LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR STATISTIK

Margret-Ruth Oelker, Gerhard Tutz

A General Family of Penalties for Combining Differing Types of Penalties in Generalized Structured Models

Technical Report Number 139, 2013
Department of Statistics
University of Munich

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

A General Family of Penalties for Combining Differing Types of Penalties in Generalized Structured Models

Margret-Ruth Oelker*† & Gerhard Tutz†

December 19, 2013

Abstract

Penalized estimation has become an established tool for regularization and model selection in regression models. A variety of penalties with specific features are available and effective algorithms for specific penalties have been proposed. But not much is available to fit models that call for a combination of different penalties. When modeling rent data, which will be considered as an example, various types of predictors call for a combination of a Ridge, a grouped Lasso and a Lasso-type penalty within one model. Algorithms that can deal with such problems, are in demand. We propose to approximate penalties that are (semi-)norms of scalar linear transformations of the coefficient vector in generalized structured models – such that penalties of various kinds can be combined in one model. The approach is very general such that the Lasso, the fused Lasso, the Ridge, the smoothly clipped absolute deviation penalty (SCAD), the elastic net and many more penalties are embedded. The computation is based on conventional penalized iteratively re-weighted least squares (PIRLS) algorithms and hence, easy to implement. Moreover, new penalties can be incorporated quickly. The approach is extended to penalties with vector based arguments; that is, to penalties with norms of linear transformations of the coefficient vector. A software implementation is available. Some illustrative examples and the model for the Munich rent data show promising results.

Keywords: Model selection; Penalties; Generalized linear model (GLM); Structured regression; Ridge; Lasso; Grouped Lasso; SCAD; Elastic net; Fused Lasso.

1 Introduction

In recent years, model selection and regularization in regression models has been an area of intensive research. Often, penalized approaches are the method of choice. Examples are Ridge regression, the Lasso, the smoothly clipped absolute deviation penalty (SCAD), the fused Lasso, the elastic net and the (adaptive) grouped Lasso, to mention only a few approaches (Hoerl and

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

†Department of Statistics, Ludwig-Maximilians-Universität Munich; address: Akademiestraße 1, 80799 Munich, Germany; phone: +49 89 2180 3351

Kennard, 1970; Tibshirani, 1996; Fan and Li, 2001; Tibshirani et al., 2005; Zou and Hastie, 2005; Yuan and Lin, 2006; Wang and Leng, 2008). The number of applications is huge: In nonparametric regression, penalties smooth wiggly functions; for example, Eilers and Marx (1996) work with Ridge penalties on higher order differences of B-spline coefficients. Meier et al. (2009) select splines with a grouped Lasso-type penalty. For wavelets and signals, L_0 penalties, or more general L_q penalties, $0 \leq q \leq 1$, are employed (Antoniadis and Fan, 2001; Rippe et al., 2012). Concerning categorical data, Bondell and Reich (2009) or Gertheiss and Tutz (2010) work with fused Lasso type penalties. Fahrmeir et al. (2010) offer a flexible framework for Bayesian regularization and variable selection; amongst others with spike and slab priors.

Various efficient algorithms to solve the resulting optimization problems are available, be it in linear models, generalized linear models (GLMs), hazard rate models or others. Least angle regression (lars; Efron et al., 2004; Hastie and Efron, 2013) offers a conceptual framework to compute the entire regularization path of the Lasso by exploiting its piecewise linear coefficient profiles. Osborne and Turlach (2011) propose a homotopy algorithm for the quantile regression Lasso and related piecewise linear problems. Meier et al. (2008) propose a coordinate-descent algorithm for the group Lasso in logistic regression problems. Goeman (2010) solves Lasso, fused Lasso and Ridge-type problems in high-dimensional models by a combination of gradient ascent optimization with the Newton-Raphson algorithm. Friedman et al. (2010) use cyclical coordinate descent algorithms, computed along a regularization path, for the elastic net and related convex penalties. Ulbricht (2010) proposes a penalized iteratively re-weighted least squares (PIRLS) algorithm for Lasso-type penalties in GLMs. The `R`-package `mgcv` (R Core Team, 2013; Wood, 2011) offers a toolbox for generalized additive models and generalized Ridge regression.

In the mentioned approaches, penalties have a specific purpose; for example, the selection of variables in a linear predictor or of smooth non-linear effects. In applications, however, frequently a combination of penalties is needed that serves different purposes. For illustration, we consider the rent data of 1488 households in the city of Munich. To model the rent, continuous covariates like the flat's size and age, as well as some explanatory factors were collected. The effect of the age of a flat is known to be non-linear (see, for example, Fahrmeir and Tutz, 2001) and can be modeled by splines with a Ridge-type penalty on the curvature. When investigating whether the effect of the residential area is linear or not, an additional grouped Lasso penalty is helpful. As some levels of categorical predictors are only sparsely occupied, ordered factors like the number of rooms of a flat require regularization, too. This can be done by employing a fused Lasso type penalty on the dummy coefficients of these covariates. Hence, the overall penalty is a sum of Ridge-, grouped Lasso- and Lasso-type penalties. We will use a generalized structured regression model with Gamma distributed response.

Although the algorithm of Friedman et al. (2010) covers Ridge- and Lasso-type penalties within one model via the elastic net, it does not allow for other penalties. The `R`-package `mgcv` allows penalized smooth functions and penalized parametric terms, but penalties of parametric terms have to be quadratic. Even though the algorithm of Ulbricht (2010) works for a family of Lasso-type penalties, we found no algorithm obviously matching the requirements of our data. As in Marx and Eilers (1998), many algorithms are based on Fisher scoring methods, which are the default approach to estimate GLMs. For quadratic penalties, a penalty matrix is added to the Fisher information matrix. See, for example, the PIRLS algorithm in `R`-package `mgcv`. For non-quadratic penalties, approximations are available and again PIRLS algorithms are ap-

plied: Fan and Li (2001) approximate the non-convex SCAD penalty quadratically. Ulbricht (2010) adopts this idea for Lasso-type penalties. Rippe et al. (2012) approximate the L_0 norm quadratically by a re-weighted Ridge penalty. Hence, to combine different penalties that employ different norms, quadratic approximations in PIRLS algorithms seem to be a natural choice.
In this paper, we show how penalties that are (semi-)norms of scalar linear transformations of the coefficient vector, can be approximated quadratically within a general model structure as in GLMs; the penalty is defined such that the Lasso, the fused Lasso, the Ridge, the SCAD, the elastic net and other regularization terms for categorical predictors are embedded. The approximation allows to combine all these penalties in one model. The estimation is based on conventional PIRLS algorithms and hence, easy to implement. The approximation is based on and generalizes the approaches of Fan and Li (2001) and Ulbricht (2010); it is not restricted to penalties that are based on (functions of) absolute values but allows for penalties with general norms. The approach is extended to penalties like the grouped Lasso; that is, to penalties with norms of vectorial linear transformations of the coefficient vector.
The rest of the paper is organized as follows: Section 2 introduces the method and its derivation; some technical remarks and the extension to vectorial linear transformations of the coefficient vector are given. Section 3 gives some illustration; established algorithms and the new approximation are compared. In Section 4, the Munich rent data is analyzed.

2 Local Quadratic Approximations in PIRLS Algorithms

We consider a general model structure as in GLMs by assuming that the mean response $\mu_i = \mathbb{E}(y_i|x_i)$ is given by

$$\mu_i = h(\eta_i).$$

Given the vector of covariates $\boldsymbol{x}_i \in \mathbb{R}^q$, y_i follows a simple exponential family (for details, see, McCullagh and Nelder, 1983; Fahrmeir and Tutz, 2001). The conditional mean of response y_i is linked to a linear predictor $\eta_i = \boldsymbol{x_i}^T\boldsymbol{\beta}$, where $\boldsymbol{\beta} \in \mathbb{R}^q$ is a coefficient vector and h is a twice continuously differentiable inverse link function, often referred to as response function. Vectors $\boldsymbol{x}_i$ build the design matrix $\boldsymbol{X} = (\boldsymbol{x}_1^T, \dots, \boldsymbol{x}_n^T)^T \in \mathbb{R}^{n \times q}$, which represents $1, \dots, p$ covariates. By a general model structure, we mean that the covariates in design matrix $\boldsymbol{X}$ can have any structure – provided that they can be parametrized as $\boldsymbol{x}_i^T\boldsymbol{\beta}$. In particular, we allow for non-parametric terms that represent unknown functions. When for example, a continuous covariate is modeled non-parametrically as $f(\boldsymbol{x}_j)$, we assume that $f(\boldsymbol{x}_j)$ is represented in $\boldsymbol{X}$ by the evaluations of some basis functions. Categorical covariates are assumed to be properly coded. We always include an intercept in the design matrix; and assume that the structure of the coefficient vector is $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_0^T, \boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_p^T)$, where entries $\boldsymbol{\beta}_j$ are vectors $\in \mathbb{R}^{k_j}$ that correspond to structures in the predictor space. A vector $\boldsymbol{\beta}_j$ can, for example, contain the coefficients of the basis functions of a smoothly modeled predictor, or the coefficients linked to the dummies of a categorical predictor. In the penalized maximum likelihood framework considered here, the objective function is

$$\mathcal{M}_{pen}(\boldsymbol{\beta}) = -l_n(\boldsymbol{\beta}) + P_\lambda(\boldsymbol{\beta}), \tag{1}$$

where $l_n(\boldsymbol{\beta})$ denotes the log-likelihood of the exponential family based on n observations and where $P_\lambda(\boldsymbol{\beta})$ denotes the penalty. The general penalty that is considered, has the form

$$P_\lambda(\boldsymbol{\beta}) = \sum_{l=1}^{L} \lambda_l p_l(\|\boldsymbol{a}_l^T \boldsymbol{\beta}\|_{\mathrm{N}_l}), \tag{2}$$

where functions p_l are penalty functions, λ_l are tuning parameters, and $\|\cdot\|_{\mathrm{N}_l}$ denotes any (semi) norm, for example $\|\xi\|_{\mathrm{N}_l} = |\xi|^r$, $r \geq 0$. $\|\cdot\|_{\mathrm{N}_l}$ is not restricted to (semi-)norms; it can be any term that is meaningful as a penalty – for example an indicator for non-zero arguments, which is often called L_0 "norm" (Donoho and Elad, 2003). Vectors $\boldsymbol{a}_l \in \mathbb{R}^q$ build transformations of the coefficient vector, for example, differences of adjacent coefficients. As in Ulbricht (2010), in principal, there can be arbitrary many restrictions L. And as proposed by Ulbricht (2010), for the penalty functions, we assume

1. $p_l : \mathbb{R}^+ \to \mathbb{R}^+$, $p_l(0) = 0$
2. $p_l(\xi)$ is continuous and strictly monotone in ξ
3. $p_l(\xi)$ is continuously differentiable for all $\xi \neq 0$, such that $p_l' = \mathrm{d}p_l(\xi)/\mathrm{d}\xi > 0$.

Together, p_l, $\|\cdot\|_{\mathrm{N}_l}$ and vectors $\boldsymbol{a}_l$ define the type of the penalty. Note, that properties like the curvature of $p_l(\|\xi\|_{\mathrm{N}_l})$ do depend on the properties of $p_l(\xi)$ and $\|\xi\|_{\mathrm{N}_l}$. When for example, $p_l(\xi)$ and $\|\xi\|_{\mathrm{N}_l}$ are convex for all l, and $p_l(\xi)$ is monotonically increasing as assumed, then the penalty is convex. The flexibility of the penalty lies in the possible choices of the three components. In the following some examples are given:

Elastic net: To penalize a scalar effect β_j by the elastic net $\lambda_l \cdot |\beta_j| + \lambda_k \cdot \beta_j^2$ (Zou and Hastie, 2005), two penalty functions are needed; one denoted by $p_l(\xi) = \xi$, $\|\xi\|_{\mathrm{N}_l} = |\xi|$ and an indicator vector $\boldsymbol{a}_l$ such that $\boldsymbol{a}_l^T \boldsymbol{\beta} = \beta_j$; the other is $p_k(\xi) = \xi$ with $\|\xi\|_{\mathrm{N}_k} = \xi^2$ and with the same indicator vector $\boldsymbol{a}_l$ as before.

Adaptive Lasso: To penalize the effect of the j-th continuous covariate with the adaptive Lasso (Zou, 2006), $\boldsymbol{a}_l$ is an indicator vector for the position of β_j, $\|\xi\|_{\mathrm{N}_l}$ is the absolute value $|\xi|$ and $p_l(\xi) = |\boldsymbol{a}_l^T \hat{\boldsymbol{\beta}}^{ML}|^{-1} \cdot \xi$, where $\hat{\boldsymbol{\beta}}^{ML}$ denotes the maximum likelihood (ML) estimate of $\boldsymbol{\beta}$.

Penalized B-splines: When the continuous covariate $\boldsymbol{x}_j$ is modeled non-parametrically, $\boldsymbol{\beta}_j$ is a sub-vector that represents coefficients on k_j basis functions, for example, of cubic B-splines. To penalize the roughness of $f(\boldsymbol{x}_j)$ as proposed by Eilers and Marx (1996), there are $k_j - 2$ penalty terms; vectors $\boldsymbol{a}_l$ build all needed second order differences $(0, \ldots, 0, 1, -2, 1, 0, \ldots, 0)^T$; for all penalty terms one employs $\|\xi\|_{\mathrm{N}_l} = \xi^2$, $p_l(\xi) = \xi$ and the identical tuning parameter λ_l.

Typically, the penalty is structured as $P_\lambda(\boldsymbol{\beta}) = \sum_{j=0}^{p} \sum_{l=1}^{L_j} \lambda_{jl} p_{jl}(\|\boldsymbol{a}_{jl}^T \boldsymbol{\beta}_j\|_{\mathrm{N}_l})$; that is, the effects of each covariate are penalized separately. We will, however, use the general form (2), which uses one index to denote the specific terms.

In common GLMs, the unpenalized optimization problem is $\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta}} -l_n(\boldsymbol{\beta})$. The equation is solved iteratively by solving the linearized problem

$$\boldsymbol{s}^{lin}(\boldsymbol{\beta}) = \boldsymbol{s}(\boldsymbol{\beta}_{(k)}) + \boldsymbol{H}(\boldsymbol{\beta}_{(k)})(\boldsymbol{\beta} - \boldsymbol{\beta}_{(k)}) = \boldsymbol{0}$$

for a given $\boldsymbol{\beta}_{(k)}$ in each step, where $\boldsymbol{s}(\boldsymbol{\beta}) = \partial l_n(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$ is the score function and $\boldsymbol{H}(\boldsymbol{\beta}) = \partial^2 l_n(\boldsymbol{\beta})/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T$ is the Hessian matrix. Rearranging gives

$$\hat{\boldsymbol{\beta}}_{(k+1)} = \hat{\boldsymbol{\beta}}_{(k)} - \boldsymbol{H}(\hat{\boldsymbol{\beta}}_{(k)})^{-1} \boldsymbol{s}(\hat{\beta}_{(k)}),$$

Norm	$\mathcal{N}_l(\xi)$	$\mathcal{D}_l(\xi)$
$\Vert\xi\Vert_1 = \vert\xi\vert$	$\sqrt{\xi^2+c}$	$(\xi^2+c)^{-1/2}\cdot\xi$
$\Vert\xi\Vert_2^2 = \xi^2$	ξ^2	2ξ
$\Vert\xi\Vert_0 = I_{\xi\neq 0}$	$\frac{2}{1+\exp(-\gamma\vert\xi\vert)}-1$	$\frac{2\gamma}{1+\exp(-\gamma\vert\xi\vert)}(1-\frac{1}{1+\exp(-\gamma\vert\xi\vert)})\frac{\xi}{\sqrt{\xi^2+c}}$
$\Vert\xi\Vert_r = \vert\xi\vert^r$	$(\xi^2+c)^{r/2}$	$r\xi(\xi^2+c)^{r/2-1}$

Table 1: Examples for approximations of norms. Column $\mathcal{N}_l(\xi)$ gives direct approximations of the L_1 norm, of the quadratic term $\|\xi\|_2^2 = \xi^2$, of the L_0 norm and of the term $\|\xi\|_r$ which is needed for Bridge penalties. Column $\mathcal{D}_l(\xi)$ gives the (approximated) derivatives of $\mathcal{N}_l(\xi)$. c is a small positive number; $c \approx 10^{-5}$ worked quite well in numerical experiments. The approximation of the L_0 norm is motivated by the logistic function, whereby γ is a large integer.

which can be transformed to a Fisher scoring algorithm or an iteratively re-weighted least squares algorithm. In order to use a PIRLS algorithm for the penalized optimization problem (1), penalized versions of the score function $\boldsymbol{s}(\boldsymbol{\beta}_{(k)})$ and the Hessian matrix $\boldsymbol{H}(\boldsymbol{\beta}_{(k)})$ or the Fisher matrix are needed. In particular, derivatives of $\mathcal{M}_{pen}(\boldsymbol{\beta})$ or close approximations are needed. To this end, non-differentiable norms $\|\cdot\|_{\mathrm{N}_l}$ are approximated. We assume that an approximation $\mathcal{N}_l(\xi, \mathcal{T})$ to each employed norm $\|\cdot\|_{\mathrm{N}_l}$ exists, such that

$$\|\xi\|_{\mathrm{N}_l} = \lim_{\mathcal{T}\to\mathcal{B}} \ \mathcal{N}_l(\xi, \mathcal{T}),$$

where $\mathcal{T}$ denotes a set of possible tuning parameters and $\mathcal{B}$ denotes the set of boundary values with $\|\xi\|_{\mathrm{N}_l} = \mathcal{N}_l(\xi, \mathcal{B})$. $\mathcal{N}_l(\xi, \mathcal{T})$ is supposed to be at least twice continuously differentiable. Moreover, we define $\mathcal{D}_l(\xi, \mathcal{T}) = \partial\mathcal{N}_l(\xi, \mathcal{T})/\partial\xi$ and assume that

$$\frac{\partial \|\xi\|_{\mathrm{N}_l}}{\partial\xi} = \lim_{\mathcal{T}\to\mathcal{B}} \ \mathcal{D}_l(\xi, \mathcal{T}),$$

for all l, $l = 1, \ldots, L$, and for all ξ for which $\partial \|\xi\|_{\mathrm{N}_l} / \partial\xi$ is defined. To keep the notation simple, we will write $\mathcal{N}_l(\xi)$ instead of $\mathcal{N}_l(\xi, \mathcal{T})$, and $\mathcal{D}_l(\xi)$ respectively. Apart from this approximation, the schedule is the same as for the unpenalized case: the penalized score function $\boldsymbol{s}_{pen}(\boldsymbol{\beta})$ is linearized by a Taylor expansion; $\boldsymbol{s}_{pen}^{lin}(\boldsymbol{\beta}) = \boldsymbol{0}$ is solved iteratively.

2.1 Examples for Approximations

Table 1 gives an idea of the approximations of different norms. As in Koch (1996) and Ulbricht (2010), the L_1 norm is approximated by $\mathcal{N}_l(\xi) = \sqrt{\xi^2+c}$ where c is a small positive number (in our experience $c \approx 10^{-5}$ works well) and controls how close the approximation and the L_1 norm are. For $c = 0$, we have $|\xi| = \sqrt{\xi^2}$. The first derivative of $\mathcal{N}_l(\xi)$ is $\xi(\xi^2+c)^{-1/2}$ which is a continuous approximation for $\mathrm{sign}(\xi)$, $\xi \neq 0$, the first derivative of the L_1 norm. For illustration, see the upper panel of Figure 1. There is no need to approximate $\|\xi\|_2^2 = \xi^2$ as it is quadratic. The approximation of the L_0 norm is motivated by the logistic function. We choose $\mathcal{N}_l(\xi) = 2(1+\exp(-\gamma|\xi|))^{-1} - 1$, where γ is a large integer. Accordingly, the derivative is $\mathcal{D}(\xi) = 2\gamma/(1+\exp(-\gamma|\xi_{(k)}|))(1-1/(1+\exp(-\gamma|\xi_{(k)}|)))(\xi_{(k)}^2+c)^{-1/2}$, whereby the absolute value is approximated like defined above. Figure 1 illustrates these approximations (left panel) and their derivatives (right panel). On top, the approximation of the L_1 norm is shown. On bottom the L_0 norm is approximated by $\mathcal{N}_l(\xi) = 2(1+\exp(-\gamma|\xi|))^{-1} - 1$. The dashed lines mark the exact norms and the exact derivatives based on sub-gradients at $\xi = 0$. For all plots, we have used $c = 0.01$ and $\gamma = 5$ for illustrative reasons.

Combining these approximations with different functions $p_l(\cdot)$ allows to approximate various known penalties. Table 2 illustrates the variety of penalties that can be approximated. The penalties are dissected in the underlying norm $\|\xi\|_{\mathrm{N}_l}$, functions p_l and expressions $\boldsymbol{a}_l^T\boldsymbol{\beta}$. Of course, other combinations of norms $\|\xi\|_{\mathrm{N}_l}$, functions p_l and $\boldsymbol{a}_l^T\boldsymbol{\beta}$ are possible. One could for example think of an adaptively weighted Ridge penalty or L_1-type penalties for coefficients of splines. The Bridge penalty $|\xi|^r$, $r \geq 0$, for a metric covariate x_j (Frank and Friedman, 1993) corresponds to $\mathcal{N}_l(\xi) = \|\xi\|_r$ with definitions $p_l(\xi) = \xi$ and $a_l^T\beta = \beta_j$; however, matters simplify a lot by employing the direct approximations for $r \in \{0, 1, 2\}$. The penalty proposed by Gertheiss and Tutz (2012) for categorical effect modifiers fits in the framework as well. In contrast to previous approaches, these penalties can be combined in one model whereby the response $y_i|x_i$ can follow any exponential family. Many models for other types of responses can be re-parametrized – such that they fit in the framework of general structured models. The Cox model (Cox, 1972) can be written as a time-discrete logit model (Fahrmeir and Tutz, 2001). Sequential models for ordinal response can be written as binary models, too (Tutz, 2011).

2.2 Approximation of the Penalty

To approximate the penalty (2), a first order Taylor expansion at $\boldsymbol{\beta}_{(k)}$ is employed. This approach extends Fan and Li (2001) and Ulbricht (2010). For the sake of simplicity, we write $\mathcal{N}_l(\cdot)$ and $\mathcal{D}_l(\cdot)$ respectively, for all penalty terms, even though not all norms have to be approximated. As in Fan and Li (2001), the approximation is

$$P_\lambda(\boldsymbol{\beta}) \approx P_\lambda(\boldsymbol{\beta}_{(k)}) + \nabla P_\lambda(\boldsymbol{\beta}_{(k)})^T \cdot (\boldsymbol{\beta} - \boldsymbol{\beta}_{(k)}),$$

where

$$\nabla P_\lambda(\boldsymbol{\beta}_{(k)})^T \cdot (\boldsymbol{\beta} - \boldsymbol{\beta}_{(k)}) = \sum_{l=1}^{L} \lambda_l \nabla p_l(\|\boldsymbol{a}_l^T\boldsymbol{\beta}_{(k)}\|_{\mathrm{N}_l})^T(\boldsymbol{\beta} - \boldsymbol{\beta}_{(k)}).$$

As Fan and Li (2001), in the following, we use the local approximation $\boldsymbol{a}_l^T\boldsymbol{\beta}/\boldsymbol{a}_l^T\boldsymbol{\beta}_{(k)} \approx 1$ for $\boldsymbol{\beta}_{(k)}$ close to $\boldsymbol{\beta}$. Moreover, $\boldsymbol{a}_l^T\boldsymbol{\beta}\boldsymbol{a}_l^T(\boldsymbol{\beta} - \boldsymbol{\beta}_{(k)})$ is approximated by $\frac{1}{2}(\boldsymbol{\beta}^T\boldsymbol{a}_l\boldsymbol{a}_l^T\boldsymbol{\beta} + \boldsymbol{\beta}_{(k)}^T\boldsymbol{a}_l\boldsymbol{a}_l^T\boldsymbol{\beta}_{(k)})$ via completing the square as proposed by Ulbricht, 2010. That gives

$$\begin{aligned}
\nabla p_l(\|\boldsymbol{a}_l^T\boldsymbol{\beta}_{(k)}\|_{\mathrm{N}_l})^T(\boldsymbol{\beta} - \boldsymbol{\beta}_{(k)}) &= \frac{\partial p_l(\|\boldsymbol{a}_l^T\boldsymbol{\beta}_{(k)}\|_{\mathrm{N}_l})}{\partial\|\boldsymbol{a}_l^T\boldsymbol{\beta}_{(k)}\|_{\mathrm{N}_l}} \cdot \frac{\partial\|\boldsymbol{a}_l^T\boldsymbol{\beta}_{(k)}\|_{\mathrm{N}_l}}{\partial\boldsymbol{a}_l^T\boldsymbol{\beta}_{(k)}} \cdot \frac{\partial\boldsymbol{a}_l^T\boldsymbol{\beta}_{(k)}}{\partial\boldsymbol{\beta}_{(k)}^T}(\boldsymbol{\beta} - \boldsymbol{\beta}_{(k)}) \\
&\approx p_l'(\|\boldsymbol{a}_l^T\boldsymbol{\beta}_{(k)}\|_{\mathrm{N}_l}) \cdot \frac{\mathcal{D}_l(\boldsymbol{a}_l^T\boldsymbol{\beta}_{(k)})}{\boldsymbol{a}_l^T\boldsymbol{\beta}_{(k)}}\boldsymbol{a}_l^T\boldsymbol{\beta} \cdot \boldsymbol{a}_l^T \cdot (\boldsymbol{\beta} - \boldsymbol{\beta}_{(k)}), \\
&\approx \frac{1}{2}p_l'(\|\boldsymbol{a}_l^T\boldsymbol{\beta}_{(k)}\|_{\mathrm{N}_l}) \cdot \frac{\mathcal{D}_l(\boldsymbol{a}_l^T\boldsymbol{\beta}_{(k)})}{\boldsymbol{a}_l^T\boldsymbol{\beta}_{(k)}}(\boldsymbol{\beta}^T\boldsymbol{a}_l\boldsymbol{a}_l^T\boldsymbol{\beta} + \boldsymbol{\beta}_{(k)}^T\boldsymbol{a}_l\boldsymbol{a}_l^T\boldsymbol{\beta}_{(k)}) \\
&= \frac{1}{2}(\boldsymbol{\beta}^T\boldsymbol{A}_l\boldsymbol{\beta} + \boldsymbol{\beta}_{(k)}^T\boldsymbol{A}_l\boldsymbol{\beta}_{(k)}),
\end{aligned}$$

where

$$\boldsymbol{A}_l = p_l'(\|\boldsymbol{a}_l^T\boldsymbol{\beta}_{(k)}\|_{\mathrm{N}_l}) \cdot \frac{\mathcal{D}_l(\boldsymbol{a}_l^T\boldsymbol{\beta}_{(k)})}{\boldsymbol{a}_l^T\boldsymbol{\beta}_{(k)}} \cdot \boldsymbol{a}_l\boldsymbol{a}_l^T.$$

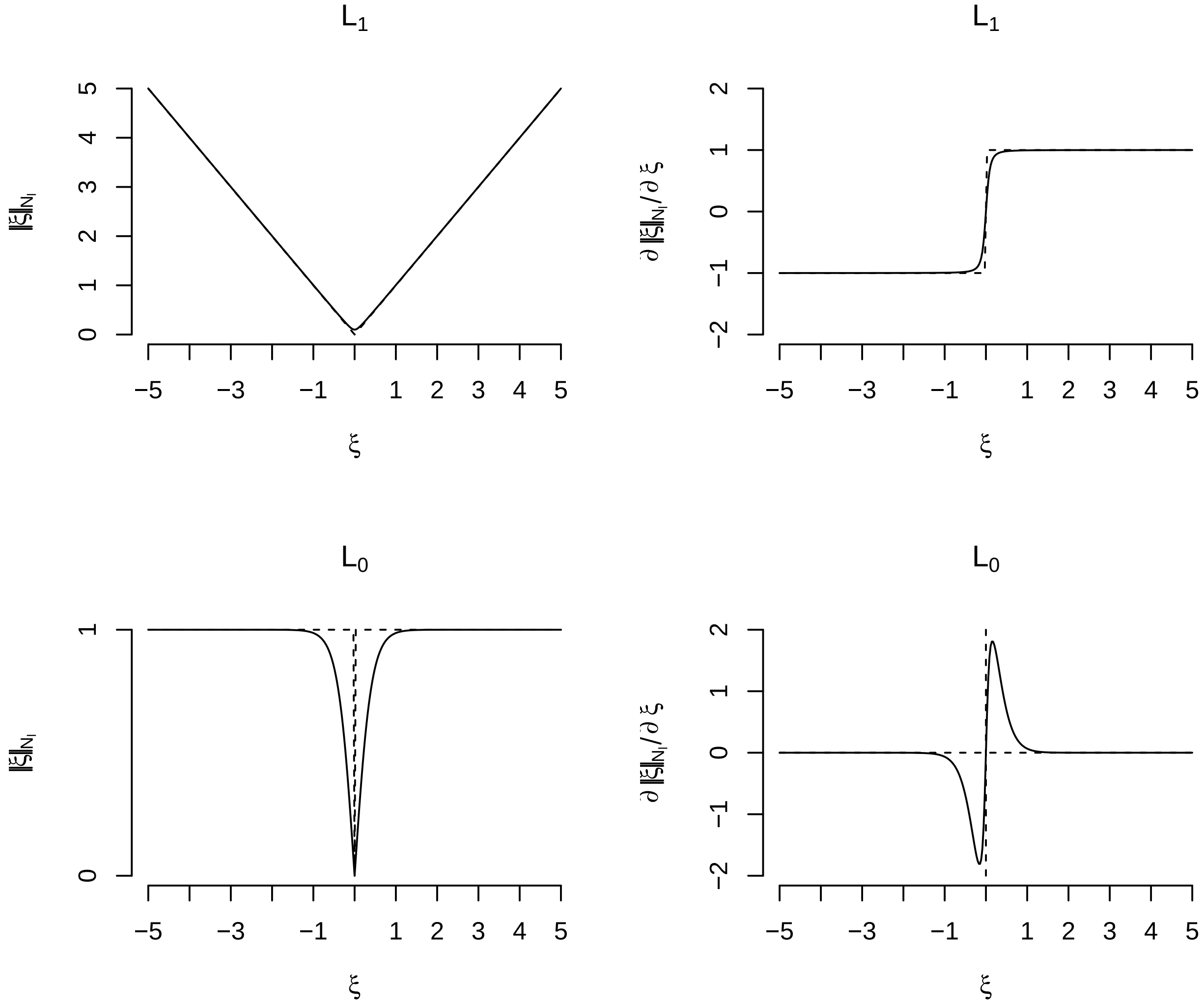

Figure 1: Graphical illustration for the approximation of the L_1 and the L_0 norm (left panel) and their derivatives (right panel) with respect to $\xi = \boldsymbol{a}_l^T\boldsymbol{\beta}$. On top, the approximation of the L_1 norm is shown. On bottom the L_0 norm is approximated by $\mathcal{N}_l(\xi) = 2(1+\exp(-\gamma|\xi|))^{-1} - 1$. The dashed lines mark the exact norms and the exact derivatives based on sub-gradients at $\xi = 0$. We use $c = 0.01$ and $\gamma = 5$. A similar figure for the approximation of the L_1 norm can be found in Ulbricht (2010).

Penalty	Covariate	Penalty Terms		
		$\|\xi\|_{\mathrm{N}_l}$	Penalty Function	$\boldsymbol{a}_l^T\boldsymbol{\beta} =$
Lasso	metric x_j	L_1	$p_l(\xi) = \xi$	β_j
Adaptive Lasso	metric x_j	L_1	$p_l(\xi) = \xi/\lvert a_l^T \hat{\beta}^{ML}\rvert$	β_j
Ridge	metric x_j	L_2^2	$p_l(\xi) = \xi$	β_j
SCAD	metric x_j	L_1	$p_l'(\xi) =$ $\left\{ I_{\xi \leq \lambda_l} + \frac{(a\lambda_l - \xi)_+}{(a-1)\lambda_l} I_{\xi > \lambda_l} \right\} I_{\xi \neq 0}$	β_j
Elastic net	metric x_j	1. L_1 2. L_2^2	$p_l(\xi) = \xi$ $p_k(\xi) = \xi$	β_j β_j
Fused Lasso	k_j ordered covariates x_j, $j = s, \ldots, t$	L_1	$p_l(\xi) = \xi$	$\beta_j - \beta_{j-1}$, $j = s+1, \ldots, t$
Penalized B-splines	$f(x_j)$, x_j metric, parametrized by k_j coeff. β_{jk}	L_2^2	$p_l(\xi) = \xi$	$\beta_{jk} - 2\beta_{j,k-1} + \beta_{j,k-2}$ $k = 3, \ldots, k_j$
Simultaneous factor selection (Bondell and Reich, 2009)	nominal factor x_j with k_j coeff. β_{jk}, $k = 1, \ldots, k_j$	L_1	$p_l(\xi) = \xi$	$\beta_{jk} - \beta_{jr}$, $k > r \geq 0$
Sparse modeling of cat. variables (Gertheiss and Tutz, 2010)	ordinal factor x_j with k_j coeff. β_{jk}, $k = 1, \ldots, k_j$	L_1	$p_l(\xi) = \xi$	$\beta_{jk} - \beta_{j,k-1}$, $k = 1, \ldots, k_j$
Penalty for signals (Rippe et al., 2012)	k_j ordered covariates x_j, $j = s, \ldots, t$	L_0	$p_l(\xi) = \xi$	$\beta_j - \beta_{j-1}$ $j = s+1, \ldots, t$

Table 2: Examples for approximations of known penalties. The elastic net is made up by two terms with separate tuning parameters λ_l and λ_k each. All the other penalties are governed by one penalty parameter λ_l, even when they are defined by several terms. The fused Lasso consists out of $k_j - 1$ terms related to divers differences. The same holds for penalized B-splines ($k_j - 2$ penalty terms), the penalties in Bondell and Reich (2009) ($\frac{1}{2}(k_j + 1)k_j$ terms) and in Gertheiss and Tutz (2010) (k_j terms). In penalty terms for factors, the coefficient $\beta_{j0} = 0$ relates to the reference category; it is considered in the differences $\boldsymbol{a}_l^T\boldsymbol{\beta}$. The SCAD penalty (Fan and Li, 2001) is defined by its derivative; parameter a, $a > 2$, is an additional tuning parameter. Fan and Li (2001) recommend $a = 3.7$. In general, the derivatives of functions p_l are amended by the factor $I_{\xi \neq 0}$ to ensure that $\boldsymbol{A}_l$ is a zero matrix $(\mathbf{0})_{(q \times q)}$ when $\left\|\boldsymbol{a}_l^T \boldsymbol{\beta}_{(k)}\right\|_{\mathrm{N}_l} = 0$.

With $\boldsymbol{A}_\lambda = \sum_{l=1}^{L} \lambda_l \boldsymbol{A}_l$, the penalty is locally quadratically approximated by

$$P_\lambda(\boldsymbol{\beta}) \quad \approx \quad P_\lambda(\boldsymbol{\beta}_{(k)}) + \frac{1}{2}(\boldsymbol{\beta}^T \boldsymbol{A}_\lambda \boldsymbol{\beta} + \boldsymbol{\beta} a_{(k)}^T \boldsymbol{A}_\lambda \boldsymbol{\beta}_{(k)}). \tag{3}$$

This approach shares the format of Fan and Li (2001) and Ulbricht (2010). However, in contrast to Ulbricht (2010), it allows to approximate arbitrary terms that are meaningful as a penalty in various ways.

The penalized versions of the score function and the Hessian matrix are $\boldsymbol{s}_{pen}(\boldsymbol{\beta}) = \boldsymbol{s}(\boldsymbol{\beta}) - \boldsymbol{A}_\lambda \boldsymbol{\beta}$ and $\boldsymbol{H}_{pen}(\boldsymbol{\beta}) = \boldsymbol{H}(\boldsymbol{\beta}) - \boldsymbol{A}_\lambda$. By using the penalized score function, one obtains essentially the same optimization problem as for usual GLMs. When solving the linearized problem $\boldsymbol{s}_{pen}^{lin}(\boldsymbol{\beta}) = \boldsymbol{0}$ iteratively, one obtains $\hat{\boldsymbol{\beta}}_{(k+1)} = \hat{\boldsymbol{\beta}}_{(k)} - \boldsymbol{H}_{pen}(\hat{\boldsymbol{\beta}}_{(k)})^{-1} \boldsymbol{s}_{pen}(\hat{\boldsymbol{\beta}}_{(k)})$. To stabilize the estimation, we use the Fisher information matrix $F(\boldsymbol{\beta}) = -\mathbb{E}(\boldsymbol{H}(\boldsymbol{\beta}))$. With the usual derivation, the corresponding PIRLS algorithm with step length parameter ν, is

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{(k+1)} &= \hat{\boldsymbol{\beta}}_{(k)} - \nu \cdot (-\boldsymbol{F}(\hat{\boldsymbol{\beta}}_{(k)}) - \boldsymbol{A}_\lambda)^{-1}(\boldsymbol{s}(\hat{\boldsymbol{\beta}}_{(k)}) - \boldsymbol{A}_\lambda \hat{\boldsymbol{\beta}}_{(k)}) \\
&= \hat{\boldsymbol{\beta}}_{(k)} - \nu \cdot (\boldsymbol{F}(\hat{\boldsymbol{\beta}}_{(k)}) + \boldsymbol{A}_\lambda)^{-1}(-\boldsymbol{s}(\hat{\boldsymbol{\beta}}_{(k)}) + \boldsymbol{A}_\lambda \hat{\boldsymbol{\beta}}_{(k)}) \\
&= \hat{\boldsymbol{\beta}}_{(k)} - \nu \cdot (\boldsymbol{X}^T \boldsymbol{W}_{(k)} \boldsymbol{X} + \boldsymbol{A}_\lambda)^{-1}[-\boldsymbol{X}^T \boldsymbol{W}_{(k)} (\underbrace{\boldsymbol{D}_{(k)}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_{(k)}) + \quad \boldsymbol{X} \hat{\boldsymbol{\beta}}_{(k)}}_{\tilde{\boldsymbol{y}}_{(k)}} - \boldsymbol{X} \hat{\boldsymbol{\beta}}_{(k)}) \quad + \boldsymbol{A}_\lambda \hat{\boldsymbol{\beta}}_{(k)}] \\
&= (1 - \nu) \cdot \hat{\boldsymbol{\beta}}_{(k)} + \nu \cdot (\boldsymbol{X}^T \boldsymbol{W}_{(k)} \boldsymbol{X} + \boldsymbol{A}_\lambda)^{-1} \boldsymbol{X}^T \boldsymbol{W}_{(k)} \tilde{\boldsymbol{y}}_{(k)}.
\end{aligned} \tag{4}$$

Assuming a simple exponential family for $y_i | \boldsymbol{x}_i$, $i = 1, \ldots, n$, allows to define $\boldsymbol{F}(\hat{\boldsymbol{\beta}} a_{(k)}) = \boldsymbol{X}^T \boldsymbol{D}_{(k)} \boldsymbol{\Sigma}_{(k)}^{-1} \boldsymbol{D}_{(k)} \boldsymbol{X} = \boldsymbol{X}^T \boldsymbol{W}_{(k)} \boldsymbol{X}$, $\boldsymbol{D}_{(k)} = \mathrm{diag}(\partial h(\eta_i(\hat{\boldsymbol{\beta}}_{(k)}))/\partial \boldsymbol{\eta})$, and $\boldsymbol{\Sigma}_{(k)} = \mathrm{diag}(\sigma_i^2(\hat{\boldsymbol{\beta}}_{(k)}))$, as well as $\boldsymbol{s}(\hat{\boldsymbol{\beta}}_{(k)}) = \boldsymbol{X}^T \boldsymbol{W}_{(k)} \boldsymbol{D}_{(k)}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_{(k)})$, $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, and $\boldsymbol{\mu}_{(k)} = (h(\boldsymbol{x}_1^T \hat{\boldsymbol{\beta}}_{(k)}), \ldots, h(\boldsymbol{x}_n^T \hat{\boldsymbol{\beta}}_{(k)}))^T$. Starting with an initial value $\hat{\boldsymbol{\beta}}_{(0)}$, this algorithm is iterated until convergence. The algorithm is determined when $|\hat{\boldsymbol{\beta}}_{(k+1)} - \hat{\boldsymbol{\beta}}_{(k)}| / |\hat{\boldsymbol{\beta}}_{(k)}| \leq \epsilon$, for fixed $\epsilon > 0$.

The parameter ν, $0 < \nu \leq 1$, in the algorithm is a step length parameter. In unpenalized Fisher scoring algorithms, $\nu = 1$; only, if it is necessary, the step length is halved. However, when the objective function is nonstandard, it can be helpful to work with $\nu < 1$ to control the convergence of the algorithm and to avoid back-fitting steps.

2.3 Some Technical Comments

Newton-type algorithms are not unconditionally convergent. When the penalized Fisher information matrix $\boldsymbol{X}^T \boldsymbol{W}_{(k)} \boldsymbol{X} + \boldsymbol{A}_\lambda$ is positive definite, the optimization problem is strictly convex and a descent direction in each iteration of algorithm (4) is guaranteed. If a solution exists, the algorithm (almost) surely converges to the optimum – independently of the initial value $\hat{\boldsymbol{\beta}}_{(0)}$. The penalized Fisher information matrix is positive definite, when the Fisher information $\boldsymbol{X}^T \boldsymbol{W}_{(k)} \boldsymbol{X}$ and the penalty matrix $\boldsymbol{A}_\lambda$ are positive definite; or when one of the two matrices is positive and the other is positive semi-definite; in some cases, the penalized Fisher information will be positive definite for a positive semi-definite Fisher information $\boldsymbol{X}^T \boldsymbol{W}_{(k)} \boldsymbol{X}$ and a positive semi-definite penalty matrix $\boldsymbol{A}_\lambda$.

For simple exponential families, like assumed here, the negative log-likelihood $l_n(\boldsymbol{\beta})$ is convex. Hence, when the number of different observations is larger than the number of parameters $(n > q)$, the Fisher information is positive definite; the penalty matrix has to be at least positive semi-definite to assure the global convergence of algorithm (4). The penalty is convex when the functions $p_l(\xi)$ and $\|\xi\|_{\mathrm{N}_l}$ are convex for all l, and $p_l(\xi)$ is monotonically increasing as assumed. This is for example, the case for the (adaptive) Lasso, the fused Lasso or the Ridge penalty. It does not apply to the SCAD penalty or the L_0 penalty of Rippe et al. (2012).

In the $n < q$ case, the Fisher information will be positive semi-definite; if the penalty matrix is positive definite, the algorithm's convergence is assured. However, when the penalized Fisher information matrix is positive semi-definite, algorithm (4) will find descent directions in each iteration; and it can happen that these directions are not unique.

In Ulbricht (2010), corresponding properties of PIRLS algorithms are described.

It is possible to estimate the degrees of freedom via the generalized hat matrix

$$\boldsymbol{H}_{(k*)} = \boldsymbol{W}_{(k*)}^{T/2}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}_{(k*)}\boldsymbol{X} + \boldsymbol{A}_\lambda)^{-1}\boldsymbol{X}^T\boldsymbol{W}_{(k*)}^{1/2}.$$

The approximated hat matrix is symmetric but not idempotent or a projection matrix. In contrast to Ulbricht (2010), we prefer to estimate the hat matrix based on the estimates of the final iteration. The estimated degrees of freedom for a given tuning are then the trace of the hat matrix:

$$\mathrm{df} = \mathrm{tr}(\boldsymbol{H}_{(k*)}).$$

In some exponential families, there is a scale parameter $\phi \neq 1$. As ϕ and $\boldsymbol{\beta}$ are orthogonal (see the mixed second derivatives $\frac{\partial l_n}{\partial\phi\partial\boldsymbol{\beta}}$ given in Claeskens and Hjort, 2008), and when the applied penalty gives consistent estimates, one can plug in $\hat{\phi}$ instead of ϕ; that corresponds to quasi-likelihood approaches.

2.4 Tuning and Computational Issues

The performance of local quadratic approximations depends on the accuracy of the approximations $\mathcal{N}_l(\xi)$ and the choice of tuning parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_L)^T$. Of course, the more precise approximations $\mathcal{N}_l(\xi)$ are, the more accurate is the proposed algorithm. The choice of the tuning parameters is more complex, because one has to find L possibly different parameters. However, penalized regression requires standardized data or data that is measured on a comparable scale. Given the data is standardized and the penalty terms are comparable, many approaches use one global tuning parameter. Bondell and Reich (2009) illustrate that weighting the penalty terms adequately gives the same effect as standardization of the predictors. This is especially helpful for categorical covariates that are hard to standardize and that can result in many penalty terms. As proposed by Bondell and Reich (2009), the penalty terms connected to a covariate x_j are weighted such that they are of order k_j, the number of (free) coefficients related to x_j. When, for example, a nominal factor with $k_j + 1$ categories is penalized by fused Lasso terms as proposed by Gertheiss and Tutz (2010), all pairwise differences of the k_j related dummy coefficients and the reference category are penalized absolutely. This results in $\frac{1}{2}(k_j + 1)k_j$ differences. Hence, the difference of the dummies β_{jl} and β_{jm} is weighted by the factor

$$\frac{2}{k_j + 1}\sqrt{\frac{n_{jl} + n_{jm}}{n}},$$

where n_{jl} and n_{jm} denote the number of observations on the levels l and m of the predictor x_j. Weights for other penalties are derived analogously.

To allow for comparisons with conventional methods, we choose the global tuning parameter λ by cross-validation. In numerical experiments, we employ K-fold cross-validation with the

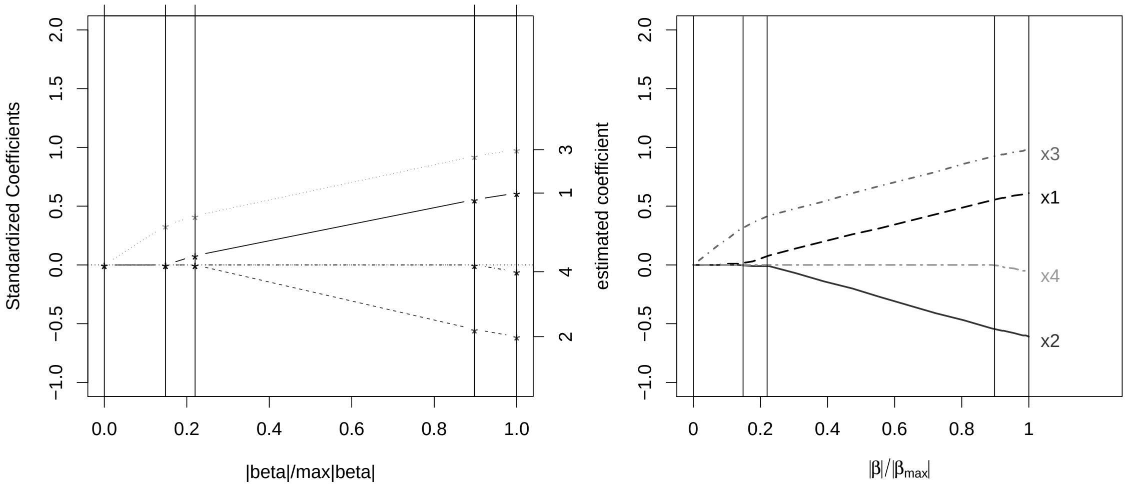

Figure 2: Coefficient paths for a linear model with an intercept and four metric predictors penalized by the Lasso. On the left, the paths are computed by the lars algorithm; on the right, the penalty is approximated quadratically. In both panels, the path related to the intercept is omitted.

predictive deviance as criterion or a generalized cross-validation criterion (GCV) as proposed by O'Sullivan et al. (1986) and used in the R package mgcv. Both approaches seem to work reasonable for the proposed approximations.

Even though the proposed algorithm can combine a variety of penalties, it is easy to implement. In principle, it can be combined with any PIRLS algorithm – given, that additional quadratic penalties may depend on estimates of the last iteration. Except for the approximation of the penalty, the computational complexity is the same as for Fisher scoring algorithms. The penalties mentioned here are implemented in the R package gvcm.cat (Oelker, 2013).

2.5 Extension to Vector-Valued Arguments

The penalties mentioned so far assume linear transformations $\boldsymbol{a}_l^T\boldsymbol{\beta}$ of the coefficient vector. That is, all norms $\|\xi\|_{\mathrm{N}_l}$ have scalars as arguments. However, penalties employing vectorial norms can be approximated in the same way. For illustration, in this section, the grouped Lasso (Yuan and Lin, 2006) is considered. The grouped Lasso penalty is defined for a subvector of coefficients $\boldsymbol{\beta}_l$ as

$$\lambda_l(\boldsymbol{\beta}_l^T\boldsymbol{K}_l\boldsymbol{\beta}_l)^{1/2} = \lambda_l\,\|\boldsymbol{\beta}_l\|_{\boldsymbol{K}_l}\,, \tag{5}$$

where the matrix $\boldsymbol{K}_l \in \mathbb{R}^{r\times r}$ is symmetric and positive (semi-)definite; typically, it is an identity matrix. The penalty implies that coefficients in $\boldsymbol{\beta}_l$ are shrunken in a way that the whole group of coefficients is set to zero. The grouped Lasso penalty (5) can be rewritten as

$$\lambda_l p_l(\|\boldsymbol{R}_l\boldsymbol{\beta}\|_2), \tag{6}$$

where $p_l(\xi) = \xi$. The norm $\|\cdot\|_2$ is the Euclidean norm and the matrix $\boldsymbol{R}_l \in \mathbb{R}^{q\times q}$ yields $\boldsymbol{\beta}^T\boldsymbol{R}_l^T\boldsymbol{R}_l\boldsymbol{\beta} = \boldsymbol{\beta}_l^T\boldsymbol{K}_l\boldsymbol{\beta}_l$. $\|\cdot\|_2$ corresponds to the norm $\|\xi\|_{\mathrm{N}_l}$ in penalty (2). It can be approximated by $\|\boldsymbol{\xi}\|_2 \approx (\boldsymbol{\xi}^T\boldsymbol{\xi} + c)^{1/2}$, where c is a small positive real number. Following the same schedule as in Subsection 2.2, an approximation of the penalty's gradient at $\boldsymbol{\beta} = \boldsymbol{\beta}_{(k)}$ is

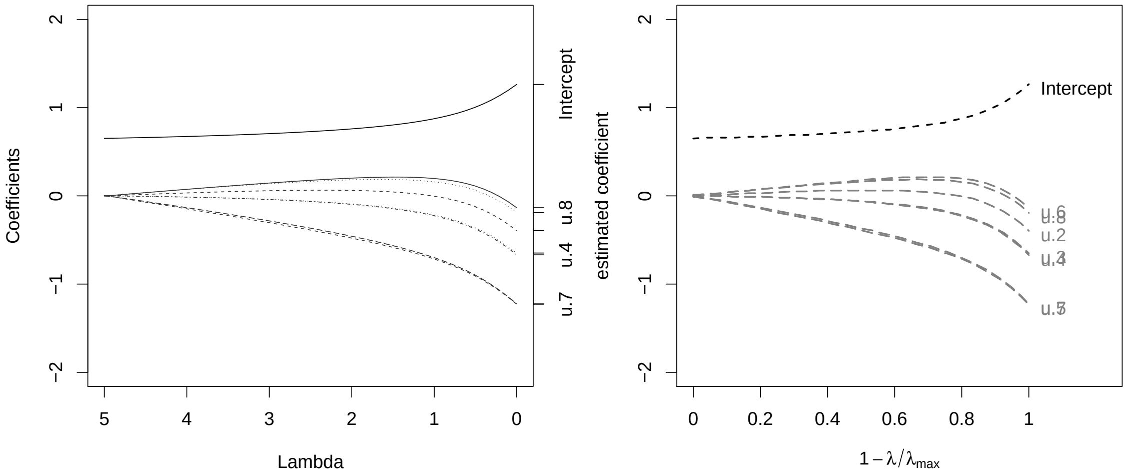

Figure 3: Coefficient paths for a logistic model with an intercept and one ordered factor (8 levels) as predictor. The coefficients are penalized by a Grouped Lasso penalty. On the left, the path is computed with `R` package `grplasso`; on the right, the proposed quadratic approximation is employed.

obtained by:

$$
\begin{aligned}
\nabla p_l(\|\boldsymbol{R}_l\boldsymbol{\beta}_{(k)}\|_2) &= \frac{\partial p_l(\|\boldsymbol{R}_l\boldsymbol{\beta}_{(k)}\|_2)}{\partial\|\boldsymbol{R}_l\boldsymbol{\beta}_{(k)}\|_2}\cdot\frac{\partial\|\boldsymbol{R}_l\boldsymbol{\beta}_{(k)}\|_2}{\partial(\boldsymbol{R}_l\boldsymbol{\beta}_{(k)})^T\boldsymbol{R}_l\boldsymbol{\beta}_{(k)}}\cdot\frac{\partial(\boldsymbol{R}_l\boldsymbol{\beta}_{(k)})^T\boldsymbol{R}_l\boldsymbol{\beta}_{(k)}}{\partial\boldsymbol{\beta}_{(k)}}\\
&\approx p_l'(\|\boldsymbol{R}_l\boldsymbol{\beta}_{(k)}\|_2)\cdot\frac{1}{2((\boldsymbol{R}_l\boldsymbol{\beta}_{(k)})^T\boldsymbol{R}_l\boldsymbol{\beta}_{(k)}+c)^{1/2}}\cdot 2\boldsymbol{R}_l^T\boldsymbol{R}_l\boldsymbol{\beta},
\end{aligned}
$$

Yielding the approximation

$$
\nabla p_l(\|\boldsymbol{R}_l\boldsymbol{\beta}_{(k)}\|_2)^T(\boldsymbol{\beta}-\boldsymbol{\beta}_{(k)}) \approx \frac{1}{2}(\boldsymbol{\beta}^T\boldsymbol{A}_l^{gr}\boldsymbol{\beta}-\boldsymbol{\beta}_{(k)}^T\boldsymbol{A}_l^{gr}\boldsymbol{\beta}_{(k)}), \tag{7}
$$

for $\boldsymbol{\beta}_{(k)}$ close to $\boldsymbol{\beta}$ and $\boldsymbol{A}_l^{gr}=p_l'(\|\boldsymbol{R}_l\boldsymbol{\beta}_{(k)}\|_2)((\boldsymbol{R}_l\boldsymbol{\beta}_{(k)})^T\boldsymbol{R}_l\boldsymbol{\beta}_{(k)}+c)^{-1/2}\boldsymbol{R}_l^T\boldsymbol{R}_l$.
In contrast to so far employed matrices $\boldsymbol{A}_l$, matrix $\boldsymbol{A}_l^{gr}$ is spanned by the product $\boldsymbol{R}_l^T\boldsymbol{R}_l$ and not by the product of vectors $\boldsymbol{a}_l\boldsymbol{a}_l^T$. Expression (7) fits exactly into the framework of approximation (3). Penalties of type (6) can be added to penalty (2) without any problems.
To implement, for example, the penalty of Gertheiss et al. (2011), $\boldsymbol{R}_l\boldsymbol{\beta}$ is a vector of differences of coefficients related to an ordinal factor of the form $\beta_{jk}-\beta_{j,k-1}$, $k=1,\ldots,k_j$. To obtain a penalty term of a comparable order, weights w_l are set to $\sqrt{r_l}$, where r_l denotes the number of differences in vector $\boldsymbol{R}_l\boldsymbol{\beta}$.
Meier et al. (2009) propose a sparsity-smoothness penalty for high-dimensional additive models; it can be written as a grouped Lasso-type penalty and fits in the framework of penalty (2), too.

3 Illustrations

In this section, the proposed method is illustrated. In order to show that the approximations work, we compare the coefficient paths of penalties computed by different algorithms. We give an example for a model with a L_0 type penalty. Finally, we illustrate how the approximation works for penalized smooth functions.

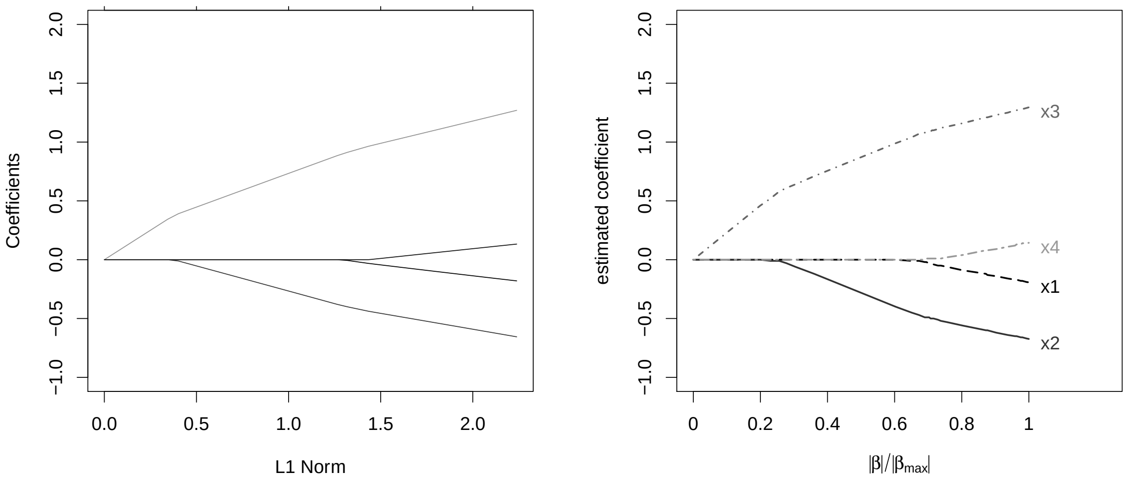

Figure 4: Coefficient paths for a logistic model with an intercept and four metric predictors; each predictor is regularized by an elastic net penalty. On the left, R package glmnet is employed for computation; on the right, the elastic net penalty is quadratically approximated. In both panels, the path related to the intercept is omitted.

3.1 Comparison of Methods

When the penalty consists of one norm only, one can compare different algorithms with the proposed quadratic approximation. Yet, the results depend on many parameters: on the choice of the tuning parameters for the approximation, on the choice of λ, on the criterion chosen for cross-validation, on the folds for cross-validation and so on. Hence to judge how the proposed approximation works, we compare the coefficient paths of different penalties (the Lasso, the grouped Lasso, the elastic net).

At first, the approximation of the Lasso is compared with the solution of the lars algorithm. We consider a linear model with four continuous predictors and $n = 400$ observations. The four predictors are drawn from a Uniform distribution on $(0, 2)$. The predictor of the model for an observation i is denoted by

$$\eta_i^{lasso} = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + x_{i4}\beta_4, \tag{8}$$

where β_0 denotes the intercept of the model. The true coefficient vector is $\boldsymbol{\beta}^{true} = (0.2, 0.7, -0.5, 1, 0)^T$. That is, there is one non-influential predictor to detect. Figure 2 shows the resulting coefficient paths. The left panel shows the solution computed by lars. There are four break points in the piecewise linear coefficient path, each marked by a vertical line. In the right panel, the coefficient path obtained with the proposed quadratic approximation is shown. The vertical lines mark the break points of the lars solution; they correspond exactly to the break points of the quadratic approximation.

In what follows, we assume a logistic model. The true predictor is

$$\eta_i^{logistic} = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + x_{i4}\beta_4 + u_{i5}^T\beta_5,$$

where u_5 is an ordered factor with 8 levels; it is dummy coded and drawn from a multinomial distribution with equal probability for each level. $x_1, \ldots, x_4$ are continuous predictors drawn from a Uniform distribution on $(0, 2)$. The data generating coefficients are $\boldsymbol{\beta}^{true} =$

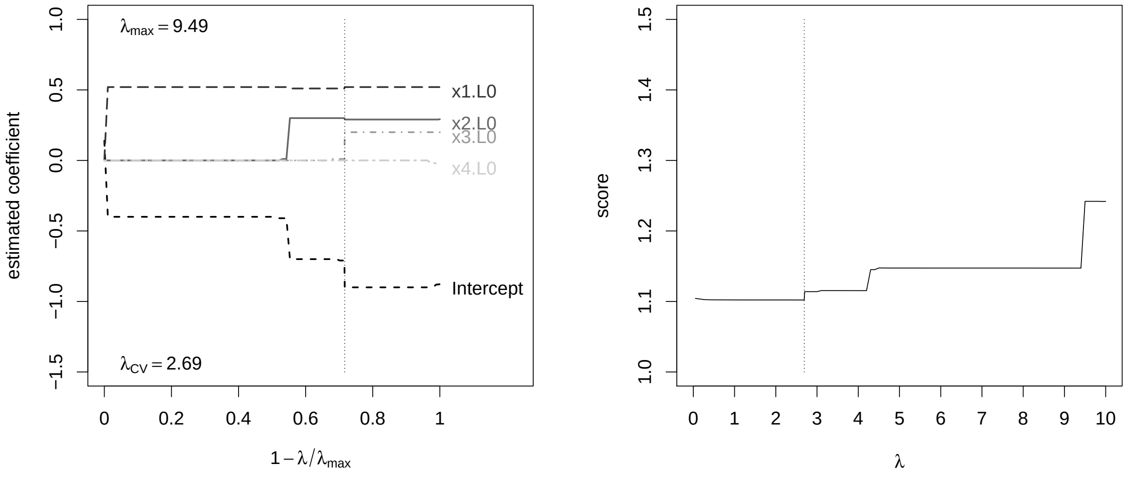

Figure 5: Coefficient paths (left panel) and GCV score (right panel) for a Poisson model with an intercept and four metric predictors; each predictor is regularized by an L_0 penalty. As the GCV score has no unique minimum, λ_{CV} is the maximal penalization parameter that minimizes the GCV score.

$(0.2, 0, -.5, 1, 0, , 0.3, .7, .7, 0.4, 0.4, 0.4, 0.4)^T$; that is, β_5 is a vector of seven coefficients corresponding to the dummies of u_5.

We consider two models: In the first one, the predictor is $\eta_i^{grouped} = \beta_0 + u_{i5}^T\beta_5$. It contains the dummy coded ordered factor only. The dummy coefficients are penalized by a grouped Lasso penalty. We compare the solution of the coordinate-descent algorithm proposed by Meier et al. (2008) in the R package `grplasso` (Meier, 2013) with the quadratic approximation. Figure 3 shows the coefficient paths. In contrast to Figure 2, the path of the intercept is added; the x-axis depends on the (scaled) values of λ instead of $\|\boldsymbol{\beta}\|_1$. Again, structure and range of the two paths are almost identical.

In the second model, the predictor is $\eta_i^{elastic} = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + x_{i4}\beta_4$; that is the influential factor u_5 is ignored. All coefficients are penalized by the elastic net. Figure 4 illustrates the resulting coefficient paths. On the left, the paths are computed by the coordinate descent algorithm of Friedman et al. (2010) available in R package `glmnet`. On the right, the paths obtained with the proposed local quadratic approximation is shown. Again, the two solutions coincide.

3.2 Penalties Based on the L_0 Norm

Apart from well known penalties like the Lasso or the elastic net that are based on the L_1 norm or on Ridge type penalties, alternative penalties are made available by our approach. In this Section, we consider a model for count data with Poisson distributed responses. The model contains an intercept and four metric covariates, whereby the predictor x_4 is non-influential: $\beta^{true} = (-1, 0.5, 0.4, 0.2, 0)^T$. The ideal penalty should uncover that the effect of this predictor is zero – without any shrinkage effects on the other coefficients. Therefore, we use the L_0 penalty

$$P_\lambda(\beta) = \lambda \sum_{l=1}^{4} \|\beta_l\|_0 \,,$$

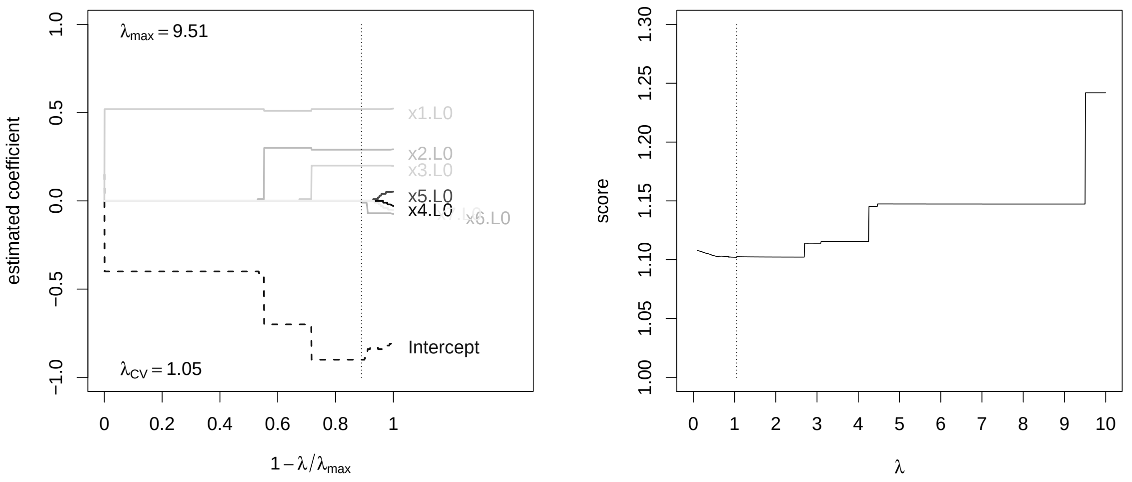

Figure 6: Coefficient paths (left panel) and GCV score (right panel) for a Poisson model with an intercept and seven metric predictors; whereby four predictors are truly non-influential. Each predictor is regularized by an L_0 penalty.

where $||\xi||_0$ denotes $I_{\xi \neq 0}$. This penalty is neither convex nor concave. The solution obtained for a set of initial values $\boldsymbol{\beta}_{(0)}$ has not to be the global optimum. However, starting with $\boldsymbol{\beta}_{(0)} = (0)^T$ works well. The tuning parameters of the approximation are set to $c = 10^{-5}$ and $\gamma = 50$. In the left panel of Figure 5, the coefficient paths for the considered model are shown. The dotted line marks the model chosen by the generalized cross-validation criterion of O'Sullivan et al. (1986). For $\lambda_{CV} = 2.69$, the coefficient related to x_4 is zero; the remaining coefficients are not shrunken, the mean squared error is 0.0225 and hence, relatively small. The right panel of Figure 5 shows the GCV score; like the coefficient paths, it is a step function. As the GCV score has no unique minimum, λ_{CV} is defined to be the maximal penalization parameter that minimizes the GCV score.

L_0 penalization is challenging, when there are several non-influential predictors, because the penalty is neither concave nor convex. Hence, to challenge the proposed approximation, the above setting is extended by three more non-influential predictors x_5, x_6 and x_7. $\boldsymbol{\beta}^{true}$ is $(-1, 0.5, 0.4, 0.2, 0, 0, 0, 0)^T$; that is, half of the coefficients is truly zero. Figure 6 shows the resulting coefficients paths (left) and the GCV score (right). For the optimal model, $\lambda_{CV} = 1.05$. All but one truly zero coefficient is detected ($\hat{\beta}_6 = -0.01$); the mean squared error is 0.0226. For $\lambda \in (1.05, 2.69]$, the true model is detected. As there are only marginal difference in the GCV score for $\lambda_{CV} = 1.05$ and $\lambda = 2.69$, one would probably choose $\lambda_{CV} = 1.69$; and hence, the right model.

3.3 Nonparametric Terms

In many applications, the effect of a continuous covariate is non-linear and one wants to allow for unspecified smooth functions in the predictor. As it is a common choice, we assume that the smooth functions are modeled by penalized cubic B-splines with equidistant knots $\kappa_1, \dots, \kappa_{M_j}$ as proposed by Eilers and Marx (1996). That is, we assume that $f_j(\boldsymbol{x}_j)$ is represented by $\boldsymbol{B}_j \boldsymbol{\beta}_j$

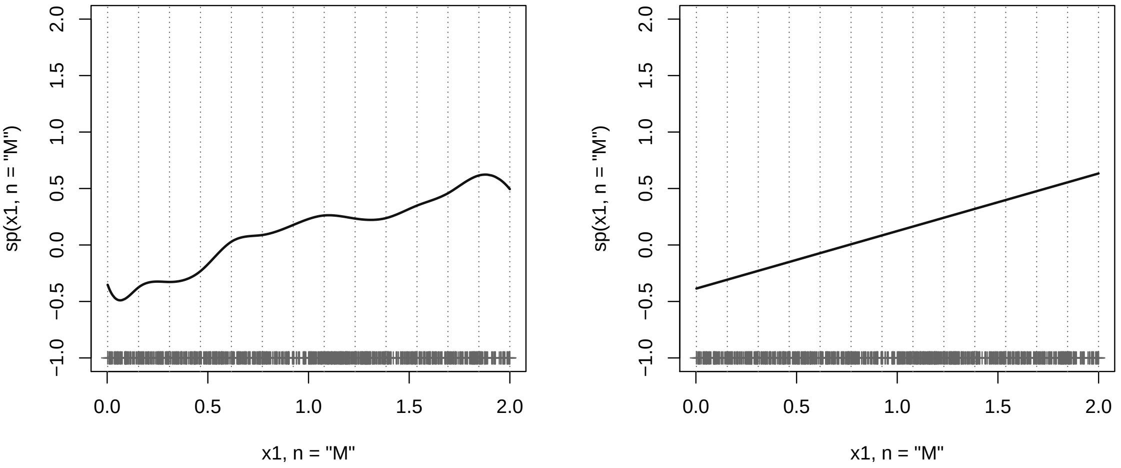

Figure 7: Resulting estimate for the predictor $\eta_i = \beta_0 + f_1(x_1)$ in a model with Poisson distributed response. The data generating effect of x_1 is linear. $f_1(x_1)$ is represented by transformed B-spline basis evaluations with 20 knots and penalized like described below. On the left, $\lambda_1 = \lambda_2 = 0$ holds; on the right, λ is chosen by cross-validation. Dotted vertical lines mark the knots of the underlying B-spline basis evaluations.

where $\boldsymbol{B}_j \in \mathbb{R}^{n\times(M_j-4)}$ is the matrix of basis function evaluations, and $\boldsymbol{\beta}_j$ is penalized by

$$\sum_{i=1}^{M_j-6} (\beta_{ji} - 2\beta_{j,i+1} + \beta_{j,i+2})^2 = \boldsymbol{\beta}_j^T (\boldsymbol{\Delta}^2)^T \boldsymbol{\Delta}^2 \boldsymbol{\beta}_j, \tag{9}$$

where $\boldsymbol{\Delta}^2 \in \mathbb{R}^{(M_j-6)\times M_j}$ denotes the matrix of second order differences with full row rank M_j-6. An attractive approach that centers the smooth function $f_j(\boldsymbol{x}_j) = \boldsymbol{B}_j\boldsymbol{\beta}_j$ for a given set of knots and that offers a decomposition of the function into a linear and a non-linear part is based on the representation of Fahrmeir et al. (2004). The coefficient vector $\boldsymbol{\beta}_j$ is decomposed into a linear part $\boldsymbol{\beta}_j^{lin} = (\beta_j^{int}, \beta_j^{slope})^T$ and into coefficients $\boldsymbol{\beta}_j^{nonlin}$ that model the deviation from the linear trend. One obtains

$$\boldsymbol{\beta}_j = \boldsymbol{\Psi}^{lin}\boldsymbol{\beta}_j^{lin} + \boldsymbol{\Psi}^{nonlin}\boldsymbol{\beta}_j^{nonlin},$$

with

$$\boldsymbol{\Psi}^{lin} = \begin{pmatrix} 1 & \kappa_1 \\ 1 & \kappa_2 \\ \vdots & \vdots \\ 1 & \kappa_{M_j-4} \end{pmatrix}$$

and with $\boldsymbol{\Psi}^{nonlin} = (\boldsymbol{\Delta}^2)^T \left(\boldsymbol{\Delta}^2(\boldsymbol{\Delta}^2)^T\right)^{-1}$. It holds, that $\boldsymbol{\Delta}^2\boldsymbol{\Psi}^{lin} = \boldsymbol{0}$ and that $\boldsymbol{\Psi}^{lin}\boldsymbol{\Delta}^2 = \boldsymbol{0}$. That is, $\boldsymbol{\Psi}^{nonlin} \in \mathbb{R}^{(M_j-4)\times(M_j-6)}$ represents the space of penalty (9); $\boldsymbol{\Psi}^{lin} \in \mathbb{R}^{(M_j-4)\times 2}$ its nullspace. β_j^{int} is incorporated in the (global) intercept β_0; β_j^{slope} and $\boldsymbol{\beta}_j^{nonlin}$ represent the centered smooth function $f_j(\boldsymbol{x}_j)$. Instead of second order differences, the penalty $\sum_{i=1}^{M_j-6}(\beta_{ji}^{nonlin})^2$ is sufficient. Hence, we obtain the same effect as Eilers and Marx (1996) by means of a structured representation with a less complex penalty.

The decomposition of Fahrmeir et al. (2004) allows to distinguish linear and nonlinear functions more easily: We apply a grouped Lasso penalty on the coefficients $\boldsymbol{\beta}_j^{nonlin}$; if the deviation from the linear trend is selected, it is selected at once. Moreover, the slope β_j^{slope} is penalized by a

	Variable	Description
	rent	the rent of the flat, response
1	numbrooms	the number of rooms in the flat, ordered factor with 6 levels, dummy coded, reference category are flats with one room
2	location	the urban district in which the flat is, nominal factor with 25 levels, dummy coded, reference is category 1, that is the city center
3	age	the age (in years) of the flat in 2007, continuous covariate
4	residentialarea	the residential area of the flat, continuous covariate

Table 3: Details of the variables in the Munich rent dataset (Fahrmeir et al., 2007).

Lasso penalty:

$$P_j(\boldsymbol{\beta}_j) = \alpha|\beta_j^{slope}| + (1-\alpha)\left\|\boldsymbol{\beta}_j^{nonlin}\right\|_2, \tag{10}$$

where α is an additional tuning parameter that allows to weigh the two parts of the penalty separately. Hence, depending on the tuning, the smooth function f_j can be estimated to be nonlinear, linear or non-influential. We consider the same Poisson data as in Section 3.2. The impact of all covariates is linear. Even though, we fit a model with the predictor

$$\eta_i = \beta_0 + f_1(x_1)$$

and penalty (10) for $f_1(\boldsymbol{x}_1)$ which is represented by β_1^{slope} and $\boldsymbol{\beta}_1^{nonlin}$. Figure 7 shows the resulting functions $f_1(\boldsymbol{x}_1)$ for $\lambda_1 = \lambda_2 = 0$ (left panel) and for cross-validated tuning (right panel). The effect is detected to be linear.

4 Rents in Munich

Most major cities and many large communities in Germany conduct surveys to construct and publish rental guides. These guides are consulted to determine suitable rents for public and private properties. To model the rent of 1488 households in the city of Munich collected in 2007 (Fahrmeir et al., 2007), continuous covariates like the flat's size and age, as well as some explanatory factors for a flat's quality and equipment are available. As the rent is positively skewed, a structured regression model with Gamma distributed response is assumed. The effect of the age of a flat is known to be non-linear (see, for example, Fahrmeir and Tutz, 2001); it is considered by a spline with a Ridge-type penalty on the effects' curvature. We want to determine whether the effect of the residential area is linear or not. This is reached by the penalty described in Section 3.3; it requires a Lasso and a grouped Lasso penalty. As some levels are only sparsely occupied, ordered factors like the number of rooms of a flat require regularization, too. We want to employ an adaptive fused Lasso type penalty on the dummy coefficients of these covariates. Table 3 gives the exact definitions of the employed predictors. For an observation i, the predictor is

$$\eta_i \;\; = \;\; \beta_0 + \boldsymbol{x}_{i1}^T\boldsymbol{\beta}_1 + \boldsymbol{x}_{i2}^T\boldsymbol{\beta}_4 + f_3(x_{i3}) + f_4(x_{i4}),$$

where transposed vectors $\boldsymbol{x}_i^T$ denote covariates that are related to more than one coefficient. The overall penalty is a sum of Ridge-, grouped Lasso- and Lasso-type penalties; it is denoted

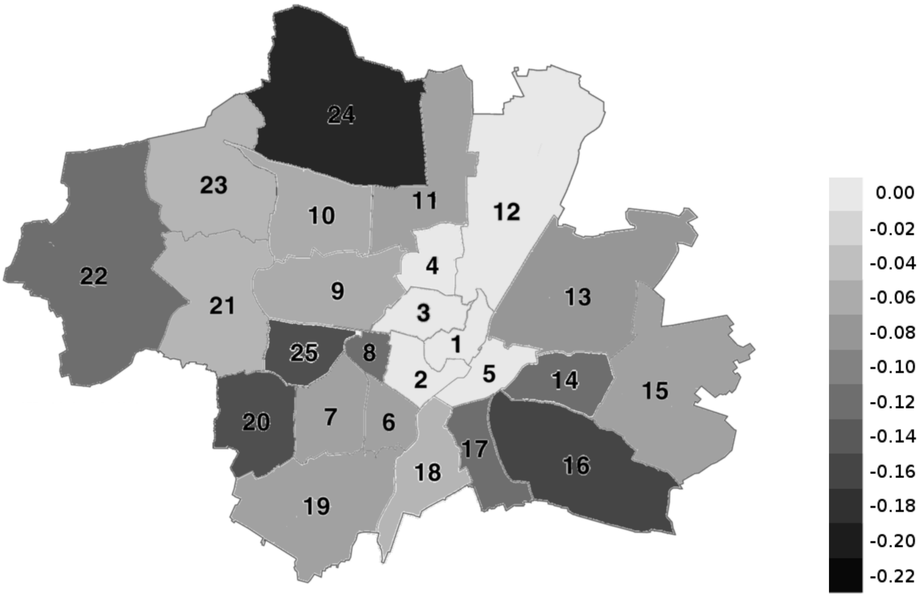

Figure 8: Graphical illustration of the impact of the nominal factor location on the rent. The map depicts the 25 districts of the city of Munich. District 1 denotes the city center, that is the reference category; the remaining districts correspond to one dummy coefficient each. The colors are allocated according to the regularized estimates. As the city center is the most expensive district, all coefficients are negative. The fused Lasso penalty on all pairwise differences of coefficients shrinks the estimates; nine clusters of districts with similar effects are detected. Districts in cluster 1: 18, 21, 23; in cluster 2: 9, 10; in cluster 3: 6, 7, 11, 15, 19; in cluster 4: 8, 14, 17, 22; in cluster 5: 20, 25. The effect of districts 2–5, and 12 is set to zero; the effect of these districts on the rent is the same as in the city center. The graphic is based on a figure of Dörrbecker (2007); it is manipulated with the GNU Image Manipulation Program (GIMP Team, 2012) and with the `R` package `EBImage` (Pau et al., 2012).

by

$$P(\boldsymbol{\beta}) = \lambda \sum_{l=1}^{4} P_l(\boldsymbol{\beta}_l),$$

where $P_1(\boldsymbol{\beta}_1) = \sum_{r=2}^{6} w_{1r}|\beta_{1r} - \beta_{1,r-1}|$ is the fused Lasso penalty for the ordered factor numbrooms with reference $\beta_{11} = 0$. $P_2(\boldsymbol{\beta}_2) = \sum_{r>s} w_{2rs}|\beta_{2r} - \beta_{2s}|$ denotes the penalty for the flats' location. In contrast to P_1 all pairwise differences are considered as the location is a nominal factor.

Weights w_{1r} and w_{2rs} contain both, the weights that account for a different number of levels/of observations on each level (see Section 2.4) and the adaptive weights $|\beta_{1r}^{ML} - \beta_{1,r-1}^{ML}|^{-1}$, $|\beta_{2j}^{ML} - \beta_{2s}^{ML}|^{-1}$ respectively (see Zou, 2006). Adaptive weights come along with quite huge penalty terms, when the according inverse differences are small. In this case, the penalty terms related to other predictors may become negligible. However, even with adaptive weights, the penalty terms of different predictors should be comparable.

To this end, one can abandon the idea of one global tuning parameter and introduce a tuning parameter λ_1 for the comparable but adaptively weighted penalty terms and the parameter λ_2 for non-adaptively weighted penalty terms. $\lambda = (\lambda_1, \lambda_2)$ is determined by cross-validation. However, to avoid multi-dimensional cross-validation, we propose to rescale adaptively weighted penalty terms such that the overall penalty of one predictor is again of order k_j, the number of (free) coefficients related to x_j. For predictor x_1, the number of rooms in a flat, we have for

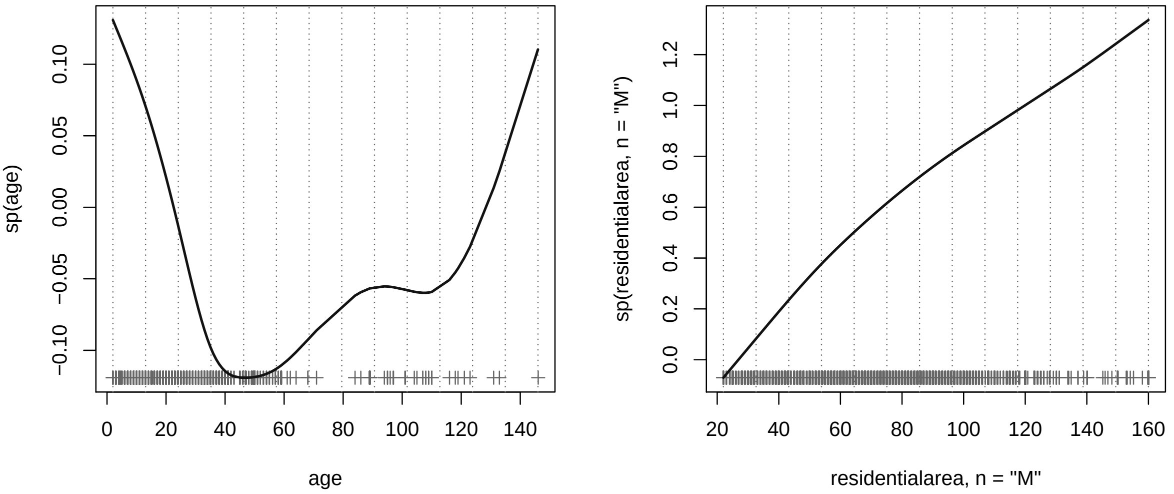

Figure 9: Estimates of functions f_3 and f_4. In the left panel, the impact of the age of a flat is illustrated. The age is measured in years; a flat built in 2007 is aged zero. On the right, the effect of the residential area is shown. The residential area is measured in square meters. The y-axis corresponds to the effect of the age, the residential area, respectively, on the predictors η_i.

example:

$$P_1(\boldsymbol{\beta}) = \frac{1}{\sum_{r=2}^{6} |\beta_{1r}^{ML} - \beta_{1,r-1}^{ML}|^{-1}} \sum_{r=2}^{6} w_{1r}^{level} \frac{|\beta_{1r} - \beta_{1,r-1}|}{|\beta_{1r}^{ML} - \beta_{1,r-1}^{ML}|},$$

where weights $w_{1r}^{level} = \sqrt{\frac{n_{1r}+n_{1,r-1}}{n}}$ adjust for the different number of observations on each level of x_1. There is no need to adjust for the number of penalty terms as they are already of order 5. Functions f_3 and f_4 are represented by decomposed cubic B-spline functions based on 20 equidistant knots, see Section 3.3. The effect of the flats' age f_3 is penalized by $P_3(\boldsymbol{\beta}_3) = \sum_{r=1}^{14} (\beta_{3r}^{nonlin})^2$; that is, by a Ridge penalty on the curvature of the function. Due to the cubic decomposed B-splines with 20 knots, penalty P_3 relates to 14 coefficients and is of order 14. Hence, no additional weighting is needed.

The coefficients related to f_4 are penalized by $P_4(\boldsymbol{\beta}_4) = \alpha|\beta_4^{slope}| + (1-\alpha)w_4\sqrt{\sum_{r=1}^{14}(\boldsymbol{\beta}_{4r}^{nonlin})^2}$ as described in Section 3.3; that is, by a Lasso penalty on the linear effect and by a grouped Lasso penalty on the deviations from this linear effect. Weight w_4 guarantees that the grouped Lasso penalty is of the right order (see Section 2.5). Parameter α is an additional tuning parameter that allows to weight the two components of the penalty. In order to separate it strictly from the global tuning parameter, it is limited to the range $(0, 1)$. As the global tuning parameter λ, it will be chosen by cross-validation.

In the resulting model, the tuning parameters are chosen by 5-fold cross-validation with the predictive deviance as loss criterion and set to $(\lambda, \alpha) = (4.55, 0.3)$. It turns out that all predictors effect upon the response. Figure 8 shows how the districts of Munich are clustered by penalty P_2. The map depicts the 25 districts of the city of Munich. District 1 denotes the city center, that is the reference category; the remaining districts correspond to one dummy coefficient each. The colors are allocated according to the regularized estimates. As the city center is the most expensive district, all coefficients are negative. The fused Lasso penalty on all pairwise differences of coefficients shrinks the estimates; five clusters of districts with similar effects are detected. The effect of the districts 2–5 and 12 is set to zero; they are fused with the city center. The cluster correspond to what one would expect. Flats in these districts are highly requested. Figure 9 illustrates smooth functions f_3 and f_4. The effect of the flats' age is actually non-linear (left panel). It captures the urban development of Munich: After World War II, many flats were constructed; flats build subsequent to the war (1945-1965), have a clearly negative impact on the rent. The more lately the flats are constructed the more expensive they become. Flats that where constructed in the beginning of the 20th century (1900-1930), seem to be of a higher value and outbalance the disadvantages of age. A few very old, extensively redecorated flats give the positive effect for flats build in the 19th century. The right panel of Figure 9 depicts the effect of the residential area. It is nearly linear. There are only small deviations from a linear trend with slope 0.01. The dummy coefficients for two and three rooms are fused with the reference category "one room". Four and more rooms rooms have a negative impact on the response; the categories for five and six rooms are fused: $\beta_4 = -0.01$, $\beta_{15} = -0.10$, $\beta_{16} = -0.10$).

Overall, the model seems to give a realistic picture of how the rents are arranged. Especially the effect of the flats' age has a close match in history. Of course, one could argue for many other models. One could spend more time on additional factors. One could think of different penalties, too. For example, the location is so far considered as a nominal factor; all pairwise differences of dummy coefficients are penalized. Instead, the penalty could take account of the spatial structure. One could consider only differences of neighbored districts or weight the differences by the length of their joint boundary.

5 Concluding Remarks

We propose a general approach to combine different types of penalties in one generalized structured regression model. It allows for example, for penalized smooth functions, Lasso-regularized predictors and categorical predictors penalized by a grouped Lasso in one predictor. The response can follow any exponential family. This is challenging because the objective function combines various potentially non-differentiable terms like norms, quadratic terms, absolute values or indicators. To solve this problem, we employ a local quadratic approximation for the penalty that is based on ideas of Fan and Li (2001) and Ulbricht (2010). The approximation is iteratively updated in a PIRLS algorithm. That gives an algorithm of similar complexity as for usual GLMs; however, coefficient paths and cross-validation scores have to be computed separately. In comparison to other (straight forward) optimization methods like Nelder-Mead or Newton-type algorithms, the PIRLS framework turns out to be more stable. We provide an implementation of the algorithm in `R` package `gvcm.cat` (Oelker, 2013).

We choose the tuning parameters by cross-validation and propose a weighting scheme to adjust for differently weighted or scaled predictors. Alternatively, tuning parameters could be estimated in a mixed model framework. Confidence regions can be constructed by bootstrap methods.

As shown in the paper, the algorithm can be easily extended to vector valued penalties like the grouped Lasso. Moreover, the approach allows to implement many penalties that have not been explored so far; only boosting and Bayesian approaches provided such a variety of penalties as proposed in this paper.

Acknowledgements

This work was partially supported by DFG project "Regularisierung für diskrete Datenstrukturen".

A Appendix: Illustration of R package gvcm.cat

R code to reproduce some of the results in Section 3.

B Appendix: Districts in the City Munich

	Number	District
Downtown	1	Altstadt, Lehel
	2	Ludwigvorstadt, Isarvorstadt
	3	Maxvorstadt
	4	Schwabing West
	5	Au, Haidhausen
	6	Sendling
	7	Sendling, Westpark
	8	Schwanthaler Höhe
	9	Neuhausen, Nymphenburg
	25	Laim
East side	13	Bogenhausen
	14	Berg am Laim
	15	Trudering, Riem
	16	Ramersdorf, Perlach
	17	Obergiesing
	18	Untergiesing, Harlaching
	19	Thalkirchen, Obersendling, Forstenried, Fürstenried, Solln
West side	10	Moosach
	11	Milbertshofen, Am Hart
	12	Schwabing - Freimann
	20	Hadern
	21	Pasing, Obermenzing
	22	Aubing, Lochhausen, Langwied
	23	Allach, Untermenzing
	24	Feldmoching, Hasenbergl

Table 4: Overview on the districts in the City of Munich (Fahrmeir et al., 2007). The numbering corresponds to the labels in Figure 8 and to the naming of the according dummy coefficients.

References

Antoniadis, A. and J. Fan (2001). Regularization of wavelet approximations. *J. Amer. Statist. Assoc. 96*(455), 939–967.

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

Claeskens, G. and N. L. Hjort (2008). Minimizing average risk in regression models. *Econometric Theory 24*(2), 493–527.

Cox, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B 34*(1), 187–220.

Donoho, D. and M. Elad (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization. *Proceedings of the National Academy of Sciences 100*(5), 2197–2202.

Dörrbecker, M. (2007). *Boroughs of Munich.* freely available under the creative commons cc-by-sa 2.5 license (http://creativecommons.org/licenses/by-sa/2.5/).

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

Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with b-splines and penalties. *Statist. Sci. 11*(2), 89–121.

Fahrmeir, L., C. Belitz, C. Biller, A. Brezger, S. Heim, A. Hennerfeind, and A. Jerak (2007). Statistische Analyse der Nettomieten. In *Mietspiegel für München 2007. Statistik, Dokumentation und Analysen. Landeshauptstadt München, Sozialreferat, Amt für Wohnen und Migration.*

Fahrmeir, L., T. Kneib, and S. Konrath (2010). Bayesian regularisation in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection. *Stat. Comput. 20*(2), 203–219.

Fahrmeir, L., T. Kneib, and S. Lang (2004). Penalized structured additive regression for space-time data: A bayesian perspective. *Statist. Sinica 14*(3), 715–745.

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

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc. 96*(456), 1348–1360.

Frank, l. E. and J. H. Friedman (1993). A statistical view of some chemometrics regression tools. *Technometrics 35*(2), 109–135.

Friedman, J. H., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw. 33*(1), 1–22. R package version 1.9-5.

Gertheiss, J., S. Hogger, C. Oberhauser, and G. Tutz (2011). Selection of ordinally scaled independent variables with applications to international classification of functioning core sets. *R. Stat. Soc. Ser. C Appl. Stat. 60*(3), 377–395.

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

Gertheiss, J. and G. Tutz (2012). Regularization and model selection with categorial effect modifiers. *Statist. Sinica 22*(3), 957–982.

GIMP Team (2012). *GNU Image Manipulation Program.* http://www.gimp.org.

Goeman, J. J. (2010). L1 penalized estimation in the cox proportional hazards model. *Biom. J. 52*(1), 70–84.

Hastie, T. and B. Efron (2013). *lars: Least Angle Regression, Lasso and Forward Stagewise.* R package version 1.2.

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

Koch, I. (1996). On the asymptotic performance of median smoothers in image analysis and nonparametric regression. *Ann. Statist. 24*(4), 1648–1666.

Marx, B. D. and P. H. C. Eilers (1998). Direct generalized additive modeling with penalized likelihood. *Journal of Computational Statistics & Data Analysis 28*, 193–209.

McCullagh, P. and J. A. Nelder (1983). *Generalized Linear Models.* Chapman & Hall, London.

Meier, L. (2013). *grplasso: Fitting user specified models with Group Lasso penalty.* R package version 0.4-3.

Meier, L., S. van de Geer, and P. Bühlmann (2008). The group Lasso for logistic regression. *R. Stat. Soc. Ser. B Stat. Methodol. 70*(1), 53–71.

Meier, L., S. van de Geer, and P. Bühlmann (2009). High-dimensional additive modeling. *Ann. Statist. 37*(6B), 3779–3821.

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

Osborne, M. R. and B. A. Turlach (2011). A homotopy algorithm for the quantile regression lasso and related piecewise linear problems. *Journal of Computational and Graphical Statistics 20*(4), 972–987.

O'Sullivan, F., B. S. Yandell, and W. J. Raynor (1986). Automatic smoothing of regression functions in generalized linear models. *J. Amer. Statist. Assoc. 81*(393), 96–103.

Pau, G., A. Oles, M. Smith, O. Sklyar, and W. Huber (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, R. C. A., J. J. Meulman, and P. H. C. Eilers (2012). Visualization of genomic changes by segmented smoothing using an l_0 penalty. *PloS One 7*(6), 1–14.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *R. Stat. Soc. Ser. B Stat. Methodol. 58*(1), 267–288.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005). Sparsity and smoothness via the fused LASSO. *R. Stat. Soc. Ser. B Stat. Methodol. 67*(1), 91–108.

Tutz, G. (2011). *Regression for Categorical Data.* Cambridge University Press, New York.

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

Wang, H. and C. Leng (2008). A note on adaptive group lasso. *Journal of Computational Statistics & Data Analysis 52*, 5277–5286.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol. 73*(1), 3–36. R package version 1.7-26.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *R. Stat. Soc. Ser. B Stat. Methodol. 68*(1), 49–67.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *J. Amer. Statist. Assoc. 101*(476), 1418–1429.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the Elastic Net. *R. Stat. Soc. Ser. B Stat. Methodol. 67*(2), 301–320.