Kennard RW (1970) Ridge regression: biased estimation for non-orthogonal problems. *Technometrics*, **12**, 55–67.

Hofner B, Hothorn T and Kneib T (2012) Variable selection and model choice in structured survival models. *Computational Statistics*, **28**, 1079–101.

Hoover DR, Rice JA, Wu CO and Yang L-P (1998) Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, **85**, 809–22.

Kauermann G and Tutz G (2000) Local likelihood estimation in varying-coefficient models including additive bias correction. *Journal of Nonparametric Statistics*, **12**, 343–71.

Koch I (1996) On the asymptotic performance of median smoothers in image analysis and nonparametric regression. *The Annals of Statistics*, **24**, 1648–66.

Leng C (2009) A simple approach for varying-coefficient model selection. *Journal of Statistical Planning and Inference*, **139**, 2138–46.

Lin Y and Zhang HH (2006) Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, **34**, 2272–97.

Lu Y, Zhang R and Zhu L (2008) Penalized spline estimation for varying-coefficient models. *Communications in Statistics - Theory and Methods*, **37**, 2249–61.

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

Meier-Dinkel L, Trautmann J, Frieden L, Tholen E, Knorr C, Sharifi AR, Bücking M, Wicke M and Mörlein D (2013) Consumer perception of boar meat as affected by labelling information, malodorous compounds and sensitivity to androstenone. *Meat Science*, **93**, 248–56.

Mörlein D, Grave A, Sharifi AR, Bücking M and Wicke M (2012) Different scalding techniques do not affect boar taint. *Meat Science*, **91**, 435–40.

Oelker M-R (2013) *gvcm.cat: regularized categorial effects/categorial effect modifiers in GLMs*. R package version 1.5.

Oelker M-R and Tutz G (2013) A general family of penalties for combining differing types of penalties in generalized structured models. *Department of Statistics: Technical Report 139*, http://epub.ub.uni-muenchen.de/14735/.

R Development Core Team (2012) *R: a language and environment for statistical computing*. Vienna, Austria, ISBN 3-900051-07–0.

Tibshirani R (1996) Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, **58**, 267–88.

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

Ulbricht J (2010) *Variable selection in generalized linear models*. Dissertation, Department of Statistics, Ludwig-Maximilians-Universität München: Verlag Dr. Hut.

Wang H and Xia Y (2009) Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association*, **104**, 747–57.

Wang L, Li H and Huang JZ (2008) Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association*, **103**, 1556–69.

Wu CO, Chiang C-T and Hoover DR (1998) Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *Journal of the American Statistical Association*, **93**, 1388–89.

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

Zou H (2006) The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418–29.