Statistical Modelling 2015; **15(5)**: 389–410

Selection and fusion of categorical predictors with L_0-Type penalties

Margret-Ruth Oelker[1], Wolfgang Pößnecker[1] and Gerhard Tutz[1]
[1]Department of Statistics, Ludwig-Maximilians-Universität München, Germany

Abstract: In regression modelling, categorical covariates have to be coded. Depending on the number of categorical covariates and on the number of levels they have, the number of coefficients can become huge. To reduce the model complexity, coefficients of similar categories should be fused and coefficients of non-influential categories should be set to zero. To this end, Lasso-type penalties on the differences of coefficients are a standard approach. However, the clustering/selection performance of this approach is sometimes poor—especially when the adaptive weights are badly conditioned or not existing. In some situations, there is no incentive to cluster similar categories. To overcome this, a L_0 penalty on the differences of coefficients is proposed, whereby the L_0 'norm' is defined as the number of non-zero entries in a vector. The proposed penalty favours to find clusters of categories that share the same effect on the response variable while the estimation accuracy is comparable to Lasso-type penalties. Numerical experiments within the framework of generalized linear models are promising. For illustration, data on the unemployment rates in Germany is analyzed.

Key words: adaptive Lasso; best subset selection; GLMs; model selection

Received January 2014; **revised** July 2014; **accepted** August 2014

1 Introduction

In the majority of regression problems, at least some of the available covariates are categorical. A categorical covariate has to be coded. Depending on the number of categorical covariates and on the number of levels they have, the number of coefficients can become huge. Hence, the accuracy of estimates can be poor. Moreover, when including categorical variables, users want to know if and how these predictors determine the response, and, in particular, which categories have to be distinguished. Typically, there are subsets of categories that have the same effect on the response variable. Recently, various approaches to obtain selection and fusion of categories by regularized estimation have been proposed: Bondell and Reich (2009) propose to apply the fused Lasso (Tibshirani *et al.*, 2005) to the coefficients of a nominal predictor; all pairwise differences of coefficients are penalized. For ordered

Address for correspondence: Margret-Ruth Oelker, Department of Statistics, Ludwig-Maximilians-Universität München, Akademiestr. 1, 80799 München, Germany.
E-mail: margret.oelker@stat.uni-muenchen.de

© 2015 SAGE Publications 10.1177/1471082X14553366

factors, it is more appropriate to penalize differences of adjacent coefficients, see Gertheiss and Tutz (2010) and Tutz and Gertheiss (2014). However, Lasso-type penalties come with shrinkage effects that depend on the coefficients' absolute values (Fan and Li, 2001). As a consequence, there are often strong shrinkage effects for large (differences of) coefficients while small (differences of) coefficients are estimated to be non-zero. When the focus is on the fusion and the selection of categories, one wants to avoid such effects. To enhance Lasso-type penalties, Zou (2006) proposes adaptive weights; each penalty term is weighted by its inverse maximum likelihood (ML) estimate. It yields asymptotically normal and consistent model selection. However, the quality of the adaptive weights depends on the quality of the ML estimate that can be poor.

As an alternative, we propose L_0 penalization for categorical effects; where the L_0 'norm' is defined as the number of non-zero entries in a vector. As Bondell and Reich (2009) or Gertheiss and Tutz (2010), we consider differences of coefficients; but instead of the absolute value, the L_0 norm is applied to the differences of coefficients. The difference between unordered and ordered factors is taken into account by using all pairwise differences or only adjacent differences. Computational issues are met by local quadratic approximations. The optimization problem is related to model selection with information criteria like the Akaike information criterion (AIC; see for example, Bozdogan, 1987) or the Bayesian information criterion (BIC; Schwarz, 1978). As the proposed penalty allows for the fusion of categories, it extends this approach. L_0 penalization is an established approach in some fields of statistics: it is applied to wavelets (Antoniadis and Fan, 2001) and to signals (see Lu and Zhang, 2010, Rippe *et al.*, 2012).

Moreover, minimizing (approximations of) constrained L_0 terms is employed to find sparse representations of signals; see, for example, Donoho and Elad (2003), Wipf and Rao (2005), Mancera and Portilla (2006) or Ge *et al.* (2011).

The article is organized as follows: Section 2 motivates L_0 penalization for categorical effects in generalized linear models. In Section 3, we introduce the method; computational issues, the relation to best subset selection and some generalizations are discussed. Section 4 investigates the numerical properties of the proposed method. In Section 5, the unemployment rates in Germany between 2005 and 2010 are analyzed. We investigate which state-specific intercepts are clustered in a model with a global temporal trend.

2 Framework and L_1-Type Fusion Penalties

In what follows, we assume a generalized linear model (GLM) with response y_i for observation i, $i = 1, \ldots, n$. As a start, we consider only one categorical covariate $\boldsymbol{x}$ with levels $0, \ldots, k$. Let the rows of the design matrix $\boldsymbol{X}$ be given by $\boldsymbol{x}_i^T = (1,\ x_{i1}, \ldots, x_{ik})$ with $x_{ir} = 1$ if x_i takes the value r and $x_{ir} = 0$ otherwise, $r = 1, \ldots, k$. This representation refers to dummy coding with category 0 as reference category, $\beta_0 = 0$. The corresponding predictor is defined as $\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_{int}, \beta_1, \ldots, \beta_k)^T$ is

the coefficient vector and β_{int} denotes the intercept. For the response $y_i|x_i$, a simple exponential family with log-likelihood $l_n(\boldsymbol{\beta})$ is assumed:

$$l_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{y_i \vartheta_i(\mu_i) - b(\vartheta_i(\mu_i))}{\varphi} + c(y_i, \varphi),$$

where $\vartheta_i(\mu_i)$ denotes the natural parameter, $b(\cdot)$ is a specific function corresponding to the type of the exponential family, $c(\cdot)$ is the log-normalization constant and φ the dispersion parameter (compare Fahrmeir and Tutz, 2001). The observations y_i are assumed to be conditionally independent. Response and predictor are linked by the response function $h(\eta_i)$ which is twice continuously differentiable with $\det(\partial h/\partial \eta_i) \neq 0 \ \forall i$. That is, we assume:

$$\mu_i = \mathbb{E}(y_i|x_i) = h(\eta_i). \tag{2.1}$$

For more details on GLMs, see, for example, Fahrmeir and Tutz (2001). Estimates $\hat{\boldsymbol{\beta}}$ are obtained by minimizing the negative log-likelihood $l_n(\boldsymbol{\beta})$. Accounting for a penalty, the objective function is defined as:

$$\mathcal{M}_{pen}(\boldsymbol{\beta}) = -l_n(\boldsymbol{\beta}) + \lambda \cdot P(\boldsymbol{\beta}), \tag{2.2}$$

where $P(\boldsymbol{\beta})$ denotes the penalty and where λ is a tuning parameter. The larger λ is, the stronger is the impact of the penalty. For $\lambda = 0$, the ML estimate is obtained. The choice of the penalty $P(\boldsymbol{\beta})$ is crucial. The Lasso (Tibshirani, 1996) penalizes the absolute values of coefficients and enforces variable selection. One obtains sparse but shrunken estimates. For dummy-coded categorical predictors, this is not the best choice; setting parameters to zero corresponds to the fusion with the reference category which can be chosen arbitrarily. Even though this problem can be handled by coding the categorical covariates differently—for example, as the deviation from a mean level ('effect coding') or as the deviation from adjacent categories ('split coding')—penalties that contain differences of parameters as proposed by (Tibshirani *et al.* (2005), Bondell and Reich (2009) or Gertheiss and Tutz (2010), are a common choice. They encourage the fusion of coefficients and thus, of categories irrespectively of the coding, and they allow one to fuse coefficients subject to more than k constraints. However, fusion-type penalties come along with some problems:

For an *ordered* categorical predictor, fusion-type penalties consider the differences of parameters that refer to adjacent categories, including the reference category 0. The corresponding Lasso-type penalty has the form:

$$P(\boldsymbol{\beta}) = \sum_{r=1}^{k} |\beta_r - \beta_{r-1}|. \tag{2.3}$$

However, the penalty does not always enforce fusion efficiently. If coefficients are ordered, for example in the form $0 = \beta_0 \leq \beta_1 \leq \ldots \leq \beta_k$, and if one is close to the true values, that is, in the range where the estimated parameters are ordered, the effective

penalty is $P(\boldsymbol{\beta}) = \sum_{r=1}^{k} |\beta_r - \beta_{r-1}| = |\beta_k - \beta_0| = |\beta_k|$. That means, the approach basically penalizes the range of the coefficients. The problem is even more obvious in an orthonormal linear model with one ordered predictor and without an intercept—that is, $\boldsymbol{X}^T\boldsymbol{X}$ is the identity matrix $\mathbb{I}_{(k+1)\times(k+1)}$. Situations like this, are for example, typical for models with categorical effect modifiers or models with group specific intercepts. In this cases, one can derive an explicit solution of the objective function (2.2) with penalty (2.3):

Proposition 1. Assume a penalized linear model with orthonormal design; that is $\boldsymbol{X}^T\boldsymbol{X} = \mathbb{I}_{(k+1)\times(k+1)}$ *where* $\boldsymbol{X} \in \mathbb{R}^{(k+1)\times(k+1)}$ *denotes the design matrix without an intercept and where* $\mathbb{I}$ *denotes the identity matrix. Let the ML estimates be ordered* $\hat{\beta}_0^{ML} < \ldots < \hat{\beta}_k^{ML}$ *and employ penalty (2.3) with a fixed penalty parameter* λ, $\lambda \geq 0$. *Then for j,* $\hat{\beta}_j^{ML} < \bar{\beta}^{ML}$, $\bar{\beta}^{ML} = \frac{1}{k+1}\sum_{j=0}^{k} \hat{\beta}_j^{ML}$, *one obtains:*

$$\hat{\beta}_j = \min\left\{\bar{\beta}^{ML},\ \max\{\hat{\beta}_l^{ML}, \hat{\beta}_j^{ML}\} + \frac{(\lambda - \lambda_l)I_{(l\geq j)}}{2(l+1)}\right\},$$

where $l = \max_{l=0,\ldots,k}(\lambda_l < \lambda)$, $\lambda_l = \sum_{u=1}^{l} 2u\left|\hat{\beta}_u^{ML} - \hat{\beta}_{u-1}^{ML}\right|$, *and with indicator function I. For* $\hat{\beta}_j^{ML} \geq \bar{\beta}^{ML}$, *one obtains analogously*

$$\hat{\beta}_j = \max\left\{\bar{\beta}^{ML},\ \min\{\hat{\beta}_l^{ML}, \hat{\beta}_j^{ML}\} - \frac{(\lambda - \lambda_l)I_{(k-l\geq j)}}{2(l+1)}\right\},$$

with $\lambda_l = \sum_{u=l}^{k-1} 2(k-u)\left|\hat{\beta}_{u+1}^{ML} - \hat{\beta}_u^{ML}\right|$ *and l as before.*

The proof of Proposition 1 is given in Supplement A. The structure of the explicit estimate reveals that the coefficients of the outer categories are always merged first. There is no shrinkage for coefficients that are not yet fused with one of the outer categories—no matter how close the corresponding ML estimates are. For the minimal penalty parameter that causes the fusion of all coefficients, the estimate of all coefficients is equal to $\bar{\beta}^{ML}$. For a fixed value of λ, the mean of the penalized estimate equals $\bar{\beta}^{ML}$ in the assumed setting. Similar results in the context of signal processing can be found in Pollak *et al.* (2005).

The left panel of Figure 1 shows the (exact) coefficient path of an exemplary model with $k = 7$. One can see that a coefficient β_r is not fused with any other β_s, $r \neq s$, unless β_r is fused with one of the outer coefficients β_0, β_k. The right panel of Figure 1 shows the same situation but the coefficient path is obtained with an L_0 norm instead of the L_1 norm in penalty (2.3). Categories with similar effects are fused—no matter which position they take in the order of ML estimates. This is the main motivation to consider L_0 penalties as an alternative when investigating categorical predictors.

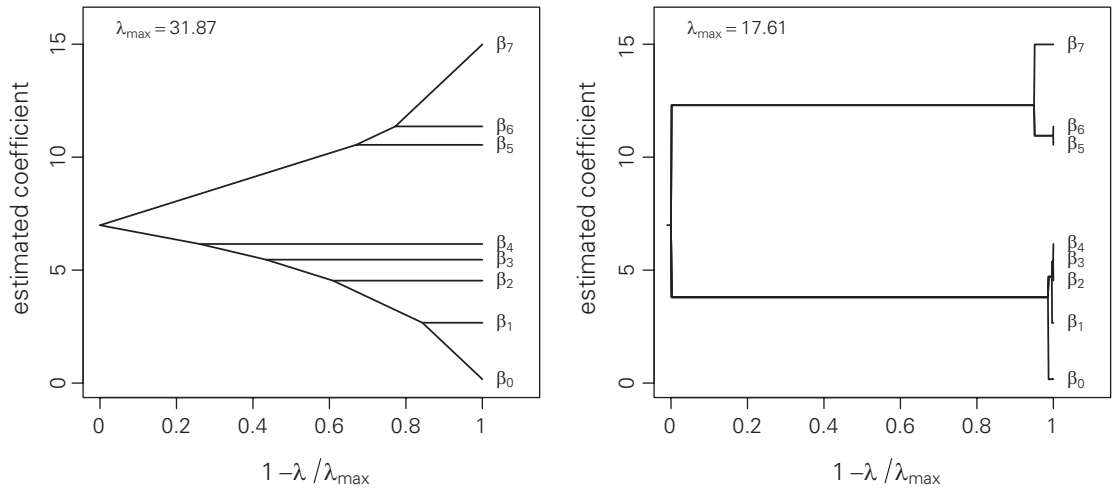

Figure 1 Coefficient paths for an orthonormal linear model with one categorical predictor ($k = 7$), dummy coded without an intercept. In the left panel, penalty (2.3) is applied; in the right panel, the L_1 norm in penalty (2.3) is replaced by the L_0 norm.
Source: Authors' own.

For a *nominal* categorical predictor, L_1 fusion-type penalties consider all pairwise differences of coefficients:

$$P(\boldsymbol{\beta}) = \sum_{r>s\geq 0} |\beta_r - \beta_s|. \tag{2.4}$$

Assume a fixed value of the tuning parameter λ and let $\beta_{(1)}, \ldots, \beta_{(k)}$ denote the (arbitrary) ordering of the solution. Then, a short transformation (see Supplement A) shows that $\sum_{r>s} |\beta_{(r)} - \beta_{(s)}| = \sum_{r=1}^{k} w_{(r)} |\beta_{(r)} - \beta_{(r-1)}|$, where $w_{(r)} = r(k - r + 1)$. For the 'outer' differences $r \in \{1, k\}$, $w_{(r)} = k$; for medium values of r, the weights $w_{(r)}$ are higher. That is, penalty (2.4) can be represented as a weighted version of penalty (2.3). The issues for nominal predictors are essentially the same as for ordered predictors. Similar to Proposition 1, one can show that the slopes of the coefficient path depend on the order of the corresponding ML estimate—even though not only the 'outer' coefficients are fused.

Efficiency of L_1-penalized estimates can be improved by using adaptive weights (Zou, 2006) that weigh each penalty term by its inverse ML estimate. This results in heavy weights on penalty terms with small ML estimates and in small weights on penalty terms with large ML estimates. When the adaptive weights of Zou (2006) are combined with fusion-type penalties, there is an incentive to fuse categories that have close ML estimates and one obtains asymptotically normal and consistent results (see for example, Gertheiss and Tutz, 2010; Oelker *et al.*, 2014). However, adaptive weighting requires ML estimates; its quality depends on the quality of the ML estimates.

3 L_0-Type Fusion Penalties

In what follows, the fusion of categories is enforced by penalizing differences of coefficients as in the approaches discussed previously, but to overcome the drawbacks of Lasso-type penalties, the L_0 norm is employed. For an *ordered* predictor, we propose

$$P_{ord}(\boldsymbol{\beta}) = \sum_{r=1}^{k} \|\beta_r - \beta_{r-1}\|_0, \tag{3.1}$$

where $\|\cdot\|_0 = |\cdot|^0$ and where we define $0^0 = 0$. In contrast to Lasso-type penalties, it does not matter whether a difference is small or huge; the penalty is reduced only if one of the differences equals zero. As a consequence, when two different values of λ, for example $\lambda_1 > \lambda_2$, yield solutions with the same set of zero and non-zero differences, $P_{ord}(\hat{\boldsymbol{\beta}}_{\lambda_1}) = P_{ord}(\hat{\boldsymbol{\beta}}_{\lambda_2})$ holds. The set of zero/non-zero differences changes for specific thresholds.

When the predictor $\boldsymbol{x}$ is *nominal*, an appropriate coefficient profile does not only relate to the coefficients of adjacent categories but to the comparison of all coefficients. The penalty considers all pairwise differences of coefficients,

$$P_{nom}(\boldsymbol{\beta}) = \sum_{r>s\geq 0} \|\beta_r - \beta_s\|_0. \tag{3.2}$$

Penalty (3.2) is more complex; with k levels, there are $k(k+1)/2$ pairwise differences; but apart from that, the effect of the penalty is the same as for ordered predictors.

Note that for large values of the tuning parameter λ, the coefficients $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k$ are set to zero as the differences in penalties (3.1) and (3.2) include the difference to the reference category $\beta_0 = 0$. As it will be seen in Section 5, it can be useful to weight the penalty terms. To this end, we introduce general weights w_r, $w_{r,s}$ respectively, and use the modified penalties:

$$P_{ord}(\boldsymbol{\beta}) = \sum_{r=1}^{k} w_r \|\beta_r - \beta_{r-1}\|_0 \quad \text{and} \quad P_{nom}(\boldsymbol{\beta}) = \sum_{r>s\geq 0} w_{r,s} \|\beta_r - \beta_s\|_0. \tag{3.3}$$

To enhance the performance, we will, for example, combine the L_0 approach with adaptive weights as employed for the L_1 penalties in Section 4.1. When analyzing the unemployment rates in Germany in Section 5, the weights allow one to account for the spatial structure of the federal states. Therefore, we define the weights $w_{r,s}$ as an indicator for a common border of two states – such that we obtain a coefficient profile that is consistent with geography.

3.1 Computational Issues

In the literature, there is a wide range of strategies to handle optimization problems that contain L_0 norms. In order to represent for example, signals sparsely, the

objective $\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_0$ subject to $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$ has to be optimized. With some assumptions on $\boldsymbol{X}$ and assuming that there is a sufficiently sparse representation of $\boldsymbol{y}$, Donoho and Elad (2003) find this representation by solving a convex optimization problem instead. Wipf and Rao (2005) derive a method based on sparse Bayesian learning including local optimality conditions to solve the same problem. In the framework of wavelets, the L_0 norm acts as penalty. The objective $\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_0$ is, for example, solved by penalty decomposition methods that are based on rank optimization procedures (see Lu and Zhang, 2010, 2013). Rippe *et al.* (2012) and Johnson (2013) minimize $\sum_{i=1}^{n}(y_i - \beta_i)^2 + \lambda \sum_{i=2}^{n} \|\beta_i - \beta_{i-1}\|_0$ to smooth segmented observations $y_1, \ldots, y_n$. While Johnson (2013) proposes a dynamic programming algorithm, Rippe *et al.* (2012) solve the problem iteratively employing a weighted Ridge penalty.

In order to minimize the objective function (2.2), we propose to approximate the L_0 norm by a modified logistic function and to derive a quasi Newton method for the approximated objective function. In detail, the L_0 norm is approximated by:

$$\|\xi\|_0 \approx \frac{2}{1 + \exp(-\gamma\sqrt{\xi^2 + c})} - 1, \tag{3.4}$$

where γ is a relatively large scalar and where c is a small, positive constant. Figure 2 gives some illustration: the circles denote the L_0 norm for a scalar argument ξ. The continuous line denotes the approximation of the L_0 norm. For $\gamma \to \infty$ and $c \to 0$, the approximation approaches the L_0 norm.

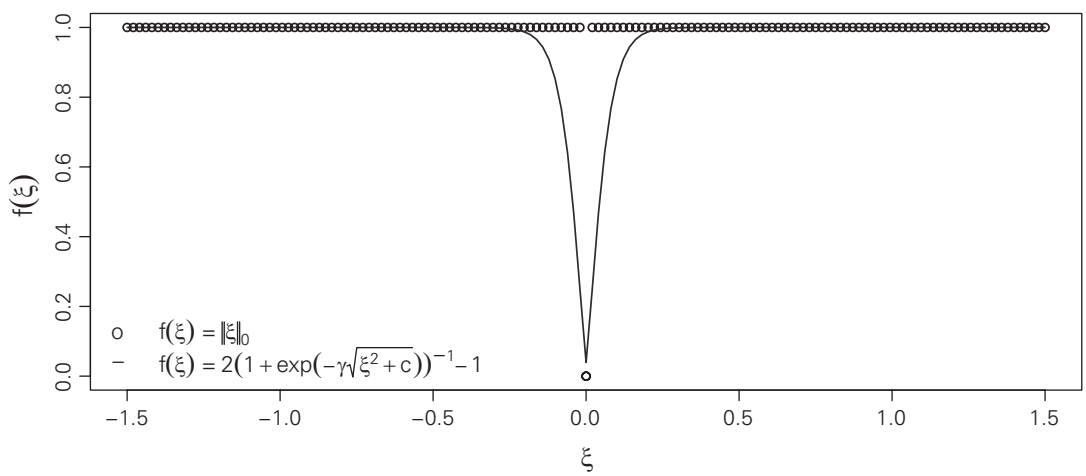

Figure 2 Graphical illustration of the approximation of the L_0 norm. $\gamma = 25$, $c = 10^{-5}$.
Source: Authors' own.

To obtain a penalized iteratively re-weighted least squares (PIRLS) algorithm, in addition to approximation (3.4), we employ a local quadratic approximation if $\hat{\boldsymbol{\beta}}_{(k+1)}$ is close to $\hat{\boldsymbol{\beta}}_{(k)}$ as proposed byFan and Li (2001) and as described

by Oelker and Tutz (2013). It allows for derivatives of the approximated objective that fit exactly in the framework of known PIRLS algorithms. Sketching the idea shortly, penalties (3.1) and (3.2) are rewritten as $P(\boldsymbol{\beta}) = \sum_{l=1}^{L} \left\| \boldsymbol{a}_l^T \boldsymbol{\beta} \right\|_0$, where vectors $\boldsymbol{a}_l$ build all needed differences of coefficients. For a nominal predictor $\boldsymbol{x}$ with five levels including the reference category, where all pairwise differences are considered, one obtains, for example,

$$\boldsymbol{A}^{nom} = (\boldsymbol{a}_1 \cdots \boldsymbol{a}_L) = \begin{pmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 0 & -1 & -1 & 0 \\ 0 & 1 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 0 & 1 & -1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}.$$

The 'diagonal part' of $\boldsymbol{A}^{nom}$ gives the differences with the reference category β_0. Finally, starting with an initial value $\hat{\boldsymbol{\beta}}_{(0)}$, one obtains the following iterative update for the current estimate $\hat{\boldsymbol{\beta}}_{(k)}$:

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

where matrix $\boldsymbol{W}_{(k)}$ denotes weights and $\tilde{\boldsymbol{y}}_{(k)}$ denotes pseudo-observations like in usual GLMs; $\boldsymbol{W}_{(k)} = \boldsymbol{D}_{(k)} \boldsymbol{\Sigma}_{(k)}^{-1} \boldsymbol{D}_{(k)}$, $\boldsymbol{D}_{(k)} = \mathrm{diag}(\partial h(\eta_i(\hat{\boldsymbol{\beta}}_{(k)}))/\partial \boldsymbol{\eta})$, $\boldsymbol{\Sigma}_{(k)} = \mathrm{diag}(\sigma_i^2(\hat{\boldsymbol{\beta}}_{(k)}))$, $\tilde{\boldsymbol{y}}_{(k)} = \boldsymbol{D}_{(k)}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_{(k)}) + \boldsymbol{X}\hat{\boldsymbol{\beta}}_{(k)}$ and $\boldsymbol{\mu}_{(k)} = h(\boldsymbol{\eta})$. The parameter ν is a step length parameter that usually equals 1; if $\nu < 1$, it allows one to control the algorithm's convergence and to stabilize the algorithm in the case of marginal solutions. Approximating the L_0 norm, smaller values of ν seem to be a good choice. Matrix $\boldsymbol{A}_\lambda$ basically contains the derivatives of the approximated L_0 norm:

$$\boldsymbol{A}_\lambda = \lambda \sum_{l=1}^{L} \left(\frac{1}{1 + \exp(-\gamma |\boldsymbol{a}_l^T \boldsymbol{\beta}_{(k)}|)} \right) \left(1 - \frac{1}{1 + \exp(-\gamma |\boldsymbol{a}_l^T \boldsymbol{\beta}_{(k)}|)} \right) \cdot \frac{2 \cdot \gamma \cdot \boldsymbol{a}_l \boldsymbol{a}_l^T}{\sqrt{(\boldsymbol{a}_l^T \boldsymbol{\beta}_{(k)})^2 + c}}.$$

The algorithm is terminated when $|\hat{\boldsymbol{\beta}}_{(k+1)} - \hat{\boldsymbol{\beta}}_{(k)}| / |\hat{\boldsymbol{\beta}}_{(k)}| \leq \epsilon$, for a fixed $\epsilon > 0$. Concerning the tuning, the constant $c > 0$ guarantees differentiability, γ determines the steepness of the logistic function. Both parameters have to be determined subject to the scale of the (coded) covariate $\boldsymbol{x}$. However, in penalized regression models, the covariates are usually scaled and/or standardized ($\mathbb{E}(\boldsymbol{x}) = 0$ and $\mathbb{V}(\boldsymbol{x}) = 1$). In such settings, $c = 10^{-5}$ works quite well in our experience. When γ is sufficiently large, the coefficients paths look like step functions; the steps occur when the coefficients are merged and as the shift of the estimates is relatively large compared to the change of λ. As long as the approximation is close enough to the L_0 norm, the concrete choice of γ has no major impact on the result's quality. However, for different tuning parameters, the scale of λ changes; if γ is too large, there may be convergence problems.

The structure of the objective function is not trivial. As the penalty is not convex, there is no guarantee that the proposed algorithm finds the global minimum of the objective function. However, the results are very plausible. Given that the tuning parameter λ is in a realistic range, the results for different initial values do not differ essentially in a majority of cases. We recommend $\hat{\boldsymbol{\beta}}_{(0)} = \mathbf{0}^T$ or to combine the default approach of the R function `glm()` ($\boldsymbol{\mu}_{(0)} = \boldsymbol{y}$ for the loss function) with the initial value $\hat{\boldsymbol{\beta}}_{(0)} = \mathbf{1}^T$ for the approximation of the penalty (referred to as 'default set of initial values'). Furthermore, the results for different initial values should be checked. Comparisons with a simulated annealing algorithm (Xiang *et al.*, 2013) that is appropriate for complex optimization problems, show that the deviations of the PIRLS algorithm from the simulated annealing are small for the relevant range of λ. The L_0 approach of Rippe *et al.* (2012) for signals works with a different approximation but also with a PIRLS algorithm and obtains similar results. Fan and Li (2001) propose to approximate the SCAD penalty that has a similar curvature, by a PIRLS algorithm; comparisons with the exact estimate in an orthonormal setting approves this procedure.

Oelker and Tutz (2013) give detailed information on the local quadratic approximation. The algorithm is implemented in the R package `gvcm.cat` (Oelker, 2013; R Core Team, 2013).

3.2 The General Case with Multiple Predictors

So far, we assumed that there is only one predictive factor $\boldsymbol{x}$. Of course, this is not the standard case and, in what follows, we assume that there are p nominal and/or ordered predictive factors $\boldsymbol{x}_j$ with $k_j + 1$ levels each. The design matrix is still denoted by $\boldsymbol{X}$, but now $\boldsymbol{X} \in \mathbb{R}^{n \times q}$, where $q = 1 + \sum_{j=1}^{p} k_j$; $\boldsymbol{X}$ contains p dummy coded predictors and an intercept. The according penalty is defined as:

$$P(\boldsymbol{\beta}) = \sum_{j=1}^{p} J_j(\boldsymbol{\beta}_j),$$

where J_j equals penalty (3.1) for ordered factors and penalty (3.2) for nominal factors. The parameter $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jk_j})^T$ denotes the vector of coefficients linked to the j-th covariate. The computational issues are not affected by this generalization.

However, the tuning should be adjusted as the penalty terms of several factors with different numbers of levels and measured on different scales (nominal/ordered) should be comparable. Bondell and Reich (2009) argue for example, that there is a bijective relation between the standardization of the data and weighting the penalty terms if one penalty term relates to one covariate and if the penalty is a norm; for example, in case of the Lasso for p continuous covariates. Bondell and Reich (2009) transfer this idea to penalties that contain pairwise differences related to nominal covariates; they propose to weigh each difference by $k_j^{-1}\sqrt{n_j^{(l)} + n_j^{(m)}}$, where $n_j^{(l)}$ denotes the number of

observations on level l of covariate j. For ordered covariates, $\sqrt{n_j^{(l)} + n_j^{(m)}}$ is appropriate; see, Gertheiss and Tutz (2010). The weights consider the number of observations per level and the number of differences in the penalty. They can be combined with the L_0 penalty easily. Depending on the concrete content, different weighting schemes can be reasonable; alternatively, one could for example, think of $J_j(\boldsymbol{\beta}_j) = const. \forall j$.

3.3 L_0 Penalization and Information Criteria

There is a relation between L_0 penalization and model selection by information criteria like the AIC or the BIC: one minimizes

$$\mathcal{IC}(\boldsymbol{\beta}) = -l_n(\boldsymbol{\beta}) + \lambda \cdot \text{df(model)}, \tag{3.5}$$

where $\lambda_{AIC} = 1$ for the AIC and $\lambda_{BIC} = \log(n)/2$ for the BIC. The degrees of freedom df(model) are the number of influential parameters in the model and therefore equal $\sum_{j=0}^{p} \|\beta_r\|_0$. Let us first consider a model with binary predictors only. Then the proposed L_0 penalty has the form $P_{bin}(\boldsymbol{\beta}) = \sum_{j=1}^{p} \|\beta_j\|_0$. Hence, it holds that

$$P_{bin}(\boldsymbol{\beta}) + 1 = \text{df(model)}.$$

That is, when the tuning parameter of the proposed L_0 approach is fixed to the values λ_{AIC} or λ_{BIC}, the objectives of the L_0 approach and of model selection based on information criteria AIC/BIC coincide. However, the computational approach differs: for model selection based on information criteria, unconstrained models with all possible subsets of coefficients are compared by criterion (3.5). The L_0 approach optimizes the approximated, constrained objective and does it without subsets.

Selection problems are much more complex if one has p categorical covariates with $k_j + 1$ levels each, because then, in addition to simple selection of relevant variables, one also wants to investigate which categories of the categorical predictor have to be distinguished. In best subset selection with categorical predictors, all models that can be built by the fusion of categories must be considered as candidate models. It is well known from cluster theory that the number of such candidate models increases strongly with the number of categories per predictor (Jain and Dubes, 1988). For a nominal covariate with three coefficients $\beta_1, \beta_2, \beta_3$, this already results in 15 combinations: $\{(), \{\beta_1\}, \{\beta_2\}, \{\beta_3\}, \{\beta_1, \beta_2\}, \{\beta_1, \beta_3\}, \{\beta_2, \beta_3\}, \{\beta_1 = \beta_2\}, \{\beta_1 = \beta_3\}, \{\beta_2 = \beta_3\}, \{\beta_1 = \beta_2, \beta_3\}, \{\beta_1 = \beta_3, \beta_2\}, \{\beta_2 = \beta_3, \beta_1\}, \{\beta_1, \beta_2, \beta_3\}, \{\beta_1 = \beta_2 = \beta_3\}\}$. Thus, model selection based on candidate models as it is used by AIC and BIC, is restricted to cases with very few categories in the predictor(s).

In this general case, the degrees of freedom are defined as the number of coefficients in a model that are unlike and unequal to zero. They are given by $\text{df(model)} = 1 + \sum_{j=1}^{p} \sum_{r=1}^{k_j} \|\beta_{jr}\|_0 \prod_{s<r} \|\beta_{jr} - \beta_{js}\|_0$—which is unequal to the proposed penalties. However, the proposed penalties can be applied to the same situations as best subset selection for categorical covariates. In contrast to best subset selection, the proposed penalties do not need candidate models and—as we will see

later on—they are nevertheless feasible for more complex models. As the tuning parameter λ can be varied, information on the order of fusions of coefficients is obtained. Hence, the proposed penalty is not only an attractive alternative to Lasso-type penalties, but as well an alternative to model selection based on information criteria.

4 Illustration and Numerical Experiments

In this section, we investigate some aspects of the proposed approach by numerical experiments. We start with an illustrative example followed by some experiments on the estimation accuracy and on the clustering/selection performance when the tuning parameter is chosen according to cross-validation criteria.

4.1 Illustrative Example

We consider a linear model with the two ordered predictors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Both predictors have four levels and are drawn from a multinomial distribution with equal probabilities for each level. The response is Gaussian; $\boldsymbol{\beta}^{true} = (\beta_{int}, \beta_{12}, \beta_{13}, \beta_{14}, \beta_{22}, \beta_{23}, \beta_{24})^T = (2, 0.8, 0.6, 0.4, -0.6, -0.6, -0.4)^T$ and $\text{Var}(y_i|\boldsymbol{X}) = 1\ \forall i$. All levels of $\boldsymbol{x}_1$ have different impact on the response whereas the levels 2 and 3 of $\boldsymbol{x}_2$ are influential but do not need to be distinguished. We generate $n = 100$ observations and consider two models: the proposed L_0 penalty, that is, penalty (3.1) for the two ordered factors, and a penalty with the same differences but with the L_1 norm instead of the L_0 norm. There is one global tuning parameter λ in both models, which is chosen by a generalized cross-validation criterion (GCV) as for example defined in O'Sullivan *et al.* (1986) and as used in the R package `mgcv` (Wood, 2011). The GVC criterion is given by $\text{GCV} = n \cdot \text{dev}/(n - \text{df(model)})^2$, where the deviance is defined as $\text{dev}(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = -\varphi(l_n(\boldsymbol{y}, \hat{\boldsymbol{\mu}}, \varphi) - l_n(\boldsymbol{y}, \boldsymbol{y}, \varphi))$, where $l_n(\cdot)$ denotes the log-likelihood; df(model) is estimated by the trace of the hat matrix $\boldsymbol{W}_{(k*)}^{T/2}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}_{(k*)}\boldsymbol{X} + \boldsymbol{A}_\lambda)^{-1}\boldsymbol{X}^T\boldsymbol{W}_{(k*)}^{1/2}$ of the final iteration $(k*)$ of the PIRLS algorithm. Figure 3 (top) shows the resulting coefficients paths and the GCV score for the L_0 penalization with tuning parameters $c = 10^{-5}$, $\nu = 0.05$, $\gamma = 60$. At the very right end of the coefficient paths, the ML estimates are displayed; at the very left end, the estimate for the minimal tuning parameter λ that gives maximal penalization is shown. As there are no shrinkage effects, for some values of λ, the estimates are the same. The coefficient path looks like a horizontal tree. The GCV score in the right panel is a step function that jumps when the estimate changes. This happens because the GVC criterion does not depend on the tuning parameter λ; for identical estimates $\hat{\beta}_{\lambda_1} = \hat{\beta}_{\lambda_2}$, the GCV scores are the same. Hence, the function has no clear minimum; we choose the maximal value of λ with minimal GCV score as λ_{CV}. The optimal model is marked by a dotted line at $\lambda_{CV} = 0.36$. In this model, levels 2 and 3 of predictor $\boldsymbol{x}_2$ have the same impact on the response ($\hat{\beta}_{22} = \hat{\beta}_{23} = -0.41$, $\hat{\beta}_{24} = -0.40$). Levels 3 and 4 of predictor $\boldsymbol{x}_1$ are falsely fused ($\hat{\beta}_{13} = \hat{\beta}_{14} = 0.30$).

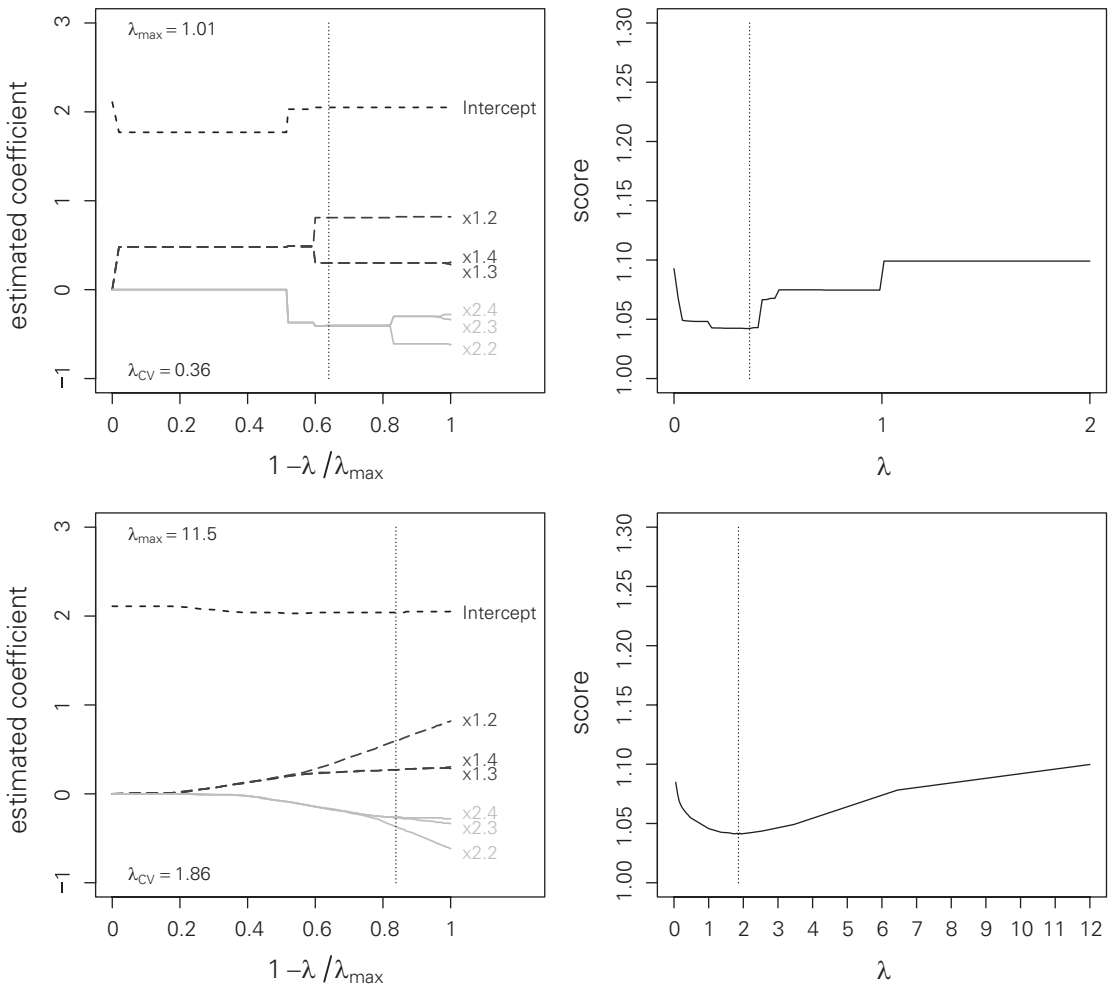

Figure 3 Illustration of L_0 penalization (top) and L_1 penalization (bottom) in a linear model with two ordered effects. The left panel shows the resulting coefficient path. The right panel shows the corresponding GCV score. The tuning for the approximation of the L_0 norm is $\gamma = 60$, $c = 10^{-5}$ and $\nu = 0.05$. The tuning for the approximation of the L_1 norm is $c = 10^{-5}$. In all panels, the dotted line marks the optimal models.
Source: Authors' own.

The estimates of the remaining effects are: $\hat{\beta}_{int} = 2.05$, $\hat{\beta}_{11} = 0.81$. In the optimal model with the same differences but with a L_1 norm in the penalty, two coefficients are falsely fused ($\hat{\beta}^{L1} = (2.04, 0.60, 0.27, 0.27, -0.37, -0.27, -0.26)^T$). Figure 3 (bottom) shows the coefficient paths and the GCV score accordingly. In contrast to the L_0 penalty, the path is characterized by steady shrinkage effects; the GCV score is a continuous function with a clear minimum. For the L_1 penalty, the shrinkage effect is slightly bigger as for the L_0 penalty: the sum of squared errors are $\widehat{\text{SSE}}_{L_1} = \sum_{i=1}^{7}(\hat{\beta}_i^{L1} - \beta_i^{true})^2 = 0.3488 \; > \; 0.1748 = \widehat{\text{SSE}}_{L_0}$.

4.2 Choice of the Tuning Parameter λ

As for every penalized approach, the choice of the tuning parameter λ is a crucial issue for L_0 penalization. In the illustrative example, we employ a GCV criterion that requires an estimate of df(model) and that gives concurrent jumps in the coefficient paths and in the GCV score. An alternative, frequently used approach to choose the tuning parameter is *K*-fold cross-validation with the predictive deviance as loss criterion. *K*-fold cross-validation relies on models estimated on different training/test data sets for different values of λ. As we approximate the L_0 norm which is not continuous, the estimate can change abruptly even if the tuning varies only slightly. The values of λ at which the estimate changes, will not be the same for all training data set. Thus, depending on the chosen folds, for *K*-fold cross-validation, the overall cross-validation score can be quite wiggly. Therefore, we compare the performance of the GCV criterion and of 5-fold cross-validation in Section 4.3.

4.3 Performance

To evaluate the overall performance of the penalties, we consider the estimation accuracy, the prediction accuracy and the error rates of the selection and clustering process. The estimation accuracy is assessed by the squared errors in terms of coefficients: $\widehat{\text{SSE}} = \frac{1}{q}\sum_{j=1}^{q}(\beta_j^{true} - \hat{\beta}_j)^2$, where $\boldsymbol{\beta}^{true}$ denotes the vector of true coefficients and $\hat{\boldsymbol{\beta}}$ the estimate of the current simulation run. The median of all squared errors is the robust estimate for the mean squared error (MSE) of a method. The prediction accuracy is assessed by the predictive deviance and referred to as MSEP. To judge the model selection process, we consider the selection and the clustering of coefficients separately; the selection of coefficients refers to the coefficients ($\beta_{jl} = 0$) whereas the clustering process refers to the differences of coefficients ($\beta_{jl} = \beta_{jm}$). We distinguish between false positive rates (fraction of truly zero coefficients that are set to non-zero, FP) and false negative rates (fraction of truly non-zero coefficients that are set to zero, FN). We focus on four settings. A setting similar to the illustrative example of Section 4.1 is considered in more detail, it is referred to as *G3*. In addition, a setting with Gaussian response and 50 nominal predictors is investigated (*G50*). Settings with Poisson distributed and with binomial distributed response are analyzed (*P8*, *B8*). For each setting, 100 replications are considered; for each replication, we compute the ML estimate, the estimate obtained with the L_1 penalty, the estimate obtained with the adaptively weighted L_1 penalty and the estimate obtained with the proposed L_0 penalty. Moreover, we combine the proposed L_0 penalty with the same adaptive weights as employed for the adaptively weighted L_1 penalty. For all penalized approaches, the tuning parameter is chosen by the GCV criterion and by 5-fold cross-validation with the predictive deviance as loss criterion (CV). In addition, a model selection method for categorical predictors is implemented. The method is based on the information criteria AIC and BIC and compares not only all possible subsets of coefficients, but as well all possibilities to fuse different numbers of levels of a categorical predictor. This method is referred to as AIC, BIC respectively.

For all settings, the tuning parameters $c = 10^{-5}$ and $\nu = 0.05$ are fixed; we employ the default set of initial values. γ is empirically chosen as described in Section 3.1 ($\gamma_{G3} = 20$, $\gamma_{G50} = 10$, $\gamma_{P8} = 20$, $\gamma_{B8} = 10$).

In the setting *G3*, there are three nominal covariates with four levels each; $\boldsymbol{\beta}^{true} = (\beta_{int}, \beta_1^T, \beta_2^T, \beta_3^T)^T = (1,\ (0, -1.5, -1.5),\ (0, 0, 2),\ (-3, -3.5, 4))^T$; in each replication, $n = 50$ observations are generated. The upper left panel of Figure 4 shows the boxplots of the squared errors. Apart from some outliers, the estimation accuracy of all considered approaches is approximately the same. This is typical: in standard situations, (adaptive) L_1 and L_0 penalization do not show substantial differences. However, as seen in Table 1, the L_0 approach produces more parsimonious and interpretable models. The methods based on information criteria are characterized by the highest FN rates. Comparing the L_1 penalization with and without adaptive weights, the adaptive weighting improves the FP rates substantially. Comparing the adaptively weighted L_1 and the adaptively weighted L_0, the clustering performance is substantially enhanced by the L_0 penalty. Again, this is typical: with the L_0 penalty, the false positive rates are lower while it can happen that the false negative rates increase slightly in comparison with L_1 penalization.

In setting *G50*, there are 50 nominal covariates with four levels each; $\boldsymbol{\beta}^{true}$ is a vector of length 151, it contains 72 non-influential coefficients and 54 truly different effects; $n = 500$. In contrast to setting *G3*, for this and the following settings, model selection based on information criteria is not feasible anymore on a default computer. The lower panel of Figure 4 depicts the squared errors of setting *G50*. It stands out that the L_0 penalized models perform slightly better than the pure L_1 penalized approaches. Regarding the FP/FN rates in Table 1, it is even more obvious that the proposed L_0 approach generates more parsimonious models. Overall, the approach 'L_0, CV' performs the best.

In setting *P8*, there are four influential nominal covariates; $\boldsymbol{\beta}^{true} = (\beta_{int}, \beta_1^T, \beta_2^T, \beta_3^T)^T = (2,\ (0, -1.2, -1.2),\ (1.4, 1.4, 0),\ (0.4, 0.6, 0.8),\ (-0.7, -1, -1.3))^T$. We assume four more non-influential, nominal covariates which are to be detected. For an observation i, the assumed predictor is $\eta_i^{model} = \beta_{int} + \sum_{j=1}^{8} x_{ij}^T \beta_j$; $n = 100$. In Figure 5, the squared errors and the PMSE of the penalized methods are distinctively smaller than of the ML estimates. In Table 1, the L_0 approach reduces the false positive rates even more as the adaptively weighted L_1 penalty. However, for the L_0 approach, the CV performs better than the GCV criterion. A possible explanation is that df(model) is not estimated as good as before. In the optimal model obtained by L_0 penalization and GCV, df(model) is estimated adequately in only eight of 100 cases (was 52 in setting *G3*).

In setting *B8*, there are four influential and four non-influential ordered predictors. In contrast to the previous settings, the distribution of the categorical covariates is not balanced; for the $n = 400$ observations, the covariates are drawn from a multinomial distribution with sampled probabilities between 0.12 and 0.44. In this setting, it can happen that the unpenalized estimate is quite extreme. As the adaptive weights depend on the quality of the ML estimate, they rely on estimates with a slight Ridge penalty for all coefficients (in the PIRLS algorithm $\boldsymbol{A}_\lambda$ is replaced by $\boldsymbol{A}_\lambda^{ridge} = \text{diag}(0.2)$).

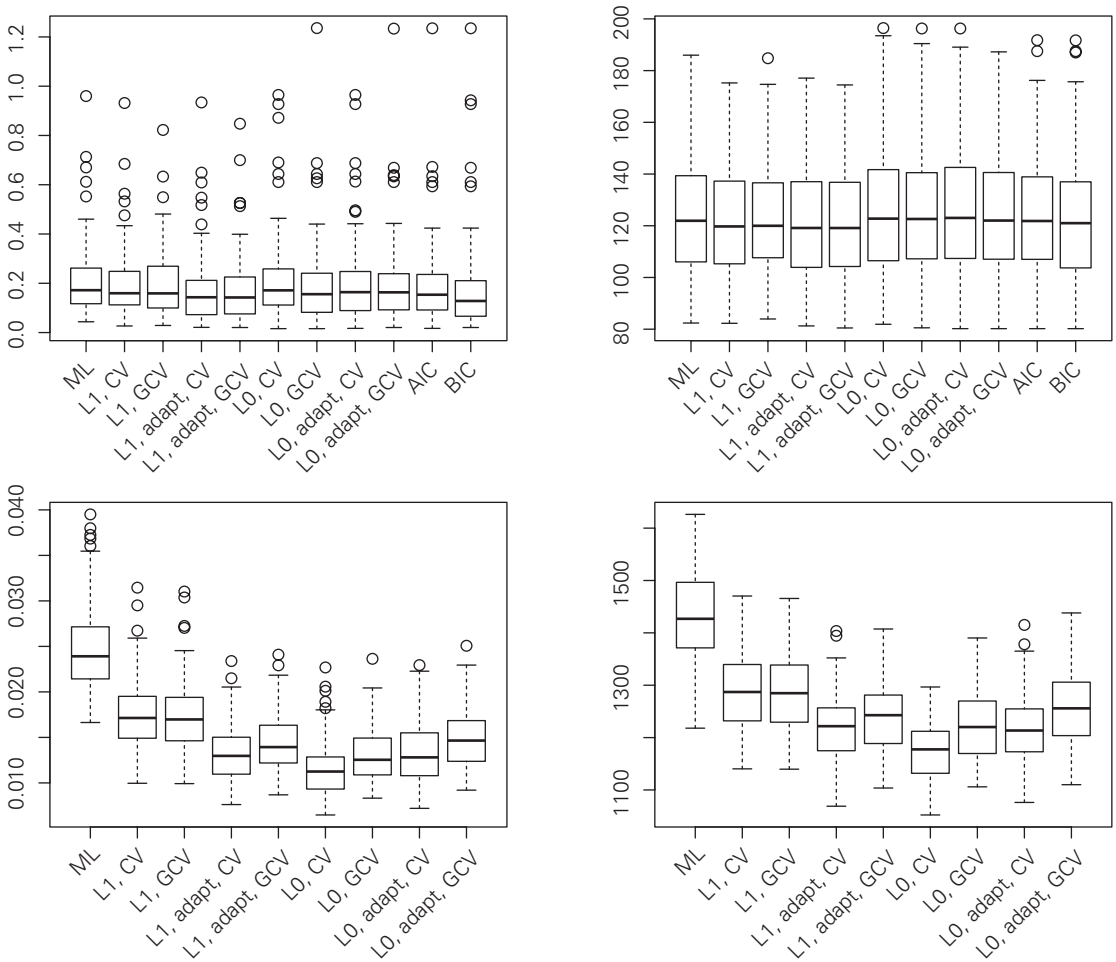

Figure 4 Results for settings *G3* (top) and *G50* (bottom): boxplots of squared errors of coefficients (left panel) and of the predictive deviances (right panel).
Source: Authors' own.

As seen in Figure 5, the estimation accuracy of the approximated L_0 penalty is slightly worse than that of the Lasso penalties. Concerning the selection and clustering performance in Table 1, it stands out that the FN rates are quite high when the L_0 penalty is combined with the CV criterion. For these approaches, the optimal values of λ are relatively large, too. Apparently, the sample size of the training data sets is too small for differentiated estimates. Hence, in settings like *B8*, the CV criterion is not recommended for L_0-type penalties. In contrast, with the GCV criterion, the clustering and selection performance of the L_0/GCV penalized models is better than that of the corresponding L_1 penalized approaches.

In general, L_0 penalized models seem to be sparser; the L_1 penalized models tend to have smaller FN rates. Even though there is less shrinkage in the coefficients paths

Table 1 Estimates of false positive (FP) and false negative (FN) rates for the selection (s) and clustering (c) performance for all considered settings.

Setting		ML	L1, CV	L1, GCV	L1, adapt, CV	L1, adapt, GCV	L0, CV	L0, GCV	L0, adapt, CV	L0, adapt, GCV	AIC	BIC
G3	FP_s	1.00	0.85	0.80	0.50	0.50	0.60	0.40	0.36	0.35	0.27	0.12
	FN_s	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01
	FP_c	1.00	0.90	0.88	0.53	0.58	0.61	0.42	0.30	0.34	0.26	0.14
	FN_c	0.00	0.02	0.03	0.07	0.06	0.09	0.11	0.15	0.12	0.15	0.22
G50	FP_s	1.00	0.74	0.72	0.37	0.47	0.13	0.51	0.23	0.44	-	-
	FN_s	0.00	0.00	0.00	0.01	0.01	0.02	0.01	0.02	0.01	-	-
	FP_c	1.00	0.77	0.76	0.42	0.52	0.21	0.60	0.28	0.50	-	-
	FN_c	0.00	0.00	0.00	0.01	0.01	0.02	0.00	0.02	0.01	-	-
P8	FP_s	1.00	0.74	0.69	0.40	0.41	0.17	0.33	0.15	0.31	-	-
	FN_s	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	-
	FP_c	1.00	0.78	0.75	0.42	0.44	0.20	0.39	0.15	0.33	-	-
	FN_c	0.00	0.01	0.01	0.02	0.02	0.05	0.02	0.06	0.03	-	-
B8	FP_s	1.00	0.74	0.83	0.53	0.69	0.37	0.65	0.35	0.56	-	
	FN_s	0.00	0.07	0.04	0.17	0.09	0.40	0.15	0.40	0.20	-	-
	FP_c	1.00	0.55	0.71	0.36	0.54	0.16	0.42	0.15	0.34	-	-
	FN_c	0.00	0.18	0.12	0.32	0.20	0.57	0.28	0.54	0.35	-	-

Source: Authors' own.

obtained with L_0 penalization, in general, the MSE of the L_0 approach is not smaller than the MSE of the L_1 approach as the estimates obtained with the L_0 approach are more sensitive to variations in the data. In standard situations, the adaptively weighted L_1 penalty and the L_0 approach perform comparably in terms of the estimation accuracy. In terms of variable selection, the L_0 approach has a higher incentive to cluster categories and reaches smaller FP while slightly enlarged FN are possible. Combining the L_0 approach with adaptive weights enhances the clustering and variable selection performance distinctly. With the L_0 penalization, we obtain stable results in settings where the computation of all possible subsets of coefficients which is needed for model selection based on information criteria, is not possible/efficient.

5 Unemployment Rates in Germany

We analyze the unemployment rates of the federal states of Germany in the years 2005 to 2010 (Weise *et al.*, 2011). The data is given in Table 2. For each of the 16 federal states, there are six annual unemployment rates observed: $(\text{state}_{it}, \text{rate}_{it})$, $i = 1, \ldots, 16$, $t = 2005, \ldots, 2010$. The aim is to find states with the same trends in the unemployment rates while accounting for the heterogeneity amongst the 16 units. Mixed models are the default approach to such data; see, for example, Molenberghs and Verbeke (2005). If one wants to model the unemployment rates by a mixed model, a potential predictor is $\eta_{it} = \beta_{int} + b_{int,i} + \beta_1 \cdot \text{time}$; whereby $b_{int,i}$, $i = 1, \ldots, 16$, are random effects for which a distribution is assumed, typically

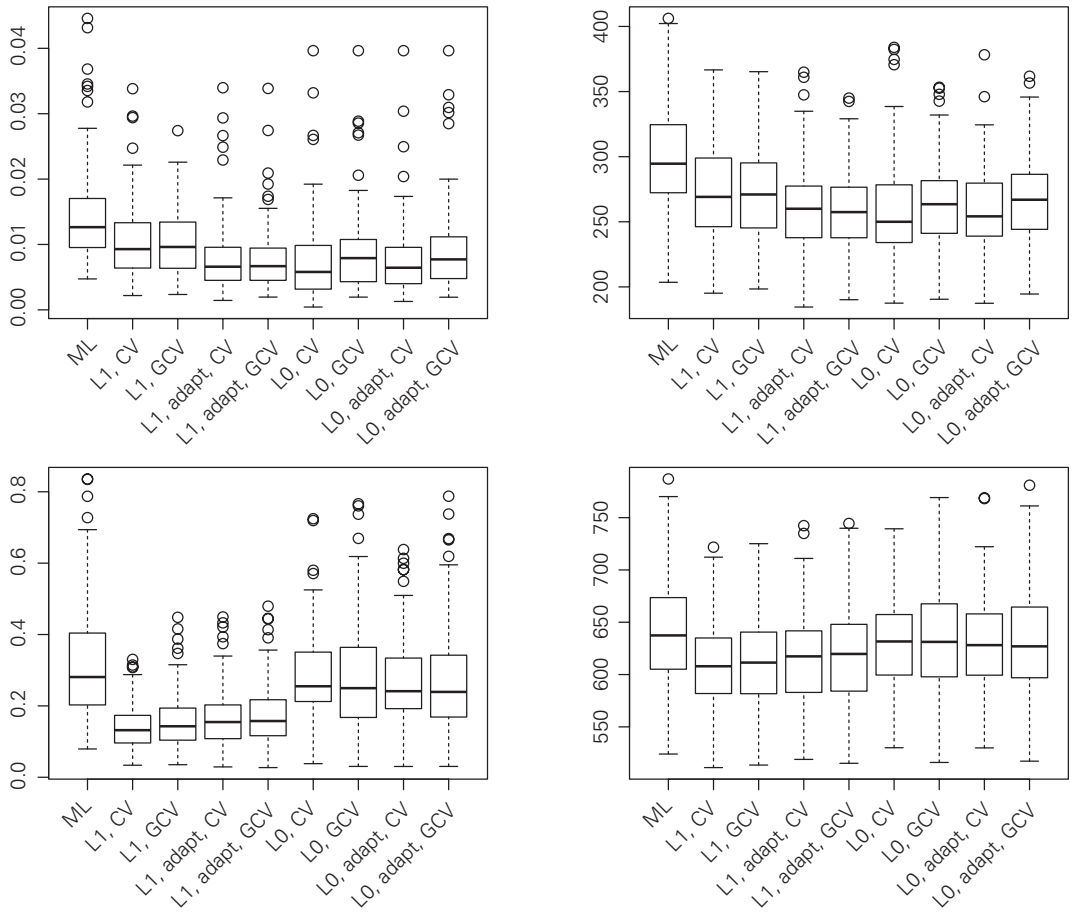

Figure 5 Results for settings *P8* (top) and *B8* (bottom): boxplots of the squared errors (left panel); boxplots of predictive deviances (right panel).
Source: Authors' own.

a normal distribution with variance σ_b^2: $b_{int,i} \sim N(0, \sigma_b^2)$. Clustering federal states with similar effects relates to identical random effects and hence, requires sophisticated distributional assumptions; for example, a mixture distribution of Gaussian components. This in turn requires elaborate estimation theory, for details, see for example Heinzl (2013). Moreover, the data is positively skewed. There are high unemployment rates, but they occur rarely. The response seems to be rather Gamma than Gaussian distributed.

Hence, we assume a fixed effects model with Gamma distributed response and a logarithmic link function. The predictor contains one intercept per federal state and a global temporal trend:

$$\eta_{it} = \beta_{int,i} + \beta_1 \cdot \text{time}, \quad \text{with} \quad i = 1, \ldots, 16. \tag{5.1}$$

Table 2 Unemployment rates for the federal states of Germany in 2005 to 2010.

Abbreviation	Federal State	2005	2006	2007	2008	2009	2010
BB	Brandenburg	18.20	17.00	14.90	13.00	12.30	11.10
BE	Berlin	19.00	17.50	15.50	13.90	14.10	13.60
BW	Baden-Würtenberg	7.00	6.30	4.90	4.10	5.10	4.90
BY	Bayern	7.80	6.80	5.30	4.20	4.80	4.50
HB	Hansestadt Bremen	16.80	14.90	12.70	11.40	11.80	12.00
HE	Hessen	9.70	9.20	7.60	6.60	6.80	6.40
HH	Hansestadt Hamburg	11.30	11.00	9.20	8.10	8.60	8.20
MV	Mecklenburg-Vorpommern	20.30	19.00	16.50	14.10	13.60	12.70
NI	Niedersachsen	11.60	10.50	8.90	7.70	7.80	7.50
NRW	Nordrhein-Westfalen	12.00	11.40	9.50	8.50	8.90	8.70
RP	Rheinland-Pfalz	8.80	8.00	6.50	5.60	6.10	5.70
SA	Sachsen	18.30	17.00	14.70	12.80	12.90	11.90
SH	Schleswig-Holstein	11.60	10.00	8.40	7.60	7.80	7.50
SL	Saarland	10.70	9.90	8.40	7.30	7.70	7.50
ST	Sachsen-Anhalt	20.20	18.30	16.00	14.00	13.60	12.50
TH	Thüringen	17.10	15.60	13.20	11.30	11.40	9.80

Source: Data from Weise *et al.* (2011).

To cluster the federal states, in a first model, the subject-specific intercepts are penalized by penalty (3.2), whereby differences to the reference category are omitted as there is none; that is, all pairwise differences of intercepts are penalized by the L_0 norm:

$$\lambda \cdot P(\boldsymbol{\beta}) = \lambda \cdot \sum_{r>s>0} \left\| \beta_{int,r} - \beta_{int,s} \right\|_0 . \tag{5.2}$$

In a second model with the same predictor, the spatial structure of the federal states is considered. Weights $w_{r,s}$ are defined as indicators for states with a common border ($w_{r,s} = 1$ if neighbored, $w_{r,s} = 0$ else). We will refer to this model as the 'spatial' model. For both models, the tuning of the algorithm is similar: $\gamma = 36$, $\gamma^{spatial} = 26$, $\nu = 0.5$. The tuning parameter λ is chosen by the generalized cross-validation criterion of O'Sullivan *et al.* (1986). It yields $\lambda_{CV} = 0.14$ and $\lambda_{CV}^{spatial} = 0.05$. Figure 6 shows the coefficient paths of the subject specific intercepts for both models. The left panel relates to the first model with penalty (5.2), the right one to the spatial model. In both models, there are basically two clusters of federal states. The upper cluster contains the states of the former German Democratic Republic (GDR) including Berlin plus the city state Bremen. Interestingly, with the spatial weights, the city state Bremen switches the cluster for a relatively large value of λ. Figure 7 illustrates the resulting clusters for the optimal choice of λ in a map of Germany. The darker the coloring, the bigger is the subject-specific intercept and the higher is the unemployment rate over the time.

In the left panel, the ML estimates are illustrated. Even though all estimates differ, the pattern of the former GDR in the north-east is clearly seen. In the middle, the subject specific intercepts are clustered by the pairwise penalty (5.2). Here, the states

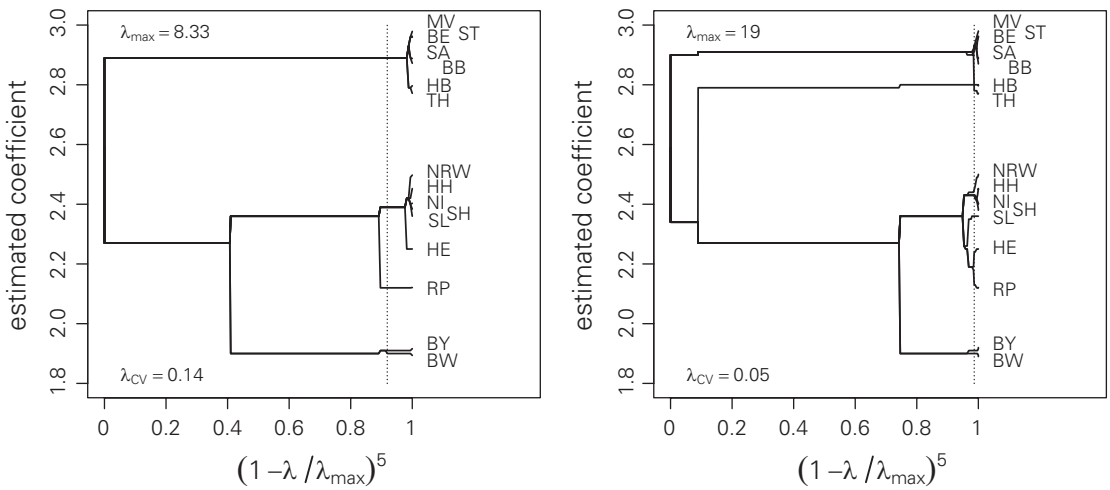

Figure 6 Unemployment rates in Germany—coefficients paths for L_0 ($\gamma = 36$, $\nu = 0.5$) penalization considering all pairwise differences (left panel) and differences of coefficients related to neighboured federal states only (right panel).
Source: Authors' own.

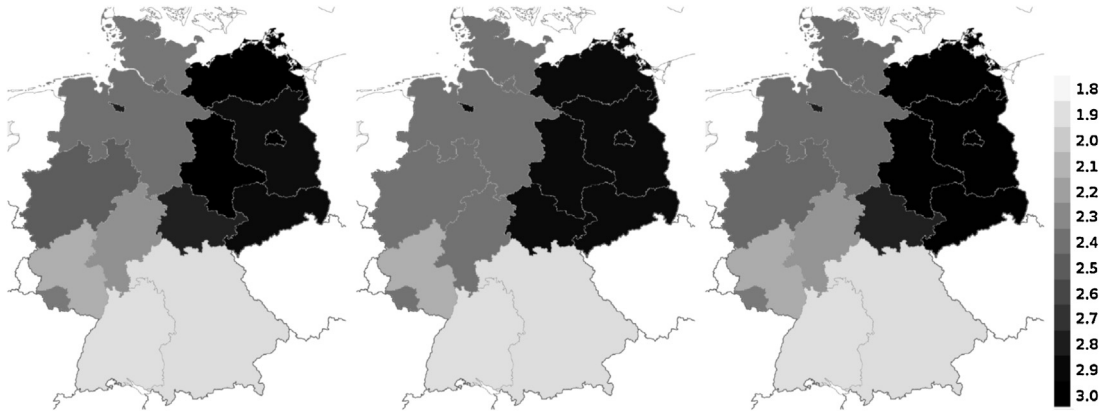

Figure 7 Unemployment rates in Germany—visualization of the specific effect of the federal states on the unemployment rates. In the very left panel, the ML estimates are shown; in the middle, all pairwise differences of coefficients are penalized by an L_0 penalty; in the very right panel, only the differences of coefficients of neighboured states are penalized.
Source: The maps are based on a figure of Wikipedia User NordNordWest (2008); they are manipulated with the GNU Image Manipulation Program (GIMP Team, 2012) and with the `R` package `EBImage` (Pau *et al.*, 2012).

of the former GDR plus Bremen form one cluster with the biggest impact on the response ($\hat{\beta}_{int,i} = 2.89$). The effects of the southern states Baden-Würtenberg and Bayern are the lowest ($\hat{\beta}_{int,BW} = 1.90, \hat{\beta}_{int,BW} = 1.91$). The remaining states except for Rheinland-Pfalz ($\hat{\beta}_{int,RP} = 2.12$) are clustered; the according intercept is $\hat{\beta}_{int,i} = 2.39$. The right panel of Figure 7 illustrates the results of the spatial model. The results

resemble the middle panel; however, the picture is more differentiated: the states of the former GDR form one cluster ($\hat{\beta}_{int,i}^{spatial} = 2.93$), but there are slightly different estimates for the states Thüringen and Bremen ($\hat{\beta}_{int,TH}^{spatial} = 2.78$, $\hat{\beta}_{int,HB}^{spatial} = 2.80$). The effects of Baden-Würtenberg and Bayern are the same as before; but in the west, only Hamburg, Niedersachen and Schleswig-Holstein are clustered ($\hat{\beta}_{int,i}^{spatial} = 2.43$). The other states have individual intercepts ($\hat{\beta}_{int,RP}^{spatial} = 2.13$, $\hat{\beta}_{int,HE}^{spatial} = 2.24$, $\hat{\beta}_{int,SL}^{spatial} = 2.36$, $\hat{\beta}_{int,NRW}^{spatial} = 2.45$). The estimates for the global temporal trend are approximately the same in all models: $\hat{\beta}_t^{ML} = -0.0875$, $\hat{\beta}_t = -0.09$, $\hat{\beta}_t^{spatial} = -0.09$. In the considered time period, the unemployment rates decreased in all states. Interestingly, Heinzl (2013) obtains similar clusters for the same data by fitting a linear mixed model with Dirichlet process mixtures using the EM algorithm. However, the computational burden for such models is higher.

6 Summary

In this article, we propose L_0 penalization for categorical effects in GLMs. The penalty works on differences of coefficients and accounts for the different amount of information contained in nominal and ordered factors. Unlike Rippe *et al.* (2012), we provide a classical regression framework for L_0 penalization. Computational issues are met by a local quadratic approximation which can be traced back to (Fan and Li, 2001). The approximation allows one to derive a PIRLS algorithm; that is, all features of Fisher scoring algorithms are sustained. It is possible to obtain coefficient paths.

Applying L_0 penalization to plain coefficients with fixed tuning has a close relation to best subset selection. As the L_0 approach allows for more flexible terms such as difference in the penalty and as it works for more complex models, L_0 penalization is an attractive alternative to model selection based on information criteria. In an illustrative example and several numerical experiments, the proposed method is competitive; however, it requires carefully tailored tuning. The range of applications for L_0 penalization is huge: applied adequately to continuous covariates, it is an alternative computational approach to model selection based on information criteria; the application to subject specific intercepts is convincing; the computational framework allows one to combine L_0 penalization easily with other types of (smooth) covariates.

Acknowledgements

This work was partially supported by DFG project 'Regularisierung für diskrete Datenstrukturen'. We thank the referees, the associate editor and the editor sincerely for their helpful comments.

SUPPLEMENT A

Proofs for Section 2.

SUPPLEMENT B

Data and R Code to replicate the results of Section 5.

References

Antoniadis A and Fan J (2001) Regularization of wavelet approximations. *J. Amer. Statist. Assoc.*, **96**, 939–67.

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

Bozdogan H (1987) Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika*, **52**, 345–70.

Donoho DL and Elad M (2003) Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization. *Proceedings of the National Academy of Sciences*, **100**, 2197–2202.

Fahrmeir L and Tutz G (2001) *Multivariate statistical modelling based on generalized linear models*. New York: Springer Verlag.

Fan J and Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, **96**, 1348–60.

Ge D, Jiang X and Ye Y (2011) A note on the complexity of l_p minimization. *Math. Program.*, **192**, 285–99.

Gertheiss J and Tutz G (2010) Sparse modelling of categorial explanatory variables. *Ann. Appl. Stat.*, **4**, 2150–80.

GIMP Team (2012) *GNU image manipulation program*. http://www.gimp.org.

Heinzl F (2013) *Clustering in linear and additive mixed models*. Dissertation, Department of Statistics, Ludwig-Maximilians-Universität München: Cuvillier Verlag Göttingen.

Jain A and Dubes R (1988) *Algorithms for Clustering Data*. New Jersey: Prentice Hall.

Johnson J (2013) A dynamic programming algorithm for the fused lasso and L_0-segmentation. *J. Comput. Graph. Statist.*, **22**, 246–60.

Lu Z and Zhang Y (2013) Sparse approximation via penalty decomposition methods. *SIAM J. Optim.*, **23**, 2448–78.

Lu Z and Zhang Y (2010) Penalty decomposition methods for l_0-norm minimization. *arXiv:1008.5372*.

Mancera L and Portilla J (2006) L0-norm-based sparse representation through alternate projections. In *International Conference on Image Processing*, pp. 2089–92. IEEE.

Molenberghs G and Verbeke G (2005) *Models for discrete longitdinal data*. New York: Springer–Verlag.

Oelker MR (2013) *gvcm.cat: Regularized categorial effects/categorial effect modifiers in GLMs*. R package version 1.6.

Oelker M-R, Gertheiss J and Tutz G (2014) Regularization and model selection with categorical predictors and effect modifiers in generalized linear models. *Statistical Modelling*, **14**, 157–77.

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 Reports 139, http://epub.ub.uni-muenchen.de/17664/*.

O'Sullivan F, Yandell BS and Raynor WJ (1986) Automatic smoothing of regression

functions in generalized linear models. *J. Amer. Statist. Assoc.*, **81**, 96–03.

Pollak I, Willsky AS and Huang Y (2005) Nonlinear evolution equations as fast and exact solvers of estimation problems. *IEEE Transactions of Signal Processing*, **53**, 484–98.

Pau G, Oles A, Smith M, Sklyar O and Huber W (2012) *EBImage: Image processing toolbox for R*. R package version 4.4.0.

R Core Team (2013) *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. R version 3.0.2 (2013-09-25).

Rippe RCA, Meulman JJ and Eilers PHC (2012) Visualization of genomic changes by segmented smoothing using an l_0 penalty. *PloS One*, **6**, 1–14.

Schwarz G (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–64.

Tibshirani R (1996) Regression shrinkage and selection via the LASSO. *R. Stat. Soc. Ser. B Stat. Methodol.*, **58**, 267–88.

Tibshirani R, Saunders M, Rosset S, Zhu J and Knight K (2005) Sparsity and smoothness via the fused LASSO. *R. Stat. Soc. Ser. B Stat. Methodol.*, **67**, 91–08.

Tutz G and Gertheiss J (2014) Rating scales as predictors–the old question of scale level and some answers. *Psychometrika*, **79**, 357–76.

Weise FJ, Alt H and Becker (Eds) (2011) *Arbeitsmarkt in Zahlen*, Nürnberg. Statistik der Bundesagentur für Arbeit.

Wikipedia User NordNordWest (2008) *Federal states of Germany*. http://commons.wikimedia.org/wiki/File:Germany_location_map.svg. Licenses: GNU Free Documentation License, Version 1.2 http://commons.wikimedia.org/wiki/Commons:GNU_Free_Documentation_License,_version_1.2, Creative Commons Attribution-Share Alike 3.0 Unported http://creativecommons.org/licenses/by-sa/3.0/deed.en.

Wipf D and Rao B (2005) L_0-norm minimization for basis selection. *Advances in Neural Information Processing Systems*, **17**, 1513–20.

Wood S (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *R. Stat. Soc. Ser. B Stat. Methodol.*, **73**, 3–36.

Xiang Y, Gubian S, Suomela B and Hoeng J (2013) Generalized simulated annealing for global optimization: the GenSA package. *R Journ*, **5**, 13–29. R package version 1.1.4.

Zou H (2006) The adaptive LASSO and its oracle properties. *J. Amer. Statist. Assoc.*, **101**, 1418–29.