LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Martin Slawski, Wolfgang zu Castell & Gerhard Tutz

Feature Selection Guided by Structural Information

Technical Report Number 051, 2009
Department of Statistics
University of Munich

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

Feature Selection Guided by Structural Information

Martin Slawski[1,2,3*] Wolfgang zu Castell[4†] Gerhard Tutz[1‡]

[1] Department of Statistics, Ludwig-Maximilians-University Munich, Germany
[2] Sylvia Lawry Centre, Munich, Germany
[3] Department of Computer Science, Saarland University, Saarbrücken, Germany
[4] Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany

Abstract

In generalized linear regression problems with an abundant number of features, lasso-type regularization which imposes an ℓ^1-constraint on the regression coefficients has become a widely established technique. Crucial deficiencies of the lasso were unmasked when Zhou and Hastie (2005) introduced the elastic net. In this paper, we propose to extend the elastic net by admitting general nonnegative quadratic constraints as second form of regularization. The generalized ridge-type constraint will typically make use of the known association structure of features, e.g. by using temporal- or spatial closeness.

We study properties of the resulting 'structured elastic net' regression estimation procedure, including basic asymptotics and the issue of model selection consistency. In this vein, we provide an analog to the so-called 'irrepresentable condition' which holds for the lasso. An oracle property is established by incorporating a scaled ℓ^1-constraint. Moreover, we outline algorithmic solutions for the structured elastic net within the generalized linear model family. The rationale and the performance of our approach is illustrated by means of simulated and real world data.

keywords: generalized linear model, regularization, sparsity, $p \gg n$, lasso, elastic net, random fields, consistency, epiconvergence, model selection, signal regression.

1 Introduction

We consider regression problems with a linear predictor. Let $\mathbb{X} = (X_1, \ldots, X_p)^\top$ be a random vector of real-valued features/predictors and let Y be a random response variable taking values in a set $\mathcal{Y} \subseteq \mathbb{R}$. Given a realization $\boldsymbol{x} = (x_1, \ldots, x_p)^\top$ of $\mathbb{X}$, a prediction $\widehat{y}$ for the response is obtained via a linear predictor

$$f(\boldsymbol{x}; \beta_0, \boldsymbol{\beta}) = \beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta}, \quad \boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top,$$

and a function $\zeta : \mathbb{R} \to \mathcal{Y}$ such that $\widehat{y} = \zeta(f(\boldsymbol{x}))$. Given an i.i.d. sample $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ from $(\mathbb{R}^p \times \mathcal{Y})^n$, an optimal set of coefficients $\widehat{\beta}_0$, $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)^\top$ can be determined by minimization of a criterion of the form

$$(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) = \operatorname*{argmin}_{(\beta_0, \boldsymbol{\beta})} \sum_{i=1}^n L(y_i, f(\boldsymbol{x}_i; \beta_0, \boldsymbol{\beta})), \tag{1.1}$$

*ms@cs.uni-sb.de
†castell@helmholtz-muenchen.de
‡tutz@stat.uni-muenchen.de

where $L : \mathcal{Y} \times \mathbb{R} \rightarrow \mathbb{R}_0^+$ is a loss function, assumed to be continuous and convex in the second argument. The loss function is chosen according to the specific prediction problem, so that large loss represents bad fit to the observed sample S. Approach (1.1) usually yields poor estimates $\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}$ if n is not one order of magnitude larger than p. In particular, if $p \gg n$, approach (1.1) is not well-defined in the sense that there exist infinitely many minimizers $\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}$. One way to cope with a small n/p ratio is to employ a regularizer $\Omega(\boldsymbol{\beta})$. A traditional approach due to Hoerl and Kennard (1970) minimizes the loss in Eq. (1.1) subject to an ℓ^2-constraint on $\boldsymbol{\beta}$. In the situation that $\boldsymbol{\beta}$ is supposed to be sparse, Tibshirani (1996) proposed, under the acronym 'lasso', to work with an ℓ^1-constraint, i.e. one maximizes the loss subject to $\Omega(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1 < s, \quad s > 0$. The latter is particularly attractive if one is interested in feature selection, since one obtains estimates $\widehat{\beta}_j, \quad j \in \{1, \ldots, p\}$, which equal exactly zero, such that feature j does not contribute to prediction, for which we say that feature j is 'not selected'. Continuous shrinkage (Fan and Li (2001)) and the existence of efficient algorithms (Efron et al. (2004), Genkin et al. (2007)) for determining the coefficients are further virtues of the lasso. Its limitations have recently been revealed by several researchers. Zhou and Hastie (2005) pointed out that the lasso degenerates in the $p \gg n$ setting, where the lasso is able to select at most n features (Rosset et al. (2004)). Furthermore, Zhou and Hastie stated that the lasso does not distinguish between 'irrelevant' and 'relevant but redundant' features. In particular, if there is a group of correlated features, then the lasso tends to select one arbitrary member of the group while ignoring the remainder. The combined regularizer of the elastic net $\Omega(\boldsymbol{\beta}) = \alpha \|\boldsymbol{\beta}\|_1 + (1-\alpha) \|\boldsymbol{\beta}\|^2, \quad \alpha \in (0,1)$ is shown to provide remedy in this regard.
A second double regularizer - tailored to one-dimensional signal regression - is employed by the fused lasso (Tibshirani et al. (2005)), who propagate $\Omega(\boldsymbol{\beta}) = \alpha \|\boldsymbol{\beta}\|_1 + (1-\alpha) \|\boldsymbol{D}\boldsymbol{\beta}\|_1$, where

$$\begin{array}{rccc} \boldsymbol{D}: & \mathbb{R}^p & \rightarrow & \mathbb{R}^{p-1} \\ & (\beta_1, \ldots, \beta_p)^\top & \mapsto & ([\beta_2 - \beta_1], \ldots, [\beta_p - \beta_{p-1}])^\top \end{array} \tag{1.2}$$

is the first forward difference operator. The total variation regularizer is meaningful whenever there is an order relation, notably a temporal one, among the features. The fused lasso has a property which can be beneficial for interpretation: it automatically clusters the features, since the sequence $\widehat{\beta}_1, \ldots, \widehat{\beta}_p$ is blockwise constant.
In this paper, we study a regularizer which is intermediate between the elastic net and the fused lasso. Our regularizer combines an ℓ^1-constraint with a quadratic form:

$$\Omega(\boldsymbol{\beta}) = \alpha \|\boldsymbol{\beta}\|_1 + (1-\alpha)\boldsymbol{\beta}^\top \boldsymbol{\Lambda} \boldsymbol{\beta}, \tag{1.3}$$

where $\boldsymbol{\Lambda} = (l_{jj'})_{1 \leq j,j' \leq p}$ is assumed to be symmetric and positive semidefinite. Setting $\boldsymbol{\Lambda} = \boldsymbol{I}$ yields the elastic net. Therefore, expression (1.3) will be referred to as structured elastic net regularizer. The inclusion of $\boldsymbol{\Lambda}$ aims at capturing the a priori association structure (if available) of the features in more generality than the fused lasso. The structured elastic net estimator is defined as

$$\begin{gathered} (\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) = \underset{(\beta_0, \boldsymbol{\beta})}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, f(\boldsymbol{x}_i; \beta_0, \boldsymbol{\beta})) \\ \text{subject to } \alpha \|\boldsymbol{\beta}\|_1 + (1-\alpha)\boldsymbol{\beta}^\top \boldsymbol{\Lambda} \boldsymbol{\beta} \leq s, \quad \alpha \in (0;1), \quad s > 0, \end{gathered} \tag{1.4}$$

which is equivalent to the Lagrangian formulation

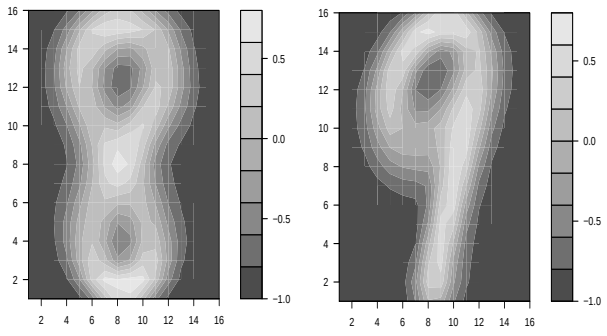
$$(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) = \underset{(\beta_0, \boldsymbol{\beta})}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, f(\boldsymbol{x}_i; \beta_0, \boldsymbol{\beta})) + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \boldsymbol{\beta}^\top \boldsymbol{\Lambda} \boldsymbol{\beta}, \quad \lambda_1, \lambda_2 > 0. \tag{1.5}$$

The rest of the paper is organized as follows: in Section 2, we discuss the choice of the matrix $\mathbf{\Lambda}$, followed by an analysis of some important properties of our proposal (1.5) in Section 3. Section 4 is devoted to asymptotics and consistency questions, motivating the introduction of the *adaptive* structured elastic net. Section 5 presents three different algorithms for computing the minimizers (1.5) in the generalized linear model family and addresses fundamental questions of model selection and -inference. The practical performance of the structured elastic net is contained in Section 6. Section 7 concludes with a brief discussion and outlook. All proofs can be found in the Appendix.

2 Structured features

2.1 Motivation

A considerable fraction of contemporary regression problems is characterized by a large number of features, which is either of the same order of magnitude as the sample size or even several orders larger ($p \gg n$). Common instances thereof are feature sets consisting of sampled signals, pixels of an image, spatially sampled data, or gene expression intensities. Beside high dimensionality of the feature space, these examples have in common that the feature set can be arranged according to an a priori association structure. If a sampled signal does not vary rapidly, the influence of nearby sampling points on the response can be expected to be similar; correspondingly, this applies to adjacent pixels of an image, or, more general, to any other form of spatially linked features. In genomics, genes can be categorized into functional groups, or one has prior knowledge of their functions and interactions within biochemical reaction chains, so called pathways.
Figures 1 and 2 display two well-known examples, phoneme- and handwritten digit classification. These examples are well-apt to illustrate the idea of the structured elastic net regularizer, since it is sensible to assume that the prediction problem is not only characterized by smoothness with respect to a given structure, but also by sparsity: in the phoneme classification example, visually only the first hundred frequencies seem to carry information relevant to the prediction problem. A similar rationale applies to the second example, where the arc of the numeral eight in the lower half of the picture is the eminent characteristic that admits a distinction from the numeral nine.



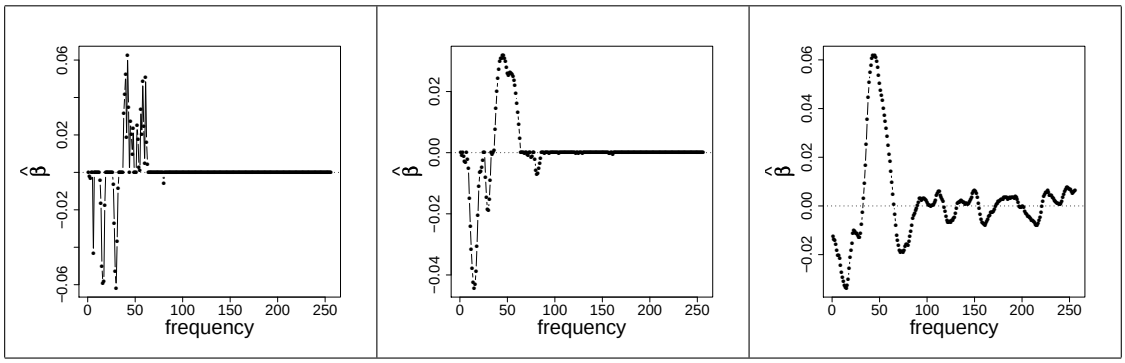


Figure 1: Phoneme data (Hastie et al. (1995)). The upper panel shows several thousand log-periodogramms of the speech frames for the phonemes 'aa' (as occuring in 'dark') and 'ao' (as occuring in 'water'). The classwise means are given by thick lines. We use linear logistic regression to predict the phoneme given a log-periodogramm. The lower panel depicts the resulting coefficients when using the lasso (left panel), a first-order difference penalty (right panel), and a combination thereof, which we term 'structured elastic net' (middle panel).

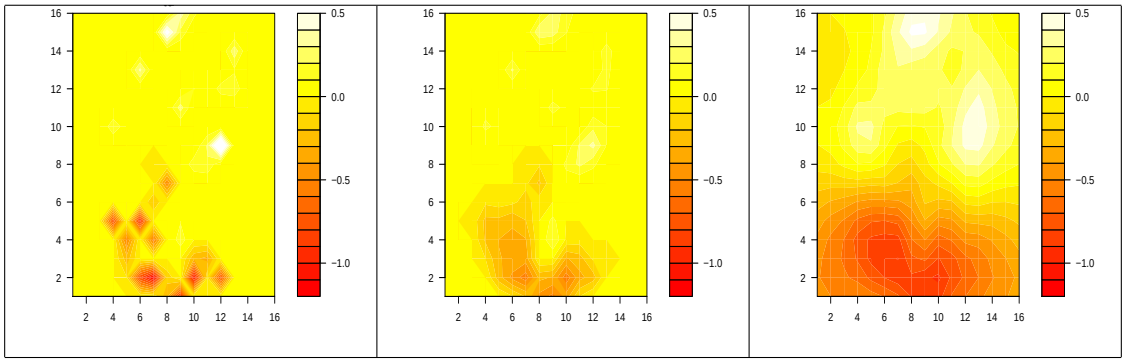


Figure 2: Handwritten digit recognition dataset (Le Cun et al. (1989)). One observation is given by a greyscale image composed of 16 × 16 pixels. The upper panel shows the contour of the pixel-wise means for the numerals '8' and '9'. We use a training set of 1500 observations of eights and nines as input for linear logistic regression. The lower panel depicts the coefficient surfaces for the lasso (left panel), a discrete Laplacian penalty according to the grid structure (right panel), and a combination, the structured elastic net (middle panel).

2.2 Gauss-Markov random fields

Given a large, but structured set of features, its structure can be exploited to cope with high dimensionality in regression estimation. The estimands $\{\beta_j\}_{j=1}^p$ form a finite set such that their prior dependence structure can conveniently be described by means of a graph $\mathcal{G} = (V, E)$, $V = \{\beta_1, \ldots, \beta_p\}$, $E \subset V \times V$. We exclude loops, i.e. $(\beta_j, \beta_j) \notin E$ for all j. The edges may additionally be weighted by a function $w : E \to \mathbb{R}$, $w((\beta_j, \beta_{j'})) = w((\beta_{j'}, \beta_j))$ for all edges in E. We will use the notation $\beta_j \sim \beta_{j'}$ to express that β_j and $\beta_{j'}$ are connected by an edge in $\mathcal{G}$. The weight function can be extended to a function on $V \times V$ by setting $w((\beta_j, \beta_{j'})) = w((\beta_{j'}, \beta_j)) = 0$ if $(\beta_j, \beta_{j'}) \notin E$.

The graph is interpreted in terms of Gauss-Markov random fields (Besag (1974); Rue and

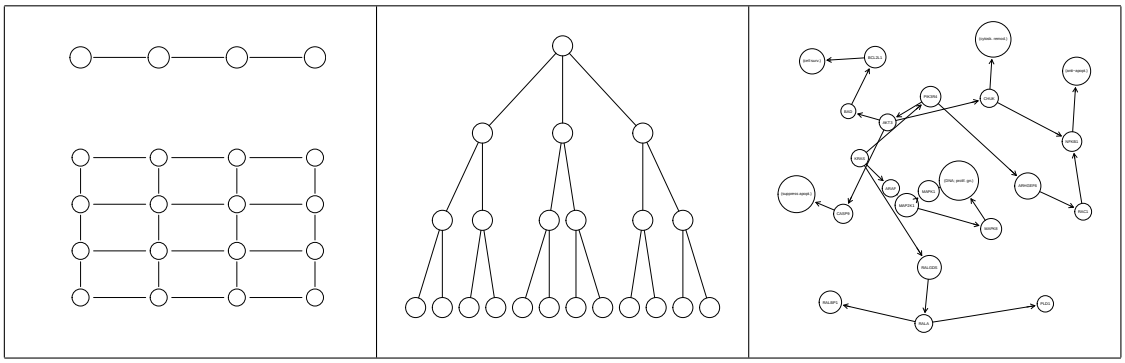


Figure 3: A collection of some graphs. A path and a grid (left panel), a rooted tree (middle panel) and an irregular graph describing a part of the so-called MAPK signaling pathway (right panel).

Held (2001)). In our setup, the pairwise Markov property reads

$$\neg\, \beta_j \sim \beta_{j'} \;\Leftrightarrow\; \beta_j \perp\!\!\!\perp \beta_{j'} |\, V \setminus \{\beta_j, \beta_{j'}\}, \tag{2.1}$$

with $\perp\!\!\!\perp$ denoting conditional independence. Property (2.1) is conform to the following choice for the precision matrix $\boldsymbol{\Lambda} = (l_{jj'})_{1 \le j,j' \le p}$:

$$l_{jj'} = \begin{cases} \sum_{k=1}^{p} |w((\beta_j, \beta_k))| & \text{if } j = j', \\ -w((\beta_j, \beta_{j'})) & \text{if } j \neq j', \end{cases} \tag{2.2}$$

which is singular in general. If $\text{sign}\{w((\beta_j, \beta_{j'}))\} \geq 0$ for all $(\beta_j, \beta_{j'})$ in E, then $\boldsymbol{\Lambda}$ as given in Eq. (2.2) is known as the combinatorial graph Laplacian in spectral graph theory (Chung (1997)). It is straightforward to verify the following properties.

- $$\boldsymbol{\beta}^\top \boldsymbol{\Lambda} \boldsymbol{\beta} = \sum_{\beta_j \sim \beta_{j'}} |w(\beta_j, \beta_{j'})| \, (\beta_j - \text{sign}\left\{w((\beta_j, \beta_{j'}))\right\} \, \beta_{j'})^2 \geq 0, \tag{2.3}$$

  where the sum is over all distinct edges in $\mathcal{G}$, and 'distinct' is understood with respect to the relation $(\beta_j, \beta_{j'}) = (\beta_{j'}, \beta_j)$ for all j, j'.

- If $\mathcal{G}$ is connected and $\text{sign}\left\{w((\beta_j, \beta_{j'}))\right\} \geq 0$ for all $(\beta_j, \beta_{j'})$ in E, the nullspace of $\boldsymbol{\Lambda}$ is spanned by the vector of ones $\mathbf{1}$.

While we have started in full generality, the choice $w((\beta_j, \beta_{j'})) \in \{0, 1\}$ for all j, j' will frequently be the standard choice in practice. In this case, the quadratic form captures local fluctuations of $\boldsymbol{\beta}$ w.r.t. $\mathcal{G}$. As a simple example, one may take $\mathcal{G}$ as the path on p vertices so that expression (2.3) equals the summed squared forward differences

$$\sum_{j=2}^{p} (\beta_j - \beta_{j-1})^2 = \|\boldsymbol{D}\boldsymbol{\beta}\|^2 = \boldsymbol{\beta}^\top \boldsymbol{D}^\top \boldsymbol{D} \boldsymbol{\beta}, \tag{2.4}$$

where $\boldsymbol{D}$ is defined in Eq. (1.2). More complex graphical structures can be generated from simple ones using the notion of Cartesian products of graphs. Given two unweighted

graphs $\mathcal{G} = (V, E)$ and $\mathcal{G}' = (V', E')$, their Cartesian product is defined as

$$\begin{aligned} \mathcal{G} \times \mathcal{G}' &= (V_\times, E_\times), \quad V_\times = V \times V', \\ E_\times &= \{((j,k),(j',k')) \in V_\times \times V_\times : \ (j,j') \in E \ \wedge \ k = k' \ \vee \ (k,k') \in E' \ \wedge j = j'\}. \end{aligned} \tag{2.5}$$

In terms of adjacency matrices $\boldsymbol{A} = (a_{jj'} = I(j \overset{\mathcal{G}}{\sim} j'))$ and $\boldsymbol{A}' = (a_{kk'} = I(k \overset{\mathcal{G}'}{\sim} k'))$, where I is the indicator function, definition (2.5) can more comfortably be expressed as

$$\boldsymbol{A}_\times = \boldsymbol{A} \otimes \boldsymbol{I}' + \boldsymbol{I} \otimes \boldsymbol{A}', \tag{2.6}$$

where $\boldsymbol{A}_\times$ denotes the adjacency matrix of $\mathcal{G} \times \mathcal{G}'$, $\boldsymbol{I}$ and $\boldsymbol{I}'$ are identity matrices with dimensions equal to the number of vertices in $\mathcal{G}$ and $\mathcal{G}'$, respectively, and $\otimes$ denotes the Kronecker product. It is easy to see that this construction yields a regular grid on $p \times p'$ vertices if $\mathcal{G}$ is chosen as p-path and $\mathcal{G}'$ as p'-path, in which case $\boldsymbol{\Lambda}$ equals the usual discretization of the Laplacian Δ acting on functions defined on $\mathbb{R}^2$. Regularizers built up from discrete differences have already seen frequent use in high-dimensional regression estimation. Examples comprise penalized discriminant analysis (Hastie et al. (1995)) and spline smoothing (Eilers and Marx (1996)).

3 Properties

3.1 Bayesian and geometric interpretation

In the setup of Section 1, consider the regularizer

$$\Omega(\boldsymbol{\beta}) = \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \boldsymbol{\beta}^\top \boldsymbol{\Lambda} \boldsymbol{\beta}, \quad \lambda_1, \lambda_2 > 0.$$

It has a nice Bayesian interpretation when the loss function L is of the form

$$L(y, f(\boldsymbol{x}; \beta_0, \boldsymbol{\beta})) = \phi^{-1}(b(f(\boldsymbol{x})) - y f(\boldsymbol{x})) + c(y, \phi), \tag{3.1}$$

i.e. the loss function equals the negative log-likelihood of a generalized linear model in canonical parametrization, which will primarily be studied in this paper. Models of this class are characterized by

$$\begin{aligned} Y|\mathbb{X} = \boldsymbol{x} \ &\sim \ \text{simple exponential family}, \\ \widehat{y} = \mathrm{E}[Y|\mathbb{X} = \boldsymbol{x}] = \mu &= \frac{d}{df} b(f(\boldsymbol{x})), \\ \mathrm{var}[Y|\mathbb{X} = \boldsymbol{x}] &= \phi \frac{d^2}{df^2} b(f(\boldsymbol{x})). \end{aligned} \tag{3.2}$$

The form (3.1) is versatile, including classical linear regression with Gaussian errors, logistic regression for classification, and Poisson regression for count data. Given a loss from the class (3.1), the regularizer $\Omega(\boldsymbol{\beta})$ can be interpreted as combined Laplace (double exponential)-Gaussian prior $p(\boldsymbol{\beta}) \propto \exp(-\Omega(\boldsymbol{\beta}))$, for which the structured elastic net estimator (1.5), provided $p(\beta_0) \propto 1$, is the maximum posterior (MAP) estimator given the sample S. Note that if $p > n$, depending on the matrix $\boldsymbol{\Lambda}$, $p(\boldsymbol{\beta})$ is not necessarily proper. It is instructive to consider two predictors, i.e. $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top$. Figure 4 gives a geometric interpretation for the basic choices $\boldsymbol{\Lambda} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$ and $\boldsymbol{\Lambda} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, corresponding

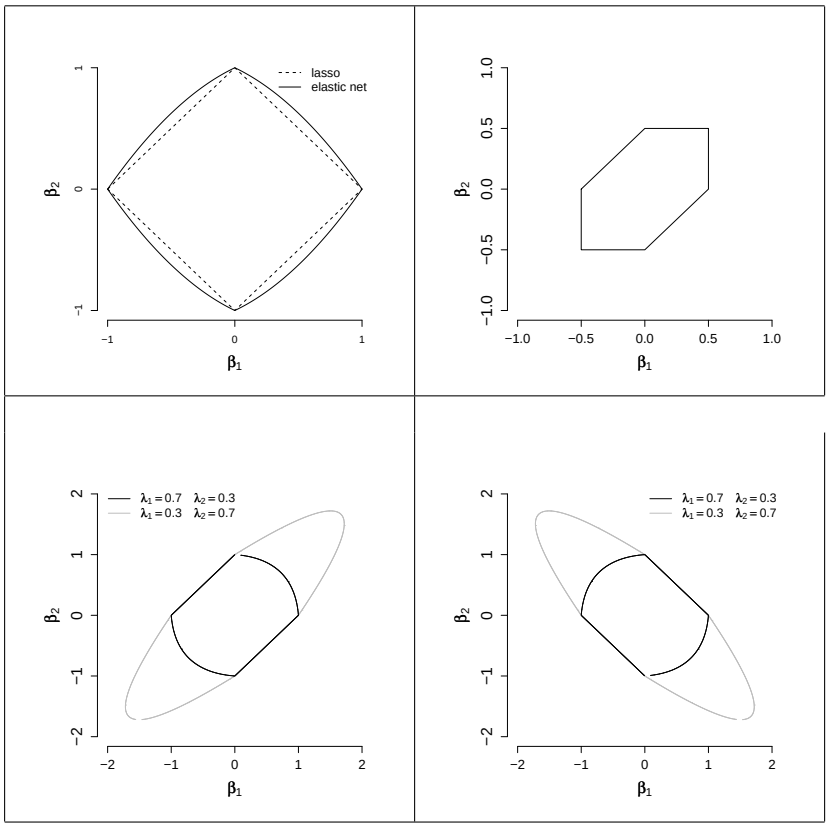


Figure 4: Geometry of the lasso and the elastic net (upper left panel), the fused lasso (upper right panel) and the structured elastic net (positive prior correlation: left panel, negative prior correlation: right panel).

to positive- and negative prior correlation, respectively.
The contour lines of the structured elastic net penalty contain elements of a diamond and an ellipsoid. The higher λ_2 in relation to λ_1, the ellipsoidal part becomes more narrower and more stretched. The sign of the off-diagonal element of $\mathbf{\Lambda}$ determines the orientation of the ellipsoidal part.

3.2 Grouping properties

For the elastic net, Zhou and Hastie (2005) provided an upper bound on the absolute distances $|\widehat{\beta}_j^{\text{elastic net}} - \widehat{\beta}_{j'}^{\text{elastic net}}|$, $j, j' = 1, \ldots, p$, in terms of the sample correlations, to which Zhou and Hastie referred to as 'grouping property'. We provide similar bounds here. For what follows, let S be a sample as in Section 1. We introduce a design matrix $\boldsymbol{X} = (x_{ij})_{\substack{1 \le i \le n \\ 1 \le j \le p}}$ and denote by $\boldsymbol{X}_j = (x_{1j}, \ldots, x_{nj})^\top$ the realizations of predictor j in S, and the response vector is defined by $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$. For the remainder of this section, we assume that the responses are centered and that the predictors are centered and standardized to unit Euclidean length w.r.t the sample S, i.e.

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} x_{ij} = 0, \quad \sum_{i=1}^{n} x_{ij}^2 = 1, \quad j = 1, \ldots, p. \tag{3.3}$$

Proposition 1. *Let $p = 2$, let the loss function be of the form (3.1), let $\rho = \langle \boldsymbol{X}_1, \boldsymbol{X}_2 \rangle$*

denote the sample correlation of $\boldsymbol{X}_1$ *and* $\boldsymbol{X}_2$*, and let* $\boldsymbol{\Lambda} = \frac{1}{2}\begin{pmatrix} 1 & s \\ s & 1 \end{pmatrix}$, $s \in \{-1, 1\}$. *If* $-s\widehat{\beta}_1\widehat{\beta}_2 > 0$*, then*

$$|\widehat{\beta}_1 + s\widehat{\beta}_1| \leq \frac{1}{2\lambda_2}\sqrt{2(1+s\rho)}\,\|\boldsymbol{y}\|\,.$$

In particular, Proposition 1 implies that if $\boldsymbol{X}_1 = -s\boldsymbol{X}_2$, $\widehat{\beta}_1 = -s\widehat{\beta}_2$.
In a similar way as done in Wang et al. (2006), we can obtain a bound for loss functions which are either uniformly Lipschitz as function of the margin (classification) or as function of the residual (regression).

Proposition 2. *In the setup of Proposition 1 and for a positive finite constant* C*, assume that one of the following conditions holds:*

(i) $L(y, f(\boldsymbol{x})) = L(m)$, $m = yf(\boldsymbol{x})$, $y \in \{-1, 1\}$, $|L(m) - L(m')| \leq C|m - m'|$ *for all* m, m'.

(ii) $L(y, f(\boldsymbol{x})) = L(r)$, $r = y - f(\boldsymbol{x})$, $|L(r) - L(r')| \leq C|r - r'|$ *for all* r, r'.

Then:

$$|\widehat{\beta}_1 + s\widehat{\beta}_2| \leq \frac{C}{\lambda_2}\sqrt{n2(1+s\rho)}.$$

Proposition 2 provides a grouping property for important loss functions, covering the hinge loss of support vector classification as well as least absolute deviation- and quantile regression.

3.3 Decorrelation

Let us now consider the important special case

$$L(y, f(\boldsymbol{x}; \boldsymbol{\beta})) = (y - \boldsymbol{x}^\top\boldsymbol{\beta})^2,$$

which corresponds to classical linear regression. The constant term β_0 is omitted, since we work with centered data. The structured elastic net estimator can then be written as

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} -2\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top[\boldsymbol{C} + \lambda_2\boldsymbol{\Lambda}]\boldsymbol{\beta} + \lambda_1\|\boldsymbol{\beta}\|_1\,, \quad \boldsymbol{C} = \boldsymbol{X}^\top\boldsymbol{X}, \\ &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} -2\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\widetilde{\boldsymbol{C}}\boldsymbol{\beta} + \lambda_1\|\boldsymbol{\beta}\|_1\,, \quad \widetilde{\boldsymbol{C}} = \boldsymbol{X}^\top\boldsymbol{X} + \lambda_2\boldsymbol{\Lambda}. \end{aligned} \tag{3.4}$$

Note that for standardized predictors, $\boldsymbol{C}$ equals the matrix of sample correlations $\rho_{jj'} = \langle \boldsymbol{X}_j, \boldsymbol{X}_{j'} \rangle$, $j, j' = 1, \ldots, p$. With a large number of predictors or elements $\rho_{jj'}$ with large $|\rho_{jj'}|$, $\boldsymbol{C}$ is known to yield severely unstable ordinary least squares (ols) estimates $\widehat{\beta}_j^{\text{ols}}$, $j = 1, \ldots, p$. If the two underlying random variables X_j and $X_{j'}$ are highly positively correlated, this will likely translate to high sample correlations of $\boldsymbol{X}_j$ and $\boldsymbol{X}_{j'}$, which in turn yields a strongly negative correlation between $\widehat{\beta}_j^{\text{ols}}$ and $\widehat{\beta}_{j'}^{\text{ols}}$ and as consequence high variances $\text{var}[\widehat{\beta}_j^{\text{ols}}]$ and $\text{var}[\widehat{\beta}_{j'}^{\text{ols}}]$. The modified matrix $\widetilde{\boldsymbol{C}}$ can be written as $\widetilde{\boldsymbol{C}} = \boldsymbol{V}_{\boldsymbol{\Lambda}}^{1/2}\boldsymbol{R}_{\boldsymbol{\Lambda}}\boldsymbol{V}_{\boldsymbol{\Lambda}}^{1/2}$, $\boldsymbol{V}_{\boldsymbol{\Lambda}} = \text{diag}(1 + \lambda_2\sum_{k=1}^p |l_{1k}|, \ldots, 1 + \lambda_2\sum_{k=1}^p |l_{pk}|)$, and the modified correlation matrix $\boldsymbol{R}_{\boldsymbol{\Lambda}}$ has entries

$$\rho_{\boldsymbol{\Lambda},jj'} = \frac{\rho_{jj'} + \lambda_2 l_{jj'}}{\sqrt{1 + \sum_{k=1}^p |l_{jk}|}\sqrt{1 + \sum_{k=1}^p |l_{j'k}|}}, \quad j, j' = 1, \ldots, p.$$

In the light of Section 2, the entries of $\boldsymbol{R}_{\boldsymbol{\Lambda}}$ combine sample- and prior correlations. Decorrelation occurs if $\rho_{jj'} \approx -\lambda_2 l_{jj'}$, i.e. if prior- and sample correlations are in accordance. The grouping- and decorrelation effect of our proposal are visualized in Figure 5 for two predictors, in which case the minimizer (3.4) can be determined analytically (see Appendix B). The figure unmasks the weakness of ordinary least squares in the case of high sample correlations, as well as the tendency of the lasso not to jointly include the two predictors.

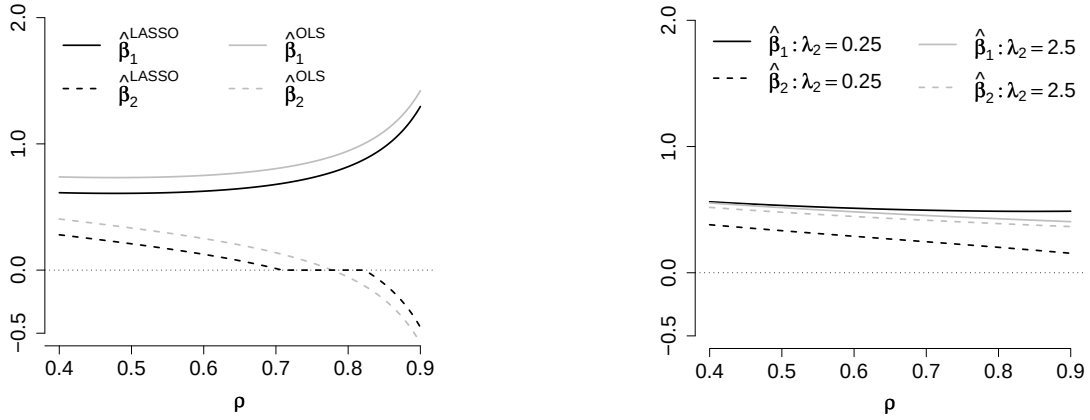


Figure 5: The case of two predictors for ordinary least squares, lasso and the structured elastic net. For both panels, the predictor-response correlations are set to $\boldsymbol{X}_1^\top \boldsymbol{y} = 0.9$ and $\boldsymbol{X}_2^\top \boldsymbol{y} = 0.7$, respectively, $\|\boldsymbol{y}\| = 1$. The correlation of the predictors $\rho = \boldsymbol{X}_1^\top \boldsymbol{X}_2$ varies from 0.4 to 0.9 (horizontal axis). The restriction $\rho \in [0.4; 0.9]$ guarantees a valid correlation structure between $\boldsymbol{X}_1, \boldsymbol{X}_2$ and $\boldsymbol{y}$. The left panel displays the behaviour of the ols estimator (grey) with $\widehat{\beta}_1^{\text{ols}}, \widehat{\beta}_2^{\text{ols}}$ becoming more and more divergent, as does the lasso (black). The right panel depicts the same situation for the structured elastic net with $\boldsymbol{\Lambda} = (l_{jj'})_{1 \leq j,j' \leq 2}$, $l_{11} = l_{22} = 1$, $l_{12} = l_{21} = -1$ for $\lambda_2 \in \{0.25, 2.5\}$ and λ_1 as for the left panel. Note that the right panel does not contradict Proposition 1, since $\rho \to 1$ implies $\boldsymbol{X}_1^\top \boldsymbol{y} \to \boldsymbol{X}_2^\top \boldsymbol{y}$, whereas are $\boldsymbol{X}_1^\top \boldsymbol{y}$ and $\boldsymbol{X}_2^\top \boldsymbol{y}$ are kept fixed here. Further note that the situation displayed in the figure does not operate in terms of true coefficients, but studies instead the presence/absence of a grouping effect in dependence of a given data situation.

3.4 Uniqueness

Let us maintain the setup of the previous subsection. As pointed out by several authors, the lasso lacks uniqueness if $p \gg n$ or if there is collinearity among $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$. In this respect, the lasso differs from the elastic net, which is always determined uniquely. Concerning our proposal, whether the minimizer (3.4) is unique for $p \gg n$ depends on $\boldsymbol{\Lambda}$. Roughly speaking, if $\boldsymbol{\Lambda}$ provides a sufficient amount of prior information, then we have a unique minimizer. Uniqueness typically fails when $\boldsymbol{\Lambda}$ has nonzero entries only for a small fraction of the predictors. A sufficient condition can be derived by the following consideration: since $\boldsymbol{\Lambda}$ is assumed to be symmetric positive semidefinite, there exists a factorization $\boldsymbol{\Lambda} = \boldsymbol{Q}^\top \boldsymbol{Q}$. The stabilized covariance matrix in Eq. (3.4) can be written as

$$\widetilde{\boldsymbol{C}} = \widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}}, \quad \widetilde{\boldsymbol{X}} = \begin{pmatrix} \boldsymbol{X} \\ \lambda_2^{1/2} \boldsymbol{Q} \end{pmatrix}.$$

Thus, if $\text{rank}(\boldsymbol{X}) + \text{rank}(\lambda_2^{1/2}\boldsymbol{Q}) \geq p$ and the rows of $\boldsymbol{X}$ combined with the rows of $\lambda_2^{1/2}\boldsymbol{Q}$ form a linearly independent set, $\widetilde{\boldsymbol{C}}$ is of full rank and hence the structured elastic net is unique. In addition, this shows that even for $p \gg n$, in principle all features can be selected.

3.5 Double shrinkage

Zhou and Hastie (2005) pointed out that the elastic net effects twofold shrinkage of the coefficient vector towards zero, one produced by the ℓ^1- and a second one by the ℓ^2-constraint. In practice, this has the consequence that the elastic net estimates are shrunken too heavily, such that they are even outperformed by the lasso. For this reason, Zhou and Hastie (2005) suggest to undo the shrinkage induced by the ℓ^2-constraint, rescaling all coefficients by the factor $(1 + \lambda_2)$. A similar rationale might be appropriate for our proposal, depending on the structure of $\boldsymbol{\Lambda}$. For instance, if $\boldsymbol{\Lambda} = \boldsymbol{D}^\top\boldsymbol{D}$ (cf. Eq. (2.4)), then the shrinkage target of $\boldsymbol{\Lambda}$ is the constant vector. In the presence of sparsity, the constant vector is roughly equal to the zero vector. Since this is already the shrinkage target of the ℓ^1-constraint, we will observe double shrinkage as for the elastic net. As a remedy, one might consider to rescale the coefficients:

$$\widehat{\beta}_j \leftarrow (1 + \lambda_2 l_{jj})\widehat{\beta}_j, \quad j = 1, \ldots, p.$$

4 Consistency

The asymptotic analysis presented in this sections closely follows the ideas of Knight and Fu (2000) and Zhou (2006). Both have studied asymptotics for the lasso in linear regression under the following conditions.

(C.1) Given a sample S of size n, the data assumed to be generated according to the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon},$$

with $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^\top$ denoting the true parameter vector. The error terms $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$ are assumed to be i.i.d. with expectation 0 and constant variance $0 < \sigma^2 < \infty$.

(C.2)

$$\boldsymbol{C}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^\top = \frac{1}{n}\boldsymbol{X}_n^\top\boldsymbol{X}_n \to \boldsymbol{C} \text{ as } n \to \infty,$$

and the limit $\boldsymbol{C}$ is strictly positive definite.

(C.3)

$$\frac{1}{n}\max_{1 \leq i \leq n} \boldsymbol{x}_i^\top\boldsymbol{x}_i \to 0 \text{ as } n \to \infty.$$

Conditions (C.2) and (C.3) are weak in the sense that both hold if the $\boldsymbol{x}_i$ are i.i.d. with finite second moments. Conditions (C.1)-(C.3) ensure $\sqrt{n}$-consistency and asymptotic normality of the ordinary least squares estimator, i.e.

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n^{\text{ols}} - \boldsymbol{\beta}^*) \xrightarrow{\text{D}} N(\boldsymbol{0}, \sigma^2\boldsymbol{C}^{-1}),$$

where here and in the following, the sub- or superscript indicates that the corresponding quantity depends on the sample size n. Using conditions (C.1)-(C.3), Knight and Fu (2000) proved for the lasso that $\widehat{\boldsymbol{\beta}}^{\text{lasso}}$ is $\sqrt{n}$-consistent for $\boldsymbol{\beta}^*$ provided $\lambda_1^n/\sqrt{n} \to \lambda_1^0 \geq 0$. Zhou (2006) has shown that while taking $\lambda_1^n = O(\sqrt{n})$ provides the optimal rate for estimation, it leads to inconsistent feature selection. Define the active set as $A = \{j : \beta_j^* \neq 0\}$ and $A^c = \{1, \ldots, p\} \setminus A$ and let δ be an estimation procedure producing an estimate $\widehat{\boldsymbol{\beta}}^\delta$. Then δ is said to consistent in feature selection if

$$\begin{aligned} \lim_{n \to \infty} \mathrm{P}(\widehat{\beta}_{j,n}^\delta \neq 0) &= 1 \quad \text{for } j \in A, \\ \lim_{n \to \infty} \mathrm{P}(\widehat{\beta}_{j,n}^\delta = 0) &= 1 \quad \text{for } j \in A^c. \end{aligned}$$

Moreover, Zhou (2006) and Zhao and Yu (2006) have shown that if $\lambda_1^n/n \to 0$ and $\lambda_1^n/\sqrt{n} \to \infty$, the lasso has to satisfy a nontrivial condition, the so-called 'irrepresentable condition', to be selection consistent. Zhou (2006) proposed the adaptive lasso, a two-step estimation procedure, to fix this deficiency. In the following, these results will be adapted to the presence of a second quadratic penalty term.

Theorem 1. *Let conditions (C.1)-(C.3) hold. Define*

$$\widehat{\beta}_n = \operatorname*{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta}\|^2 + \lambda_1^n \|\boldsymbol{\beta}\|_1 + \lambda_2^n \boldsymbol{\beta}^\top \boldsymbol{\Lambda} \boldsymbol{\beta}.$$

Assume that $\lambda_1^n/\sqrt{n} \to \lambda_1^0 \geq 0$ *and* $\lambda_2^n/\sqrt{n} \to \lambda_2^0 \geq 0$. *Consider the random function*

$$\begin{aligned} V(\boldsymbol{u}) = &-2\boldsymbol{u}^\top \boldsymbol{w} + \boldsymbol{u}^\top \boldsymbol{C} \boldsymbol{u} \\ &+ \lambda_1^0 \sum_{j=1}^p u_j \operatorname{sign}(\beta_j^*) I(\beta_j^* \neq 0) + |u_j| I(\beta_j^* = 0) \\ &+ 2\lambda_2^0 \boldsymbol{u}^\top \boldsymbol{\Lambda} \boldsymbol{\beta}^*, \qquad \boldsymbol{w} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{C}). \end{aligned}$$

Then $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*) \xrightarrow{D} \operatorname{argmin} V(\boldsymbol{u})$.

Theorem 1 is analogous to Theorem 2 in Knight and Fu (2000) and establishes $\sqrt{n}$-consistency of $\widehat{\boldsymbol{\beta}}_n$, provided λ_1^n and λ_2^n are $O(\sqrt{n})$. Theorem 1 admits a straightforward extension to the class of generalized linear models (cf. Eq. (3.2)). Let the true model be defined by

$$\mathrm{E}[Y|\mathbb{X} = \boldsymbol{x}] = b'(f(\boldsymbol{x}; \boldsymbol{\beta}^*)), \quad f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}^*.$$

For the sake of a clearer presentation, we assume that $\beta_0^* = 0$. We study the estimator

$$\widehat{\boldsymbol{\beta}}_n = \operatorname*{argmin}_{\boldsymbol{\beta}} 2\phi^{-1} \sum_{i=1}^n b(f(\boldsymbol{x}_i; \boldsymbol{\beta})) - y_i f(\boldsymbol{x}_i; \boldsymbol{\beta}) + \lambda_1^n \|\boldsymbol{\beta}\|_1 + \lambda_2^n \boldsymbol{\beta}^\top \boldsymbol{\Lambda} \boldsymbol{\beta}. \tag{4.1}$$

We work with the following regularity conditions.

(G.1) The expected Fisher information

$$\mathcal{I} = \mathrm{E}[\phi^{-1} b''(f(\mathbb{X}); \boldsymbol{\beta}^*) \mathbb{X}\mathbb{X}^\top]$$

is finite and strictly positive definite.

(G.2) There exists a function M and an open neighbourhood U of $\boldsymbol{\beta}^*$ such that for all $\boldsymbol{\beta} \in U$

$$|b'''(f(\boldsymbol{x}); \boldsymbol{\beta})| \leq M(\boldsymbol{x}) < \infty \text{ for all } \boldsymbol{x},$$

and

$$\mathrm{E}[M(\mathbb{X})\, |X_j X_k X_l|] < \infty \quad \forall 1 \leq j, k, l \leq p.$$

Condition (G.2) is necessary for a Taylor expansion argument, as explained in the Appendix.

Theorem 2. *Let conditions (G.1) and (G.2) hold, consider the estimator (4.1) and let* $\lambda_1^n/\sqrt{n} \to \lambda_1^0 \geq 0$ *and* $\lambda_2^n/\sqrt{n} \to \lambda_2^0 \geq 0$. *Consider the random function*

$$\begin{aligned} W(\boldsymbol{u}) = &-2\boldsymbol{u}^\top \boldsymbol{w} + \boldsymbol{u}^\top \mathcal{I} \boldsymbol{u} \\ &+ \lambda_1^0 \sum_{j=1}^{p} u_j \operatorname{sign}(\beta_j^*) I(\beta_j^* \neq 0) + |u_j| I(\beta_j^* = 0) \\ &+ 2\lambda_2^0 \boldsymbol{u}^\top \boldsymbol{\Lambda} \boldsymbol{\beta}^*, \qquad \boldsymbol{w} \sim N(\boldsymbol{0}, \mathcal{I}). \end{aligned}$$

Then $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*) \xrightarrow{D} \operatorname{argmin} W(\boldsymbol{u})$.

Now let us turn to the question of selection consistency. In the situation of Theorem 1, if λ_1^n and λ_2^n both are $O(\sqrt{n})$, then $\widehat{\boldsymbol{\beta}}_n$ cannot be selection consistent. To see this, let $\widehat{\boldsymbol{u}}$ denote the minimizer of $V(\boldsymbol{u})$. Without loss of generality, let us assume that $\boldsymbol{\beta}^* = ([\boldsymbol{\beta}_A^*]^\top, [\boldsymbol{\beta}_{A^c}^*]^\top)^\top$, where here and in the following, the subscripts A and A^c refer to active and inactive set, respectively. Selection consistency requires that $\widehat{\boldsymbol{u}}$ is of the form $\widehat{\boldsymbol{u}} = (\widehat{\boldsymbol{u}}_A^\top, \boldsymbol{0}^\top)^\top$. Evaluation of $V(\boldsymbol{u})$ at $\widehat{\boldsymbol{u}}$ yields:

$$V(\widehat{\boldsymbol{u}}) = -2\widehat{\boldsymbol{u}}_A^\top \boldsymbol{w}_A + \widehat{\boldsymbol{u}}_A^\top \boldsymbol{C}_A \widehat{\boldsymbol{u}}_A + \lambda_1^0 \widehat{\boldsymbol{u}}_A^\top \boldsymbol{s}_A + 2\lambda_2^0 \widehat{\boldsymbol{u}}_A^\top \boldsymbol{\Lambda}_A \boldsymbol{\beta}_A^*, \;\; \boldsymbol{s}_A = (\operatorname{sign}(\beta_j^*),\ j \in A)^\top.$$

Since $\widehat{\boldsymbol{u}}_A$ is a minimizer, differentiation yields

$$\widehat{\boldsymbol{u}}_A = \boldsymbol{C}_A^{-1} \left(\boldsymbol{w}_A - \frac{\lambda_1^0}{2} \boldsymbol{s}_A - \lambda_2^0 \boldsymbol{\Lambda}_A \boldsymbol{\beta}_A^* \right). \tag{4.2}$$

Now consider the partitioning schemes

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{C}_A & \boldsymbol{C}_{AA^c} \\ \boldsymbol{C}_{A^cA} & \boldsymbol{C}_{A^c} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_A & \boldsymbol{\Lambda}_{AA^c} \\ \boldsymbol{\Lambda}_{A^cA} & \boldsymbol{\Lambda}_{A^c} \end{pmatrix}. \tag{4.3}$$

Since $\widehat{\boldsymbol{u}}_{A^c} = \boldsymbol{0}$, the Karush-Kuhn-Tucker (KKT) conditions imply that

$$|\lambda_2^0 \boldsymbol{\Lambda}_{A^cA} \boldsymbol{\beta}_A^* + \boldsymbol{C}_{A^cA} \widehat{\boldsymbol{u}}_A - \boldsymbol{w}_{A^c}| \leq \frac{\lambda_1^0}{2} \boldsymbol{1}, \tag{4.4}$$

where the inequality is interpreted componentwise. Substituting the right hand side of Eq. (4.2) into Eq. (4.4) and rearranging terms, one obtains

$$\left| \left[\boldsymbol{\Lambda}_{A^cA} - \boldsymbol{C}_{A^cA} \boldsymbol{C}_A^{-1} \boldsymbol{\Lambda}_A \right] \lambda_2^0 \boldsymbol{\beta}_A^* + \boldsymbol{C}_{A^cA} \boldsymbol{C}_A^{-1} \left(\boldsymbol{w}_A - \frac{\lambda_1^0}{2} \boldsymbol{s}_A \right) - \boldsymbol{w}_{A^c} \right| \leq \frac{\lambda_1^0}{2} \boldsymbol{1}. \tag{4.5}$$

Since $\boldsymbol{w}$ is random quantity, the system of inequalities only holds with a certain probability, implying inconsistency of selection in general. In addition, the system (4.5) shows that selection consistency depends on $\boldsymbol{C}$, $\boldsymbol{\Lambda}$ and also $\boldsymbol{\beta}_A^*$. Selection consistency can be achieved if one lets λ_1^n, λ_2^n grow more strongly and if the quantities $\boldsymbol{C}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{\beta}_A^*$ fulfill a nontrivial condition, which can be seen as analog to the irrepresentable condition of the lasso.

Theorem 3. *In the setup of Theorem 1, let* $\lambda_1^n/n \to 0$, $\lambda_1^n/\sqrt{n} \to \infty$, $\lambda_2^n/\lambda_1^n \to R$, $0 < R < \infty$. *Then, if selection consistency holds, the following condition must be fulfilled: there exists a sign vector* $\boldsymbol{s}_A$ *such that*

$$|-\boldsymbol{C}_{A^cA}\boldsymbol{C}_A^{-1}(\boldsymbol{s}_A + 2R\boldsymbol{\Lambda}_A\boldsymbol{\beta}_A^*) + 2R\boldsymbol{\Lambda}_{A^cA}\boldsymbol{\beta}_A^*| \leq \mathbf{1},$$

where the inequality is interpreted componentwise.

While this condition is interesting from a theoretical point of view, it is impossible to check in practice, since $\boldsymbol{\beta}_A^*$ is unknown.
Selection consistency can be achieved by a two-step estimation strategy introduced in Zhou (2006) under the name adaptive lasso, which replaces ℓ^1-regularization uniform in β_j, $j = 1, \ldots, p$, by a weighted variant $J(\boldsymbol{\beta}) = \sum_{j=1}^p \omega_j|\beta_j|$, where the weights $\{\omega_j\}_{j=1}^p$ are determined adaptively as function of an 'initial estimator' $\widehat{\boldsymbol{\beta}}^{\text{init}}$:

$$\omega_j = |\widehat{\beta}_j^{\text{init}}|^{-\gamma}, \quad \gamma > 0, \quad j = 1, \ldots, p. \tag{4.6}$$

In terms of selection consistency this strategy turns out to be favourable for our proposal, too.

Theorem 4. *In the setup of Theorem 1, define*

$$\widehat{\boldsymbol{\beta}}_n^{adaptive} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta}\|^2 + \lambda_1^n \sum_{j=1}^p \omega_j|\beta_j| + \lambda_2^n\boldsymbol{\beta}^\top\boldsymbol{\Lambda}\boldsymbol{\beta},$$

where the weights are as in Eq. (4.6), and suppose that the initial estimator satisfies

$$r_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*) = O_{\mathrm{P}}(1), \quad r_n \to \infty \text{ as } n \to \infty.$$

Furthermore, suppose that

$$r_n^\gamma\lambda_1^n n^{-1/2} \to \infty, \quad \lambda_1^n n^{-1/2} \to 0, \quad \lambda_2^n n^{-1/2} \to \lambda_2^0 \geq 0$$

as $n \to \infty$. *Then*

$$\begin{aligned} &(1)\ \sqrt{n}(\widehat{\boldsymbol{\beta}}_{A,n}^{adaptive} - \boldsymbol{\beta}_A^*) \xrightarrow{D} N(-\lambda_2^0\boldsymbol{C}_A^{-1}\boldsymbol{\Lambda}_A\boldsymbol{\beta}_A^*, \boldsymbol{C}_A^{-1}),\\ &(2)\ \lim_{n\to\infty} \mathrm{P}(\widehat{\boldsymbol{\beta}}_{A^c,n}^{adaptive} = \mathbf{0}) = 1. \end{aligned}$$

Theorem 4 implies that the adaptive structured elastic net $\widehat{\boldsymbol{\beta}}^{\text{adaptive}}$ is an oracle estimation procedure (Fan and Li (2001)) if the bias term in (1) vanishes, which is the case if $\boldsymbol{\beta}_A^*$ resides in the nullspace of $\boldsymbol{\Lambda}_A$. Interestingly, if $\boldsymbol{\Lambda}$ equals the combinatorial graph Laplacian (cf. Section 2.1), this happens if and only if $\boldsymbol{\beta}_A^*$ has constant entries and A specifies a connected component in the underlying graph.
Concerning the choice of the initial estimator, the ridge estimator has worked well for us in practice, provided the ridge parameter is chosen appropriately. While γ may be treated as a tuning parameter, we have set γ equal to 1 in all our data analyses. Finally, we remark that while Theorem 4 applies to linear regression, it can be extended to hold for generalized linear models, in a similar way as the extension of Theorem 1 to Theorem 2.

5 Computation

This section discusses aspects concerning computation and model selection for the structured elastic net estimator when the loss function is the negative log-likelihood of a generalized linear model (3.1).

5.1 Data augmentation

From the discussions in Subsections 3.3 and 3.4, it follows that the structured elastic net for squared loss, assuming centered data, can be recast as lasso on augmented data

$$\widetilde{\boldsymbol{X}} = \begin{pmatrix} \boldsymbol{X} \\ \lambda_2^{1/2}\boldsymbol{Q} \end{pmatrix}_{(n+p)\times p}, \quad \widetilde{\boldsymbol{y}} = \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{pmatrix}_{(n+p)\times 1}, \quad \boldsymbol{\Lambda} = \boldsymbol{Q}^\top\boldsymbol{Q},$$

and hence algorithms available for computing the lasso, notably LARS (Efron et al. (2004)) may be applied, which computes for fixed λ_2 and varying λ_1 the piecewise linear solution path $\widehat{\boldsymbol{\beta}}(\lambda_1;\lambda_2)$. In order to fit arbitrary regularized generalized linear models, the augmented data representation has to be modified. Without regularization, estimators in generalized linear models are obtained by iteratively computing weighted least squares estimators:

$$\begin{aligned}
\begin{pmatrix} \widehat{\beta}_0^{(k+1)} \\ \widehat{\boldsymbol{\beta}}^{(k+1)} \end{pmatrix} &= \left([\mathbf{1}\ \boldsymbol{X}]^\top \boldsymbol{W}^{(k)} [\mathbf{1}\ \boldsymbol{X}]\right)^{-1} [\mathbf{1}\ \boldsymbol{X}]^\top \boldsymbol{W}^{(k)} \boldsymbol{z}^{(k)}, \\
\boldsymbol{z}^{(k)} &= \boldsymbol{f}^{(k)} + [\boldsymbol{W}^{(k)}]^{-1}(\boldsymbol{y} - \boldsymbol{\mu}^{(k)}), \\
\boldsymbol{f}^{(k)} &= (f_1^{(k)}, \ldots, f_n^{(k)})^\top, \quad f_i^{(k)} = \widehat{\beta}_0^{(k)} + \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}^{(k)}, \quad i = 1, \ldots, n, \\
\boldsymbol{\mu}^{(k)} &= (\mu_1^{(k)}, \ldots, \mu_n^{(k)})^\top, \quad \mu_i^{(k)} = b'(f_i^{(k)}), \quad i = 1, \ldots, n, \\
\boldsymbol{W}^{(k)} &= \operatorname{diag}(w_1^{(k)}, \ldots, w_n^{(k)}), \quad w_i^{(k)} = \phi^{-1} b''(f_i^{(k)}), \quad i = 1, \ldots, n.
\end{aligned} \tag{5.1}$$

Note that the design matrix additionally includes a constant term $\mathbf{1}$. Turning back to the structured elastic net, an adaptation of the augmented data approach iteratively determines

$$\begin{pmatrix} \widehat{\beta}_0^{(k+1)} \\ \widehat{\boldsymbol{\beta}}^{(k+1)} \end{pmatrix} = \underset{(\beta_0,\boldsymbol{\beta})}{\operatorname{argmin}} \sum_{i=1}^{n+p} \widetilde{w}_i^{(k)} \left(\widetilde{z}_i^{(k)} - \widetilde{\boldsymbol{x}}_i^\top \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta} \end{pmatrix} \right)^2 + \lambda_1 \|\boldsymbol{\beta}\|_1,$$

with

$$\begin{aligned}
&\widetilde{w}_i^{(k)} = w_i^{(k)}, \quad i = 1, \ldots, n, \text{ as in Eq. (5.1)}, \quad \widetilde{w}_i^{(k)} = 1, \quad i = (n+1), \ldots, (n+p), \\
&\widetilde{z}_i^{(k)} = z_i^{(k)} \quad i = 1, \ldots, n, \text{ as in Eq. (5.1)}, \quad \widetilde{z}_i^{(k)} = 0, \quad i = (n+1), \ldots, (n+p), \\
&\widetilde{\boldsymbol{x}}_i = (1\ \boldsymbol{x}_i^\top)^\top, \quad i = 1, \ldots, n, \quad \widetilde{\boldsymbol{x}}_i = (0\ \sqrt{\lambda_2}\boldsymbol{q}_i^\top)^\top, \quad i = (n+1), \ldots, (n+p),
\end{aligned}$$

with $\boldsymbol{q}_i^\top$ denoting the i-th row of $\boldsymbol{Q}$.

5.2 Cyclical coordinate descent

Recently, cyclical coordinate descent (CCD) approaches for obtaining the lasso solution in generalized linear models have gained much popularity, e.g. Genkin et al. (2007), Wu and Lange (2008), Friedman et al. (2007). The latter have studied convex functions of the form

$$F(\boldsymbol{\beta}) = S(\boldsymbol{\beta}) + J(\boldsymbol{\beta}), \quad \boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top, \quad J(\boldsymbol{\beta}) = \sum_{j=1}^{p} \iota_j(\beta_j), \tag{5.2}$$

to be minimized w.r.t. $\boldsymbol{\beta}$. We suppose that $F(\boldsymbol{\beta})$ is convex, that $S(\boldsymbol{\beta})$ is smooth and that $J(\boldsymbol{\beta})$ is continuous and separable in the β_j. CCD optimizes one coordinate at a time, with

the remaining ones kept fixed. Convergence analysis for this setup is studied in Tseng (2001). As reported in Friedman et al. (2007), the CCD approach turns out be remarkably efficient as well as versatile, allowing the computation of the lasso, the elastic net and several related procedures along a fine grid of λ_1-values, using the current estimate as 'warm start' for the next grid point. While CCD could be applied to the augmented data representation of the previous subsection to compute the structured elastic net estimator, we prefer a CCD algorithm directly adapted to our specific problem, without the need to determine the root $\boldsymbol{Q}$. Note that such an algorithm is possible, since the structured elastic net criterion matches the structure of $F(\boldsymbol{\beta})$ in Eq. (5.2) with $\iota_j = \lambda_1|\beta_j|,\ \ j = 1, \ldots, p$.We first state the algorithm for squared loss, and straightforward modifications admit the extension to generalized linear models.
Consider the structured elastic net estimator for $L(y, f(\boldsymbol{x}; \beta_0, \boldsymbol{\beta})) = (y - \beta_0 - \boldsymbol{x}^\top\boldsymbol{\beta})^2$. The aim is to determine an estimate $\widehat{\beta}_j$ given $\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}_{-j},\ \widehat{\boldsymbol{\beta}}_{-j} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_{j-1}, \widehat{\beta}_{j+1}, \ldots, \widehat{\beta}_p)^\top$. Likewise, the j-th row of $\boldsymbol{\Lambda}$ can be divided into l_{jj} and $\boldsymbol{l}_{-j} = (l_{jj'})_{j'\neq j}^\top$. The KKT conditions imply that if $\widehat{\beta}_j = 0$

$$\left| -2\boldsymbol{X}_j^\top \left(\boldsymbol{y} - \widehat{\beta}_0 - \sum_{j'\neq j} \boldsymbol{X}_{j'}\widehat{\beta}_{j'} \right) + 2\lambda_2 \boldsymbol{l}_{-j}^\top \widehat{\boldsymbol{\beta}}_{-j} \right| \leq \lambda_1, \tag{5.3}$$

and otherwise,

$$-2\boldsymbol{X}_j^\top(\boldsymbol{y} - \widehat{\beta}_0 - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) + 2\lambda_2 l_{jj}\widehat{\beta}_j + 2\lambda_2 \boldsymbol{l}_{-j}^\top \widehat{\boldsymbol{\beta}}_{-j} + \lambda_1 \operatorname{sign}(\widehat{\beta}_j) = 0. \tag{5.4}$$

Solving for $\widehat{\beta}_j$ in Eq. (5.4) yields

$$\begin{aligned} \widehat{\beta}_j &= \frac{\boldsymbol{X}_j^\top(\boldsymbol{r}_j - \boldsymbol{l}_{-j}^\top \widehat{\boldsymbol{\beta}}_{-j}) - \frac{\lambda_1}{2}\operatorname{sign}(\widehat{\beta}_j)}{\|\boldsymbol{X}_j\|^2 + \lambda_2 l_{jj}}, \\ \boldsymbol{r}_j &= \boldsymbol{y} - \widehat{\beta}_0 - \sum_{j'\neq j} \boldsymbol{X}_{j'}\widehat{\beta}_{j'}. \end{aligned} \tag{5.5}$$

Combining Eq. (5.3) with Eq. (5.4) yields the update formula

$$\widehat{\beta}_j \leftarrow \frac{\left[\left| \boldsymbol{X}_j^\top \boldsymbol{r}_j - \boldsymbol{l}_{-j}^\top \widehat{\boldsymbol{\beta}}_{-j} \right| - \frac{\lambda_1}{2} \right]_+ \operatorname{sign}(\boldsymbol{X}_j^\top \boldsymbol{r}_j - \boldsymbol{l}_{-j}^\top \widehat{\boldsymbol{\beta}}_{-j})}{\|\boldsymbol{X}_j\|^2 + \lambda_2 l_{jj}}, \quad j = 1, \ldots, p, \quad [z]_+ = \max(z, 0).$$

For the intercept, we have the update

$$\widehat{\beta}_0 \leftarrow \frac{\sum_{i=1}^n y_i - \sum_{j=1}^p \boldsymbol{X}_j \widehat{\beta}_j}{n}.$$

For generalized linear models, the algorithm involves one outer loop in which the quantities of iteratively weighted least squares (5.1) are updated, and one inner loop for CCD as just described. To be more precise, given estimates $\widehat{\beta}_0^{(k)}, \widehat{\boldsymbol{\beta}}^{(k)}$, we compute the $\boldsymbol{W}^{(k)}$ and $\boldsymbol{z}^{(k)}$ as given in Eq. (5.1). Defining the working response $\widetilde{\boldsymbol{z}}^{(k)} = [\boldsymbol{W}^{(k)}]^{1/2}\boldsymbol{z}^{(k)}$, $\widetilde{\boldsymbol{X}}^{(k)} = [\boldsymbol{W}^{(k)}]^{1/2}\boldsymbol{X}$ and a modified intercept $\widetilde{\mathbf{1}}^{(k)} = [\widetilde{\boldsymbol{W}}^{(k)}]^{1/2}\mathbf{1}$, the data $(\widetilde{\boldsymbol{z}}^{(k)}, [\widetilde{\mathbf{1}}^{(k)}\ \widetilde{\boldsymbol{X}}^{(k)}])$ can be plugged into the CCD algorithm for squared loss. Once an inner loop has converged, the new estimates $\widehat{\beta}_0^{(k+1)}, \widehat{\boldsymbol{\beta}}^{(k+1)}$ are used to obtain $(\widetilde{\boldsymbol{z}}^{(k+1)}, [\widetilde{\mathbf{1}}^{(k+1)}\ \widetilde{\boldsymbol{X}}^{(k+1)}])$.

5.3 Goeman's algorithm

The third algorithm we provide is an adaptation of a recent proposal in Goeman (2007), which combines full gradient descent- with Newton-Raphson-steps in subdomains where the gradient of the structured elastic net objective function is continuous. For a loss function of the type (3.1), the gradient and the Hessian of the differentiable part of the structured elastic net criterion are given by

$$\begin{pmatrix} \nabla_{\beta_0} \\ \nabla_{\boldsymbol{\beta}} \end{pmatrix} = \begin{pmatrix} -\mathbf{1}^\top(\boldsymbol{y} - \boldsymbol{\mu}(\beta_0, \boldsymbol{\beta})) \\ -\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{\mu}(\beta_0, \boldsymbol{\beta})) + 2\lambda_2 \boldsymbol{\Lambda} \end{pmatrix},$$
$$\nabla^2_{\beta_0,\boldsymbol{\beta}} = \begin{pmatrix} \nabla^2_{\beta_0} & \frac{\partial}{\partial \boldsymbol{\beta}^\top}\nabla_{\beta_0} \\ \frac{\partial}{\partial \beta_0}\nabla_{\boldsymbol{\beta}} & \nabla^2_{\boldsymbol{\beta}} \end{pmatrix} = \begin{pmatrix} \mathbf{1}^\top \boldsymbol{W}(\beta_0, \boldsymbol{\beta})\mathbf{1} & \mathbf{1}^\top \boldsymbol{W}(\beta_0, \boldsymbol{\beta})\boldsymbol{X} \\ \boldsymbol{X}^\top \boldsymbol{W}(\beta_0, \boldsymbol{\beta})\mathbf{1} & \boldsymbol{X}^\top \boldsymbol{W}(\beta_0, \boldsymbol{\beta})\boldsymbol{X} + 2\lambda_2 \boldsymbol{\Lambda} \end{pmatrix}. \tag{5.6}$$

Let the superscript $^{(k)}$ refer to the iteration counter and let A denote the currently active set, i.e. $A = \{j \in \{1, \ldots, p\} : \widehat{\beta}_j^{(k)} \neq 0\}$. In principle, the idea is to take gradient descent steps until it is favourable to switch to Newton-Raphson. For the former, one looks for a suitable descent direction $\boldsymbol{v}$. For $j \in A$, the ℓ^1-part of the regularizer is differentiable as well: combined with the gradient in Eq. (5.6), this yields descent directions

$$v_0 \leftarrow -\nabla_{\widehat{\beta}_0^{(k)}}, \quad \boldsymbol{v}_A \leftarrow -[\nabla_{\widehat{\boldsymbol{\beta}}^{(k)}}]_A - \lambda_1 \operatorname{sign}(\widehat{\boldsymbol{\beta}}_A),$$

where the evaluation of the sign function here and in the following is understood componentwise. For $j \in A^c$, we have to distinguish two cases. If $|[\nabla_{\widehat{\boldsymbol{\beta}}^{(k)}}]_j| > \lambda_1$, one enlarges the active set, i.e. $A \leftarrow A \cup \{j\}$, and sets $v_j \leftarrow -[\nabla_{\widehat{\boldsymbol{\beta}}^{(k)}}]_j - \lambda_1 \operatorname{sign}([\nabla_{\widehat{\boldsymbol{\beta}}^{(k)}}]_j)$. Otherwise, there is no move for coordinate $j : v_j \leftarrow 0$. We define $\boldsymbol{v} = (v_1, \ldots, v_p)^\top$. The next issue to consider is the determination of an appropriate step length t for the gradient descent update $\widehat{\boldsymbol{\beta}}^{(k+1)} \leftarrow \widehat{\boldsymbol{\beta}}^{(k)} + t\boldsymbol{v}$. One has to take care that the stepsize is sufficiently small in order to remain within a subdomain of gradient continuity, i.e. one has to ensure that $\operatorname{sign}(\widehat{\boldsymbol{\beta}}^{(k+1)}) = \operatorname{sign}(\widehat{\boldsymbol{\beta}}^{(k)})$. This naturally imposes an upper bound on the step size, denoted by t^{edge}:

$$0 < t^{\text{edge}} = \min_{j \in A} \left\{ -\frac{\widehat{\beta}_j^{(k)}}{v_j} : \quad \operatorname{sign}(\widehat{\beta}_j^{(k)}) = -\operatorname{sign}(v_j) \right\}. \tag{5.7}$$

In a subdomain of gradient continuity, assuming $t < t^{\text{edge}}$, one can perform a quadratic Taylor approximation of the objective function $F(\beta_0, \boldsymbol{\beta})$ around the current estimates $\widehat{\beta}_0^{(k)}, \widehat{\boldsymbol{\beta}}^{(k)}$:

$$F(\widehat{\beta}_0^{(k)}, \widehat{\boldsymbol{\beta}}^{(k)} + t\boldsymbol{v}) \approx F(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}^{(k)}) - t\boldsymbol{v}^\top \boldsymbol{v} + \frac{1}{2} t^2 \boldsymbol{v}^\top \nabla^2_{\widehat{\boldsymbol{\beta}}^{(k)}} \boldsymbol{v}. \tag{5.8}$$

Differentiating the r.h.s. of approximation (5.8) with respect to t and setting the result equal to zero, it follows that the minimum of the Taylor approximation is achieved by choosing the stepsize

$$t^{\text{opt}} = \frac{\boldsymbol{v}^\top \boldsymbol{v}}{\boldsymbol{v}^\top \nabla^2_{\widehat{\boldsymbol{\beta}}^{(k)}} \boldsymbol{v}},$$

provided the denominator is nonzero. At this place, one checks whether $t^{\text{opt}} < t^{\text{edge}}$. If this is the case, the gradient descent update is performed with step size $t = t^{\text{topt}}$, i.e.

$$\widehat{\boldsymbol{\beta}}^{(k+1)} \leftarrow \widehat{\boldsymbol{\beta}}^{(k)} + t^{\text{opt}} \boldsymbol{v}.$$

Otherwise, the active set is reduced: $A \leftarrow A \setminus \{j_0\}$ and $\widehat{\beta}_{j_0}^{(k+1)} \leftarrow 0$, where j_0 is the index for which the minimum in Equation (5.7) is attained, and the gradient descent step is performed with stepsize t^{edge} for the reduced active set. For the intercept, one may use $\widehat{\beta}_0^{(k+1)} = \widehat{\beta}_0^{(k)} + t_0 v_0$, $t_0 = 1/\nabla^2_{\widehat{\beta}_0^{(k)}}$.
Near the minimum, Newton-Raphson can considerably speed up convergence. Newton-Raphson-steps are only possible within a single subdomain of gradient continuity. As Newton-Raphson-steps are computationally expensive, they should be avoided if they are likely to fail. The Newton-Raphson-update restricted to the active set is given by

$$\widehat{\boldsymbol{\beta}}_A^{(k+1)} \leftarrow \widehat{\boldsymbol{\beta}}_A^{(k)} - [\nabla^2_{\widehat{\boldsymbol{\beta}}^{(k)}}]_A^{-1} \left\{ [\nabla_{\widehat{\boldsymbol{\beta}}^{(k)}}]_A + \lambda_1 \operatorname{sign}(\widehat{\boldsymbol{\beta}}_A^{(k)}) \right\}. \tag{5.9}$$

The Newton-Raphson-step is accepted only if $\operatorname{sign}(\widehat{\boldsymbol{\beta}}_A^{(k+1)}) = \operatorname{sign}(\widehat{\boldsymbol{\beta}}_A^{(k)})$. Otherwise, a subdomain of gradient continuity has been left. Following Goeman (2007), a Newton-Raphson step should be attempted only if $t^{\text{opt}} < t^{\text{edge}}$ in a preceding gradient descent step. Moreover, Newton-Raphson-steps are neither favourable in the very first iterations nor when the active set has not yet changed since the last failed attempt.
If $p \gg n$, it is possible that the size of the active set exceeds by far the sample size. In this case, the Newton-Raphson-steps given by the update (5.9) are inefficient. The matrix inversion requires $O(|A|^3)$ operations, but it is immediately seen that a reduction to $O(n^3)$ operations can be achieved by reparametrization. We set $\widehat{\boldsymbol{\beta}}_A^{(k)} = \boldsymbol{X}_A^\top \widehat{\boldsymbol{\gamma}}^{(k)}$, so that the gradient has the form

$$\nabla_{\widehat{\boldsymbol{\gamma}}^{(k)}} = -\boldsymbol{X}_A \boldsymbol{X}_A^\top (\boldsymbol{y} - \boldsymbol{\mu}) + \lambda_1 \boldsymbol{X}_A \operatorname{sign}(\boldsymbol{X}_A^\top \widehat{\boldsymbol{\gamma}}^{(k)}) + 2\lambda_2 \boldsymbol{X}_A \boldsymbol{\Lambda}_A \boldsymbol{X}_A^\top \widehat{\boldsymbol{\gamma}}^{(k)} = \boldsymbol{X}_A [\nabla_{\widehat{\boldsymbol{\beta}}^{(k)}}]_A,$$

and for the Hessian, one obtains

$$\nabla^2_{\widehat{\boldsymbol{\gamma}}^{(k)}} = \boldsymbol{X}_A \boldsymbol{X}_A^\top \boldsymbol{W}_A \boldsymbol{X}_A \boldsymbol{X}_A^\top + 2\lambda_2 \boldsymbol{X}_A \boldsymbol{\Lambda}_A \boldsymbol{X}_A^\top = \boldsymbol{X}_A [\nabla^2_{\widehat{\boldsymbol{\beta}}^{(k)}}]_A \boldsymbol{X}_A^\top.$$

For the Newton-Raphson-steps, this has the following algorithmic implications.

- One checks if $|A| \gg n$, where '$\gg$' means that the ratio $|A|/n$ exceeds a factor related to the specific problem.
- If yes, one reparametrizes $\widehat{\boldsymbol{\beta}}_A^{(k)}$ to $\widehat{\boldsymbol{\gamma}}^{(k)}$ by solving the linear equation $\boldsymbol{X}_A \boldsymbol{X}_A^\top \widehat{\boldsymbol{\gamma}}^{(k)} = \boldsymbol{X}_A \widehat{\boldsymbol{\beta}}_A^{(k)}$. Otherwise, one proceeds according to the update (5.9).
- One applies Newton-Raphson to $\widehat{\boldsymbol{\gamma}}^{(k)}$, using $\nabla_{\widehat{\boldsymbol{\gamma}}^{(k)}}$, $\nabla^2_{\widehat{\boldsymbol{\gamma}}^{(k)}}$ given above to obtain $\widehat{\boldsymbol{\gamma}}^{(k+1)}$.
- One backtransforms $\widehat{\boldsymbol{\beta}}_A^{(k+1)} = \boldsymbol{X}_A^\top \widehat{\boldsymbol{\gamma}}^{(k+1)}$.
- The Newton-Raphson step is accepted only if $\operatorname{sign}(\widehat{\boldsymbol{\beta}}_A^{(k+1)}) = \operatorname{sign}(\widehat{\boldsymbol{\beta}}_A^{(k)})$.

5.4 Comparison

Cyclical coordinate descent is conceptually simple and involves only vector operations, also avoiding the direct storage of the matrix $\boldsymbol{X}^\top \boldsymbol{X}$. If organized properly, one inner loop involves $O(p(n+p))$ operations. Although we do not give a proof of convergence, the algorithm is well-founded since it can be embedded into the framework of Tseng (2001). This is contrary to Goeman's algorithm, for which no convergence analysis has been performed yet. Several parts resort to heuristics, but it generally works well in practice and has turned out to be faster than CCD, which can be rather slow for large p.

5.5 Degrees of freedom

The 'degrees of freedom' of some estimation procedure δ are an integral part of model selection criteria such as GCV, C_p or the AIC. Therefore, it would be desirable to determine the degrees of freedom of the structured elastic net as a function of the hyperparameters λ_1 and λ_2. Given a sample $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ such that $\boldsymbol{\mu} = \mathrm{E}[\boldsymbol{y}|\boldsymbol{X}]$ and $\mathrm{var}[\boldsymbol{y}|\boldsymbol{X}] = \sigma^2\boldsymbol{I}$, denote by $\widehat{\boldsymbol{\mu}}$ the fit when applying δ to S. In the framework of Stein's unbiased risk estimation (Stein (1981)), the degrees of freedom of $\widehat{\boldsymbol{\mu}}$ are given by

$$\mathrm{df}(\widehat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} \mathrm{cov}(\widehat{\mu}_i, y_i)/\sigma^2. \tag{5.10}$$

This definition can easily be evaluated if $\widehat{\boldsymbol{\mu}} = \boldsymbol{A}\boldsymbol{y}$ for some matrix $\boldsymbol{A}$ independent of $\boldsymbol{y}$, in which case $\mathrm{df}(\widehat{\boldsymbol{\mu}}) = \mathrm{tr}(\boldsymbol{A})$. The lasso fit is a nonlinear function of $\boldsymbol{y}$, which considerably complicates the evaluation of (5.10). A heuristic due to Tibshirani (1996), see also Fan and Li (2001), represents the lasso fit for linear models as weighted ridge fit:

$$\widehat{\boldsymbol{\mu}} \approx [\mathbf{1}\ \boldsymbol{X}]([\mathbf{1}\ \boldsymbol{X}]^\top[\mathbf{1}\ \boldsymbol{X}] + \lambda_1\boldsymbol{\Omega})^{-1}[\mathbf{1}\ \boldsymbol{X}]^\top\boldsymbol{y} = \boldsymbol{A}\boldsymbol{y},$$

where $\boldsymbol{\Omega} = \mathrm{diag}(0, (\widehat{\beta}_1^{\mathrm{lasso}})^2/|\widehat{\beta}_1^{\mathrm{lasso}}|, \ldots, (\widehat{\beta}_p^{\mathrm{lasso}})^2/|\widehat{\beta}_p^{\mathrm{lasso}}|)$, with the convention $0^{-1} = 0$. The degrees of freedom according to Eq. (5.10) are then computed as the trace of $\boldsymbol{A}$. This heuristic can be modified for the structured elastic net fit via

$$\widehat{\boldsymbol{\mu}} = [\mathbf{1}\ \boldsymbol{X}]([\mathbf{1}\ \boldsymbol{X}]^\top[\mathbf{1}\ \boldsymbol{X}] + \lambda_1\boldsymbol{\Omega} + \lambda_2\widetilde{\boldsymbol{\Lambda}})^{-1}[\mathbf{1}\ \boldsymbol{X}]^\top\boldsymbol{y}, \quad \widetilde{\boldsymbol{\Lambda}} = \begin{pmatrix} 0 & \mathbf{0}^\top \\ \mathbf{0} & \boldsymbol{\Lambda} \end{pmatrix}.$$

A similar formula can be applied for the adaptive structured elastic net by rescaling the entries of $\boldsymbol{\Omega}$. For generalized linear models, we can make use of the iteratively weighted least squares approximation (5.1):

$$\widehat{\boldsymbol{\mu}} \approx [\boldsymbol{W}(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}})]^{1/2}[\mathbf{1}\ \boldsymbol{X}]([\mathbf{1}\ \boldsymbol{X}]^\top\boldsymbol{W}(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}})[\mathbf{1}\ \boldsymbol{X}] + \lambda_1\boldsymbol{\Omega} + \lambda_2\widetilde{\boldsymbol{\Lambda}})^{-1}[\mathbf{1}\ \boldsymbol{X}]^\top[\boldsymbol{W}(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}})]^{1/2}\boldsymbol{z}.$$

5.6 Standard errors

The heuristic of the previous section also turns out to be useful to obtain approximate standard errors for the estimated coefficients. For linear models, we have

$$(\widehat{\beta}_0\ \widehat{\boldsymbol{\beta}}^\top)^\top \approx \underbrace{([\mathbf{1}\ \boldsymbol{X}]^\top[\mathbf{1}\ \boldsymbol{X}] + \lambda_1\boldsymbol{\Omega} + \lambda_2\widetilde{\boldsymbol{\Lambda}})^{-1}}_{\boldsymbol{\Gamma}}[\mathbf{1}\ \boldsymbol{X}]^\top\boldsymbol{y},$$

concluding that one may use standard errors

$$\mathrm{se}(\widehat{\beta}_j) = \widehat{\sigma}\sqrt{(\boldsymbol{\Gamma}[\mathbf{1}\ \boldsymbol{X}]^\top[\mathbf{1}\ \boldsymbol{X}]\boldsymbol{\Gamma})_{jj}}, \quad j = 0, \ldots, p,$$

where $\widehat{\sigma}$ denotes an estimator of the standard deviation of the error terms. Likewise for generalized linear models, we have the approximation

$$(\widehat{\beta}_0\ \widehat{\boldsymbol{\beta}}^\top)^\top \approx \underbrace{([\mathbf{1}\ \boldsymbol{X}]^\top\boldsymbol{W}(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}})[\mathbf{1}\ \boldsymbol{X}] + \lambda_1\boldsymbol{\Omega} + \lambda_2\widetilde{\boldsymbol{\Lambda}})^{-1}}_{\boldsymbol{\Upsilon}(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}})}[\mathbf{1}\ \boldsymbol{X}]^\top\boldsymbol{W}(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}})\boldsymbol{z},$$

and consequently

$$\mathrm{se}(\widehat{\beta}_j) = \widehat{\phi}\sqrt{\{\boldsymbol{\Upsilon}(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}})[\mathbf{1}\ \boldsymbol{X}]^\top\boldsymbol{W}(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}})[\mathbf{1}\ \boldsymbol{X}]\boldsymbol{\Upsilon}(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}})\}_{jj}}, \quad j = 0, \ldots, p.$$

Note that this approach is not adequate for those $\widehat{\beta}_j$ equal to zero, since it always yields $\widehat{\mathrm{se}}(\widehat{\beta}_j) = 0$. An alternative is the bootstrap, whose validity for computing standard errors in lasso-type estimation procedures has, up to our knowledge, not yet been studied.

5.7 Determining the hyperparameters

We propose to resort to the standard technique of cross-validation, though it is computationally expensive. For each value on a sufficiently fine grid of (λ_1, λ_2)-values, we compute the cross-validated loss

$$\mathrm{CV}(\lambda_1, \lambda_2) = \sum_{l=1}^{k} \sum_{i:\, (\boldsymbol{x}_i, y_i) \in S_l} L(y_i, f(\boldsymbol{x}_i; \widehat{\beta}_0^{S_{-l}}, \widehat{\boldsymbol{\beta}}^{S_{-l}}))$$

by randomly dividing S into cross-validation folds $S_1, \ldots, S_k$ of roughly equal size, defining $S_{-l} = S \setminus S_l,\ l = 1, \ldots, k$, and denoting by $\widehat{\beta}_0^{S_{-l}}, \widehat{\boldsymbol{\beta}}^{S_{-l}}$ the structured elastic net estimates using the sample S_l. Alternatively, one may compute a model selection criterion such as the GCV, C_p, AIC on the basis of the formula of degrees of freedom given in Section 5.5.

6 Data Analysis

6.1 One-dimensional signal regression

In one-dimensional signal regression, as described, e.g., in Frank and Friedman (1993), one aims at the prediction of a response given a sampled signal $\boldsymbol{x}^\top = (x(t))_{t=1}^T$, where the indices $t = 1, \ldots, T$, refer to different ordered sampling points. For a sample $S = \{(\{x_1(t)\}_{t=1}^T, y_1), \ldots, (\{x_n(t)\}_{t=1}^T, y_n)\}$ of pairs consisting of sampled signals and responses, we consider prediction models of the form

$$\widehat{y}_i = \zeta\left(\widehat{\beta}_0 + \sum_{t=1}^{T} x_i(t)\widehat{\beta}(t)\right), \quad i = 1, \ldots, n.$$

6.1.1 Simulation study

Similarly to Tutz and Gertheiss (2009), we simulate signals $x(t),\ t = 1, \ldots, T,\ \ T = 100$, according to

$$\begin{aligned} \{x(t)\} &\sim \sum_{k=1}^{5} b_k \sin(t\pi(5 - b_k)/50 - m_k) + \tau(t), \\ \{b_k\} &\sim U(0; 5), \quad \{m_k\} \sim U(0; 2\pi), \quad \{\tau(t)\} \sim N(0, 0.25), \end{aligned}$$

with $U(a; b)$ denoting the uniform distribution on the interval $(a; b)$. For the coefficient function $\beta^*(t),\ t = 1, \ldots, T$, we examine two cases. In the first case, referred to as 'bump setting', we use

$$\beta^*(t) = \begin{cases} -\left\{(30 - t)^2 + 100\right\}/200 & t = 21, \ldots, 39, \\ \left\{(70 - t)^2 - 100\right\}/200 & t = 61, \ldots, 80, \\ 0 & \text{otherwise.} \end{cases}$$

In the second case, referred to as 'block setting',

$$\boldsymbol{\beta}^* = (\underbrace{0, \ldots, 0}_{20 \text{ times}}, \underbrace{0.5, \ldots, 0.5}_{10 \text{ times}}, \underbrace{1, \ldots, 1}_{10 \text{ times}}, \underbrace{0.5, \ldots, 0.5}_{10 \text{ times}}, \underbrace{0.25, \ldots, 0.25}_{10 \text{ times}}, \underbrace{0, \ldots, 0}_{40 \text{ times}})^\top.$$

The form of the signals and coefficient functions are displayed in Figure 6.

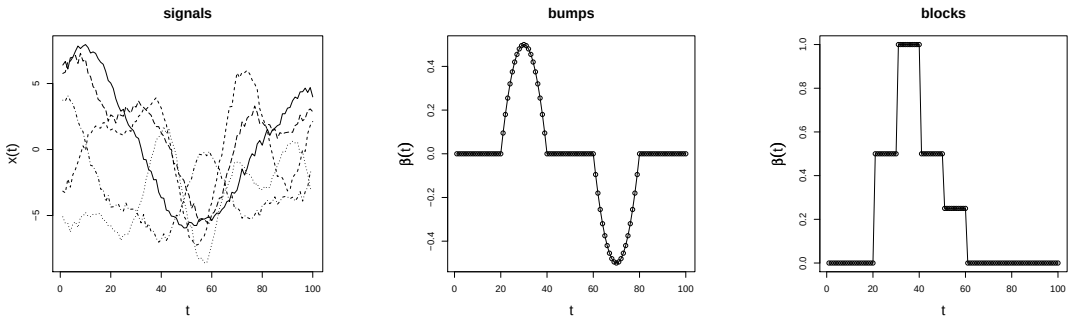


Figure 6: The setting of the simulation study. A collection of five signals (left panel), the coefficient functions for 'bump'- (middle panel) and 'block' setting (right panel), respectively.

For both settings, data are simulated according to

$$y = \sum_{t=1}^{T} x(t)\beta^*(t) + \epsilon, \quad \epsilon \sim N(0, 5),$$

For each out of 50 iterations, we simulate $i = 1, \ldots, 500$ i.i.d. realizations and divide them into three parts: a training set of size 200, a validation set of size 100 and a test set of size 200. Hyperparameters of the methods listed below are optimized by means of the validation set. As performance measures, we compute the absolute distance $L^1(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \left\| \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right\|_1$ of true- and estimated coefficients and the mean squared prediction error on the test set. For methods with built-in feature selection, we additionally evaluate the goodness of selection in terms of sensitivity and specificity, defined by

$$\text{sensitivity}(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) = \frac{|\{t : \ \widehat{\beta}(t) \neq 0\} \cap \{t : \ \beta^*(t) \neq 0\}|}{|\{t : \ \beta^*(t) \neq 0\}|},$$
$$\text{specificity}(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) = \frac{|\{t : \ \widehat{\beta}(t) = 0\} \cap \{t : \ \beta^*(t) = 0\}|}{|\{t : \ \beta^*(t) = 0\}|}.$$

For each of the two setups, the simulation is repeated 50 times. The following methods are compared:

- ridge regression,
- generalized ridge regression with a first difference penalty,
- P-splines according to Eilers and Marx (1999),
- lasso,
- fused lasso,
- elastic net,
- structured elastic net with a first difference penalty,
- adaptive structured elastic net, where the weights $\{\omega(t)\}$ are chosen according to the ridge estimator of the same iteration as $\omega(t) = 1/|\widehat{\beta}^{\text{ridge}}(t)|$.

Performance measures are averaged over 50 iterations and displayed in Table 1 (bump setting) and Table 2 (block setting), respectively.

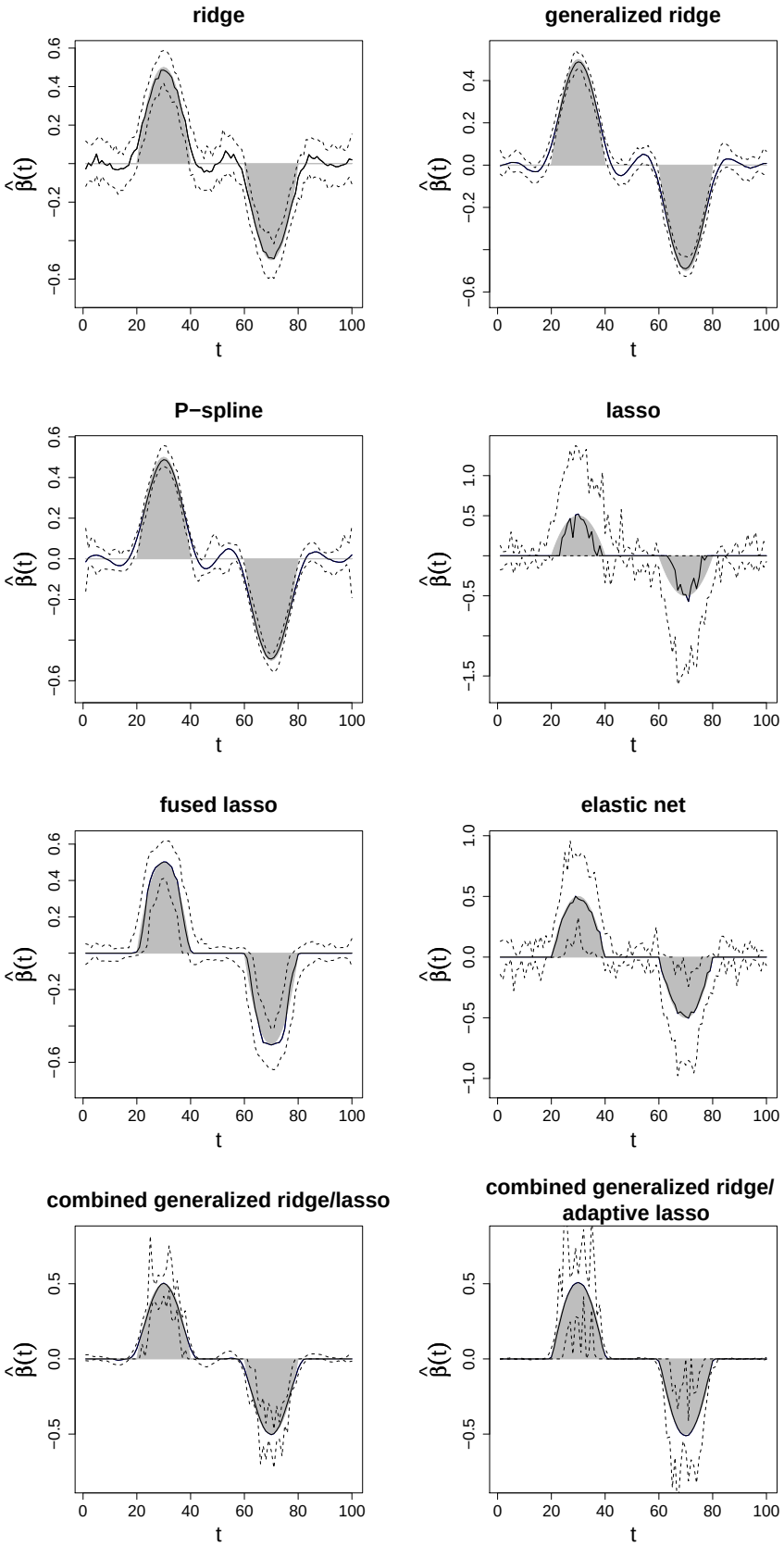


Figure 7: Estimated coefficient functions for the bump setting. The pointwise median curve over 50 iterations is represented by a solid line, pointwise 0.05- and 0.95-quantiles are drawn in dashed lines.

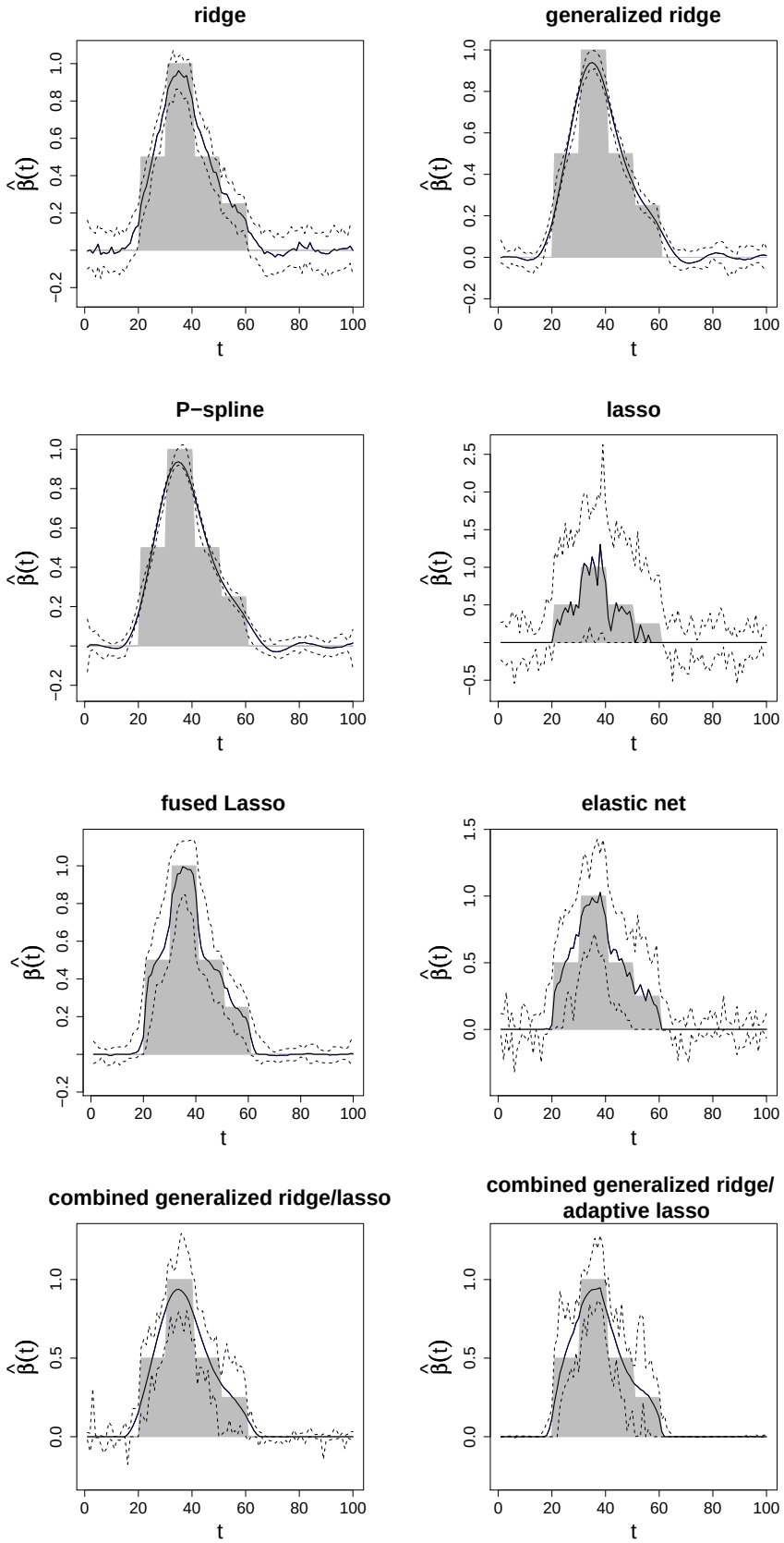


Figure 8: Estimated coefficient functions for the block setting. The pointwise median curve over 50 iterations is represented by a solid line, pointwise 0.05- and 0.95-quantiles are drawn in dashed lines.

method	$L^1(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*)$	PE	sensitivity	specificity
ridge	0.249 (5.9×10^{-4})	5.35 (0.078)		
g.ridge	0.238 (9.9×10^{-4})	5.32 (0.076)		
P-spline	0.241 (16.0×10^{-4})	5.30 (0.077)		
lasso	0.271 (23.9×10^{-4})	5.72 (0.079)	0.62 (8.9×10^{-3})	0.65 (0.016)
fused lasso	0.235 (7.2×10^{-4})	5.30 (0.075)	0.96 (5.5×10^{-3})	0.51 (0.010)
enet	0.246 (29.9×10^{-4})	5.46 (0.081)	0.93 (0.013)	0.69 (0.032)
s.enet	**0.232** (7.6×10^{-4})	5.30 (0.078)	**0.98** (7.8×10^{-3})	0.59 (0.029)
ada.s.enet	**0.232** (15.0×10^{-4})	**5.25** (0.075)	0.91 (21.0×10^{-3})	**0.82** (0.020)

Table 1: Results for the bump setting, averaged over 50 simulations. We make use of the following abbreviations: 'PE' for 'mean squared prediction error', 'g. ridge' for 'generalized ridge', 'enet' for 'elastic net', 's.enet' for 'structured elastic net' and 'ada.s.enet' for 'adaptive structured elastic net'. Standard errors are given in parentheses. For each column, the best performance is emphasized in boldface.

For the bump setting, Figure 7 shows that the at least the crude shape of the coefficient function is estimated by all compared methods in a satisfactory way, except for the lasso. Due to a favourable signal-to-noise ratio, even simplistic approaches such as ridge- or generalized ridge regression show competitive performance with respect to prediction of future observations. In pure numbers, the estimation of $\boldsymbol{\beta}^*$ is satisfactory as well. However, the lack of sparsity results into 'noise fitting' for those parts where $\beta^*(t)$ is zero. For the two settings examined here, the P-spline approach does not improve over generalized ridge regression, because the two coefficient functions are not overly smooth. The elastic net considerably improves over the lasso and visually also over ridge regression, but it lacks smoothness. Its numerical inferiority to ridge regression results from double shrinkage as discussed in Subsection 3.5. The performance of the structured elastic net is not fully satisfactory. In particular, at the changepoints from zero- to nonzero parts, there is a tendency to widen unnecessarily the support of the nonzero sections. This shortcoming is removed by the adaptive structured elastic net, thereby confirming the theoretical result concerning selection consistency. This quality seems to be supported by the eminent performance with respect to sensitivity and specificity. The success of the adaptive strategy is also founded on the good performance of the ridge estimator providing the component-specific weights $\omega(t)$. The block setting is actually tailored to the fused lasso, whose output are piecewise constant coefficient functions. Nevertheless, it is not optimal, as the shrinkage of the ℓ^1-penalty acts on all coefficients, including those different from zero. As a result, the fused lasso is outperformed by the adaptive structured elastic net with respect to prediction, though the structure part is seen to be not fully appropriate in the block setting. As opposed to the bump setting, fitting the block function seems to be much more difficult to accomplish in general.

method	$L^1(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*)$	PE	sensitivity	specificity
ridge	0.082 (3.4×10^{-3})	5.41 (0.080)		
g.ridge	0.064 (1.9×10^{-3})	5.35 (0.078)		
P-spline	0.065 (1.9×10^{-3})	5.34 (0.077)		
lasso	0.207 (3.6×10^{-3})	6.12 (0.089)	0.73 (7.5×10^{-3})	0.62 (0.014)
fused lasso	**0.058** (1.9×10^{-3})	5.34 (0.076)	**0.99** (7×10^{-4})	0.51 (0.009)
enet	0.094 (5.0×10^{-3})	5.47 (0.072)	0.95 (6.4×10^{-3})	0.73 (0.083)
s.enet	0.070 (5.0×10^{-3})	5.38 (0.080)	**0.99** (3.3×10^{-3})	0.60 (0.027)
ada.s.enet	0.061 (3.2×10^{-3})	**5.32** (0.69)	0.97 (8.0×10^{-3})	**0.83** (0.018)

Table 2: Results for the block setting, averaged over 50 simulations. For annotation, see Table 1.

6.1.2 Accelerometer data

The 'Sylvia Lawry Centre for Multiple Sclerosis Research e.V.', Munich, kindly provided us with two accelerometer records of two healthy female persons, aged between 20 and 30. They were equipped with a belt containing an accelerometer integrated into the belt buckle before walking several minutes on a flat surface at a moderate speed. The output are triaxial (vertical, horizontal, lateral) acceleration measurements at roughly 25,000 sampling points per person. Following Daumer et al. (2007) human gait, if defined as temporal evolution of three-dimensional accelerations of the center of mass of the body, is supposed to be a quasi-periodic process. Every period defines one gait cycle / double step, which starts with the heel strike and ends with the heel strike of the same foot. A single step ends with the heel strike of the other foot. Therefore, a double step can be seen as natural unit. As consequence, decomposition of the raw signal into pieces, each representing one double step, is an integral part of data preprocessing, not described in further detail here. Overall, we extract $i = 1, \ldots, n = 406$ double steps, 242 from person B ($y = 0$) and 164 from person A ($y = 1$), ending up with a sample $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where each $\boldsymbol{x}_i = (x_i(t)),\ \ t = 1, \ldots, T = 102$ stores the observed vertical acceleration within double step i, $\ i = 1, \ldots, n$. For simplicity, we neglect the dependence of consecutive double steps within the same person and treat them as independent realizations. Horizontal- and lateral acceleration are not considered, since they do not carry information relevant to our prediction problem. We aim at the prediction of the person (A or B) given a double step pattern, and additionally at the detection of parts of the signal apt for discriminating between the two persons. We randomly divide the complete sample into a learning set of size 300 and a test set of size 106, and subsequently carry out logistic regression on the training set, using the structured elastic net with a squared first difference penalty. Hyperparameters are determined by ten-fold cross-validation, and the resulting logistic regression model is used to obtain predictions for the test set. The fused lasso with the hinge loss of support vector machines is used as competitor. A collection of results is assembled in Figures 9, 10 and Table 3, from which one concludes that classification is an easy task, since (nearly) perfect misclassification rates

on the test set are achieved. Concerning feature selection, the results of the structured elastic net are comparable to those of the fused lasso.

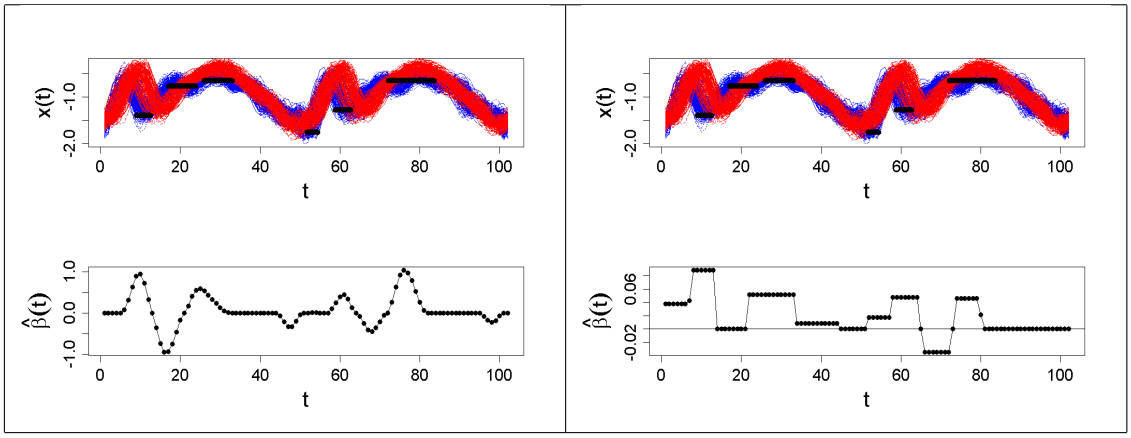


Figure 9: Coefficient functions for structured elastic net-regularized logistic regression (left panel) and the fused lasso support vector machine (right panel). Within each panel, the upper panel displays the overlayed double step patterns of the complete sample (406 double steps). Colours refer to persons, and visual differences between the two classes have been drawn manually as black bars.

fused lasso				
t_1	t_2	test error	# support vectors	# nonzero coefficients
2.5	0.5	0	124	46
structured elastic net				
λ_1	λ_2	test error	degrees of freedom	# nonzero coefficients
1.5	5	1	7.85	61

Table 3: Results of step classification for the fused lasso support vector machine and structured elastic net-regularized logistic regression. For the fused lasso, t_1 denotes the bound imposed on the 1-norm of $\boldsymbol{\beta}$ corresponding to λ_1, and t_2 denotes the bound on the absolute differences $\sum_{t=2}^{T} |\beta(t) - \beta(t-1)|$. For estimating the degrees of freedom, we make use of the heuristic suggested in Section 5.

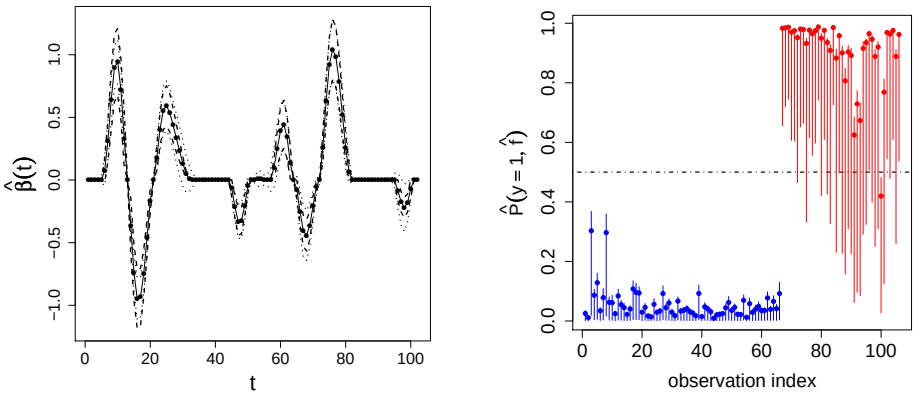


Figure 10: Left panel: the coefficient function of structured elastic net logistic regression (black points) and pointwise ±2 standard error curves, once computed according to heuristic proposed in Section 5 (dashed lines) and once using 100 bootstrap iterations with fixed hyperparameters (dotted lines). Right panel: predicted probabilities for class 1, evaluated on the test set, indicated by filled points. The bars quantify uncertainty: lower- and upper end correspond to 0.1- and 0.9-quantile, respectively, computed from the bootstrap.

6.2 Surface fitting

Figure 11 depicts the surface to be fitted on a 20×20 grid. The surface can be represented by a discrete function $\beta^*(t,u), \quad t,u = 1,\dots,20$. It consists of three non-overlapping truncated Gaussians of different shape and one plateau function. We have

$$\begin{aligned}
\beta^*(t,u) &= B(t,u) + G_1(t,u) + G_2(t,u) + G_3(t,u), \\
B(t,u) &= \frac{1}{2} I(t \in \{10,11,12\}, u \in \{3,4\}), \\
G_1(t,u) &= \max\left\{0, \exp\left(-(t-3 \;\; u-8)\begin{pmatrix} 3 & 0 \\ 0 & 0.25 \end{pmatrix}\begin{pmatrix} t-3 \\ u-8 \end{pmatrix}\right) - 0.2\right\}, \\
G_2(t,u) &= \max\left\{0, \exp\left(-(t-7 \;\; u-17)\begin{pmatrix} 0.75 & 0 \\ 0 & 0.75 \end{pmatrix}\begin{pmatrix} t-7 \\ u-17 \end{pmatrix}\right) - 0.2\right\}, \\
G_3(t,u) &= \max\left\{0, \exp\left(-(t-15 \;\; u-14)\begin{pmatrix} 0.5 & -0.25 \\ -0.25 & 0.5 \end{pmatrix}\begin{pmatrix} t-15 \\ u-14 \end{pmatrix}\right) - 0.2\right\}
\end{aligned} \tag{6.1}$$

Similarly to the simulation study in Subsection 6.1.1, we simulate a noisy version of the surface according to

$$y(t,u) = \beta^*(t,u) + \epsilon(t,u), \quad \{\epsilon(t,u)\} \overset{\text{i.i.d.}}{\sim} N(0, 0.25^2), \quad t,u = 1,\dots,20.$$

For each of the 50 runs, we simulate two instances of $y(t,u)$. The first one is used for training and the second one for hyperparameter tuning. The mean squared error for estimating $\boldsymbol{\beta}^*$ is computed and averaged over 50 runs. Results are summarized in Figure 12 and Table 4. We compare ridge, generalized ridge with a difference penalty according to the grid structure, lasso, fused lasso with a total variation penalty along the grid, structured- and adaptive structured elastic net with the same difference penalty as for generalized ridge. The elastic net coincides - up to a constant scaling factor - with the lasso/soft thresholding in the orthogonal design case and is hence not considered.

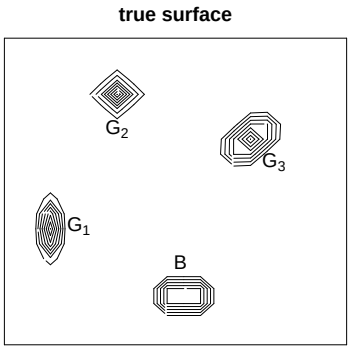


Figure 11: Contours of the surface according to Eq. (6.1).

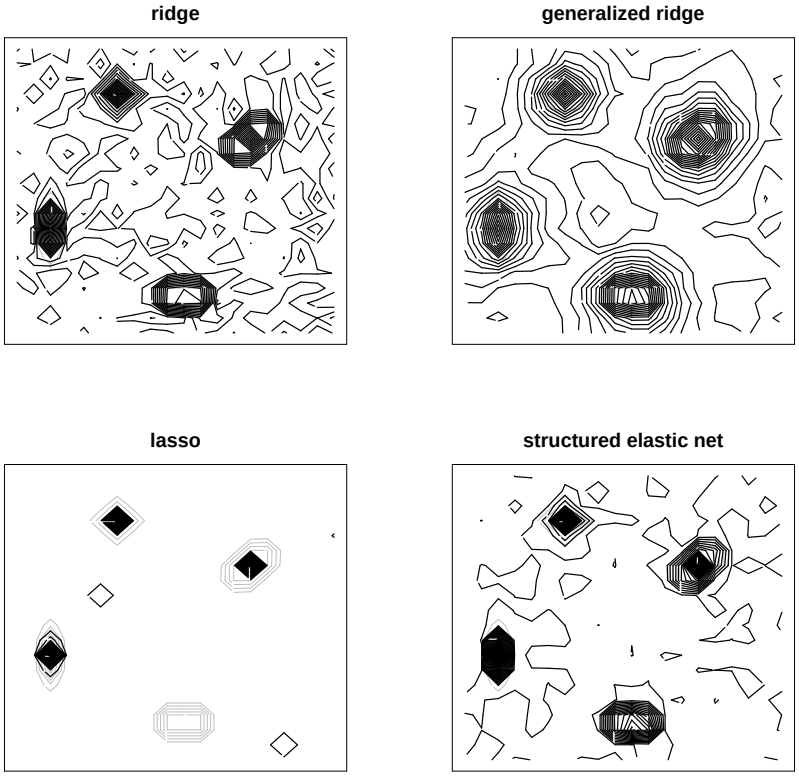


Figure 12: Contours of the estimated surfaces for four selected methods, averaged pointwise over 50 runs.

method	PE	B	G_1	G_2	G_3	zero
ridge	1.20 (0.01)	0.31	0.28	0.20	0.34	0.07
g.ridge	1.17 (0.04)	0.18	0.20	0.14	0.21	0.44
lasso	1.31 (0.01)	0.37	0.32	0.22	0.39	**0.01**
fused lasso	0.67 (0.02)	**0.14**	0.12	**0.08**	**0.15**	0.18
s.enet	0.88 (0.02)	0.22	0.16	0.12	0.23	0.18
ada.s.enet	**0.56** (0.02)	0.15	**0.09**	**0.08**	0.18	0.06

Table 4: Results of the simulation, averaged over 50 iterations (standard errors in parantheses). The columns labeled B, G_1, G_2, G_3 and 'zero' contain the mean prediction error for the corresponding region of the surface. The abbreviations equal those in Table 1. The prediction error has been rescaled by 100.

6.3 Leukaemia cancer data

The dataset of Golub et al. (1999) constitutes one of the milestones in molecular classification of cancer. It consists of gene expression intensities for 7129 genes of 38 leukaemia patients, from which 27 are diagnosed acute lymphoblastic leukaemia (ALL), and the remaining patients are diagnosed acute myeloid leukaemia (AML). In addition, there is an independent test set of 34 samples. A major challenge in cancer research is the detection of pathways governing or influencing the rise and development of cancer. While this rationale is used to motivate the elastic net, it lacks the explicit integration of information on pathways, as available, e.g., in the KEGG database (Kanehisa and Goto (2000)). The latter represents metabolic pathways as graphs in which chemical reactions form the vertex set, and the edges are labeled by proteins/genes taking part in or catalyzing these reactions. Li and Li (2008) incorporate this information by employing what they term 'network-constrained regularization and variable selection', a special case of the structured elastic net, where the matrix $\mathbf{\Lambda}$ is chosen as the *normalized* graph Laplacian (Chung (1997)), a rescaled version of the combinatorial Laplacian introduced in Section 2.2. Our approach for cancer class prediction making use of knowledge about KEGG pathways operates in a similar manner. We are able to match 2761 out of 7129 genes in the KEGG database, i.e. these 2761 genes occur in at least one KEGG pathway. Next, we construct an unweighted association graph by connecting two genes if and only if they share at least one common pathway. The resulting graph consists of eight connected components, from which the largest has a vertex set of size 2688. Following Li and Li (2008), we compute the normalized graph Laplacian, whose rank equals the number of vertices minus the number of connected components. The normalized graph Laplacian constitutes the matrix $\mathbf{\Lambda}$ of structured net-regularized logistic regression model involving all 7129 genes. Since we are given only 38 observations, it follows from the augmented data representation that the structured elastic net degenerates in our situation. Concerning the aim of our analysis, we are primarily interested in the amount the structured elastic net differs from the elastic net in terms of variable selection. We do not observe a notable difference concerning prediction error on the test set, which is not surprising for three reasons. Firstly, prediction is not a hard task for the dataset at hand, because even simple approaches such as the nearest

enet		$\lambda_2 \overset{+}{\rightarrow}$		
# selected	12 (26)	19 (41)	27 (56)	64 (156)
$\|\widehat{\boldsymbol{\beta}}\|_1$	4.49 (9.27)	4.64 (9.29)	4.65 (9.34)	4.66 (9.90)
s.enet				
# selected	19 (44)	23 (73)	27 (91)	36 (135)
$\|\widehat{\boldsymbol{\beta}}\|_1$	0.016 (9.18)	0.004 (9.39)	0.003 (9.60)	0.003 (10.02)

Table 5: The table displays the number of selected pathway genes for the elastic net and the structured elastic net and the 1-norm of the corresponding coefficient subvector for selected values of λ_2, increasing from left to right and λ_1 kept fixed. The numbers in parentheses refer to all genes, including the non-pathway genes.

shrunken centroid classifier (Tibshirani et al. (2003)) yield good results. Secondly, as it becomes clear from Table 5, the genes *not* mapped to any pathway are the relevant ones. Finally, we are sufficiently self-critical to admit that the 38 bits of information contained in the response are probably not enough to judge accurately the influence of pathways. Concerning selection of variables occurring in pathways, we observe striking differences when comparing the elastic net and its structured counterpart. Table 5 shows that for large λ_2, the importance of pathway-based predictors tends to zero for the structured elastic net, while it remains roughly constant for the elastic net. As shown in Figures 13 and 14, the number of pathway genes increases for both methods when increasing λ_2, and the increase is stronger for the elastic net. To judge the importance of pathways, we additionally apply subset selection based on gene set enrichment analysis (GSA, Efron and Tibshirani (2007)), in which the members of the same pathways are treated as one group and the non-pathway genes are treated as groups consisting of one member only. We include a group as a whole in a ridge-regularized logistic regression model if it is given a high score in GSA. The latter is applied repeatedly in ten-fold cross-validation on the training set to select the optimal number of groups. The results of this procedure confirm the observations made for the structured elastic net, because we end up with a competitive prediction model that only includes non-pathway genes.

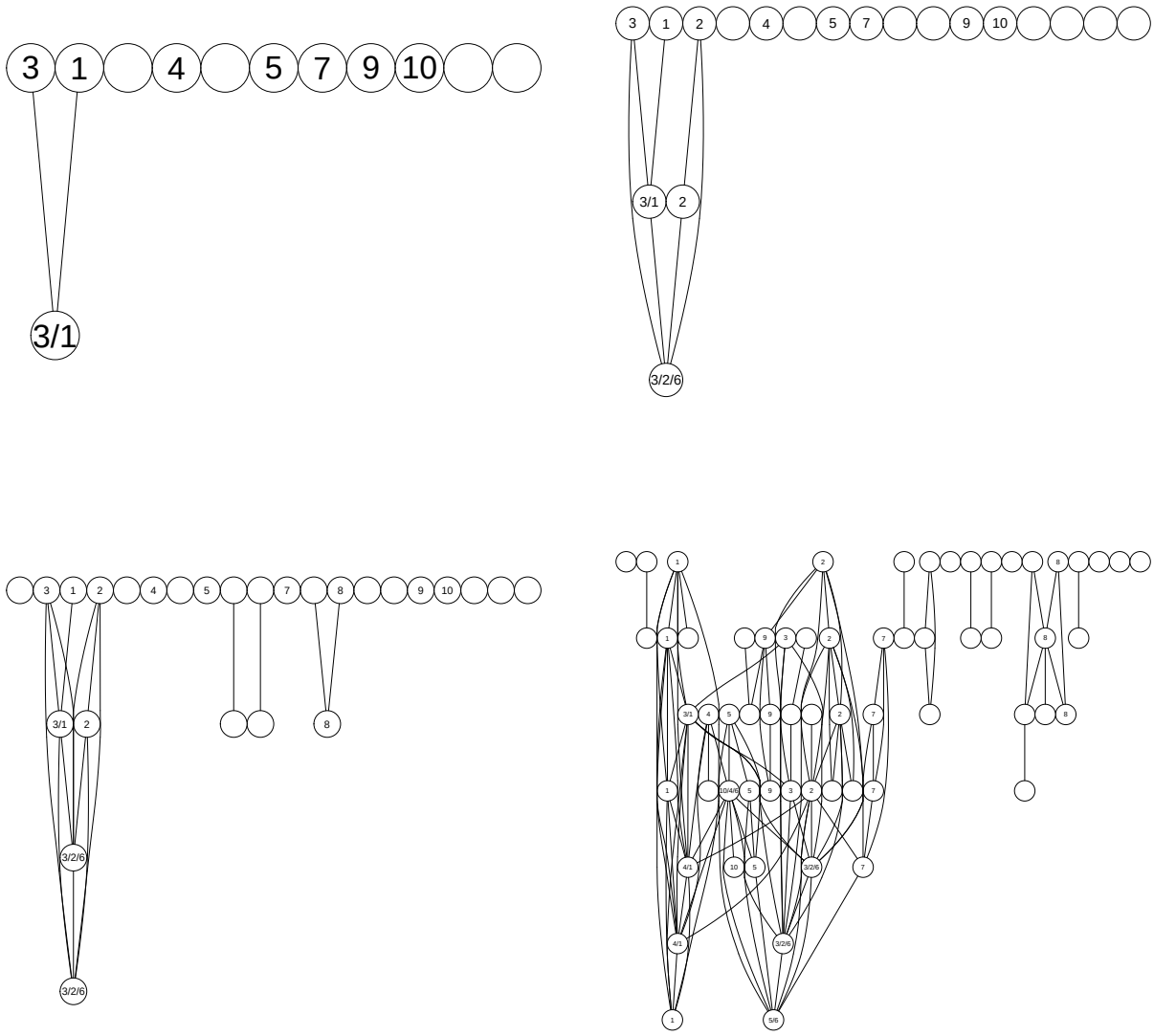


Figure 13: Selected pathway genes of the elastic net in the situation of Table 5, with λ_2 increasing from top to bottom and left to right. The numerals refer to frequently occurring pathways: 1 - 'hematopoietic cell lineage', 2 - 'epithelial cell signaling', 3 - 'cytokine-cytokine receptor interaction', 4 - 'ECM receptor interaction', 5 - 'cell cycle', 6 - 'bladder cancer', 7 - 'arachidonic acid metabolism', 8 - 'Tryptophan metabolism', 9 - 'calcium signaling pathway', 10 - 'focal adhesion'.

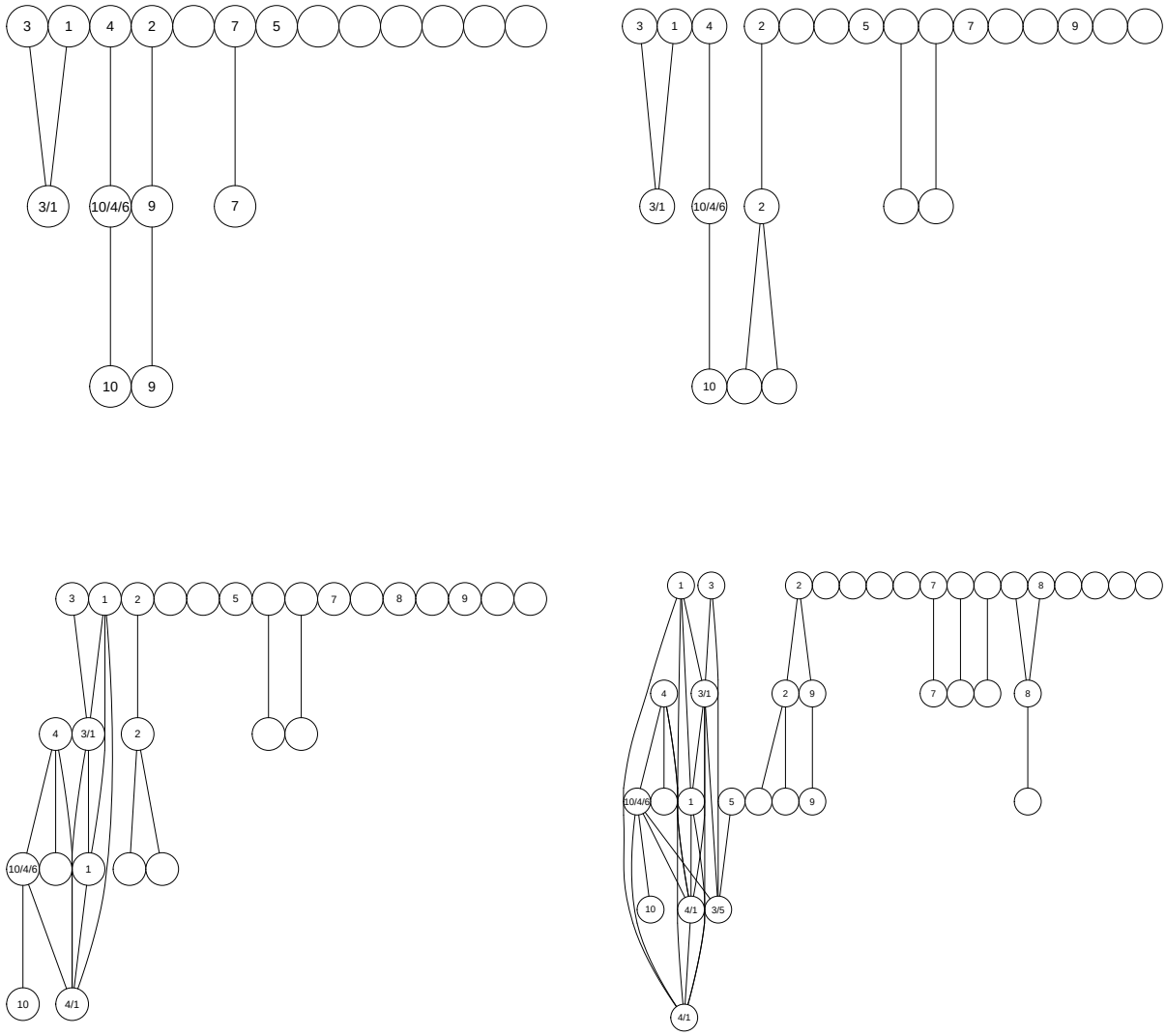


Figure 14: Selected pathway genes of the structured elastic net in the situation of Table 5, with λ_2 increasing from top to bottom and left to right. The numerals refer to frequently occurring pathways: 1 - 'hematopoietic cell lineage', 2 - 'epithelial cell signaling', 3 - 'cytokine-cytokine receptor interaction', 4 - 'ECM receptor interaction', 5 - 'cell cycle', 6 - 'bladder cancer', 7 - 'arachidonic acid metabolism', 8 - 'Tryptophan metabolism', 9 - 'calcium signaling pathway', 10 - 'focal adhesion'.

7 Discussion

The structured elastic net is proposed as a procedure for coefficient selection and -smoothing. We have established a general notion of structured features, for which the structured elastic net is able to take advantage of prior knowledge as opposed to the lasso and the elastic net, which are both purely data-driven. The structured elastic net may also be regarded as a computationally more convenient alternative to the fused lasso. Conceptually, generalizing the fused lasso by computing the total variation of the coefficients along a graph is straightforward. However, due to the non-differentiability of the structure part of the fused lasso, computation may be intractable even for moderately sized graphs.
Turning to the drawbacks of the structured elastic net, we have outlined in Section 5 that model selection and computation of standard errors and in turn the quantification of uncertainty is notoriously difficult. A Bayesian approach promises to be superior in this regard. The lasso can be treated within a Bayesian inference framework (Park and Casella (2008)), while the quadratic part of the structured elastic net regularizer is already motivated from a Bayesian perspective in this paper.
With regard to to possible directions of future research, we consider to study the structured elastic net in combination with other loss functions, e.g. the hinge loss of support vector machines or the check loss for quantile regression. Concerning the application to genomic data, we work on an extension to survival data, in particular to the Cox proportional hazards model. The asymptotic analysis in this paper is basic in the sense that it is bound to strong assumptions, and the role of the structure part of the regularizer and its interplay with the true coefficient vector is not well understood yet, leaving some room for more profound investigations.

Acknowledgments

Martin Slawski was partly supported by the Porticus Foundation for Clinical Medicine and Bioinformatics. We thank the Sylvia Lawry Centre for its support with the accelerometer data example, in particular Martin Daumer for numerous discussions, Christine Gerges and Kathrin Thaler for producing the data. We thank Jelle Goeman for one helpful discussion about his algorithm and making his code publicly available as `R` package. We are grateful to Angelika van der Linde and Daniel Sabanés-Bové for pointing us to several errors and typos in earlier drafts.

A Proofs

A.1 Proof of Proposition 1

Using the well-known expression for the gradient $\frac{\partial L}{\partial \boldsymbol{\beta}}$ in generalized linear models, the Karush-Kuhn-Tucker (KKT) conditions imply that

$$\begin{aligned} -\boldsymbol{X}_1^\top(\boldsymbol{y}-\widehat{\boldsymbol{\mu}}) + \lambda_2(\widehat{\beta}_1 + s\widehat{\beta}_2) + \lambda_1 \operatorname{sign}(\widehat{\beta}_1) &= 0, \\ -\boldsymbol{X}_2^\top(\boldsymbol{y}-\widehat{\boldsymbol{\mu}}) + \lambda_2(\widehat{\beta}_2 + s\widehat{\beta}_1) + \lambda_1 \operatorname{sign}(\widehat{\beta}_2) &= 0, \end{aligned}$$

where $\widehat{\boldsymbol{\mu}} = \widehat{\mathrm{E}}[\boldsymbol{y}|\boldsymbol{X}_1, \boldsymbol{X}_2]$. Adding the second equation to the first equation multiplied by s yields

$$-(s\boldsymbol{X}_1 + \boldsymbol{X}_2)^\top(\boldsymbol{y}-\widehat{\boldsymbol{\mu}}) + 2\lambda_2(\widehat{\beta}_2 + s\widehat{\beta}_1) = 0,$$

implying

$$
\begin{aligned}
|\widehat{\beta}_1 + s\widehat{\beta}_2| &= \frac{1}{2\lambda_2}|(s\boldsymbol{X}_1 + \boldsymbol{X}_2)^\top(\boldsymbol{y} - \widehat{\boldsymbol{\mu}})| \\
&\leq \frac{1}{2\lambda_2}\left\|s\boldsymbol{X}_1 + \boldsymbol{X}_2\right\|\left\|\boldsymbol{y} - \widehat{\boldsymbol{\mu}}\right\| \\
&\leq \frac{1}{2\lambda_2}\sqrt{2(1+s\rho)}\left\|\boldsymbol{y}\right\|,
\end{aligned}
$$

noting that $\|s\boldsymbol{X}_1 + \boldsymbol{X}_2\| = (\|\boldsymbol{X}_1\|^2 + \|\boldsymbol{X}_2\|^2 + 2s\langle \boldsymbol{X}_1, \boldsymbol{X}_2\rangle)^{1/2} = (2(1+s\rho))^{1/2}$, since $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are standardized.

A.2 Proof of Proposition 2

Consider the case $s = -1$ and define another set of coefficients by

$$
\widetilde{\beta}_0 = \widehat{\beta}_0, \quad \widetilde{\beta}_1 = \widetilde{\beta}_2 = \frac{1}{2}(\widehat{\beta}_1 + \widehat{\beta}_2).
$$

First observe that

$$
\begin{aligned}
&\sum_{i=1}^{n} L(y_i, f(\boldsymbol{x}_i; \widetilde{\beta}_0, \widetilde{\boldsymbol{\beta}})) + \frac{\lambda_2}{2}(\widetilde{\beta}_1 - \widetilde{\beta}_2)^2 + \lambda_1 \left\|\widetilde{\boldsymbol{\beta}}\right\|_1 \\
&- \sum_{i=1}^{n} L(y_i, f(\boldsymbol{x}_i; \widehat{\beta}_0, \widehat{\boldsymbol{\beta}})) - \frac{\lambda_2}{2}(\widehat{\beta}_1 - \widehat{\beta}_2)^2 - \lambda_1 \left\|\widehat{\boldsymbol{\beta}}\right\|_1 \geq 0.
\end{aligned} \tag{A.1}
$$

For the part involving the loss function, we obtain

$$
\begin{aligned}
&\sum_{i=1}^{n} \left\{ L(y_i, f(\boldsymbol{x}_i; \widetilde{\beta}_0, \widetilde{\boldsymbol{\beta}})) - L(y_i, f(\boldsymbol{x}_i; \widehat{\beta}_0, \widehat{\boldsymbol{\beta}})) \right\} \\
&\leq \sum_{i=1}^{n} \left| L(y_i, f(\boldsymbol{x}_i; \widetilde{\beta}_0, \widetilde{\boldsymbol{\beta}})) - L(y_i, f(\boldsymbol{x}_i; \widehat{\beta}_0, \widehat{\boldsymbol{\beta}})) \right| \\
&\leq C \sum_{i=1}^{n} |\boldsymbol{x}_i^\top(\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})| \\
&= C \sum_{i=1}^{n} \left| \frac{1}{2}(x_{i1} - x_{i2})(\widehat{\beta}_2 - \widehat{\beta}_1) \right| \\
&= \frac{C}{2}|\widehat{\beta}_2 - \widehat{\beta}_1| \left\|\boldsymbol{X}_1 - \boldsymbol{X}_2\right\|_1.
\end{aligned} \tag{A.2}
$$

For the ℓ^1-regularizer, we have

$$
\left\|\widetilde{\boldsymbol{\beta}}\right\|_1 - \left\|\widehat{\boldsymbol{\beta}}\right\|_1 = |\widehat{\beta}_1 + \widehat{\beta}_2| - |\widehat{\beta}_1| + |\widehat{\beta}_2| \leq 0. \tag{A.3}
$$

For the quadratic regularizer one obtains the difference $\frac{1}{2}(\widehat{\beta}_1 - \widehat{\beta}_2)^2$. Combining this with (A.1) - (A.3), we obtain

$$
\frac{C}{2}|\widehat{\beta}_1 - \widehat{\beta}_2| \left\|\boldsymbol{X}_1 - \boldsymbol{X}_2\right\|_1 - \frac{\lambda_2}{2}(\widehat{\beta}_1 - \widehat{\beta}_2)^2 \geq 0,
$$

from which it follows that

$$
\begin{aligned}
|\widehat{\beta}_1 - \widehat{\beta}_2| &\leq \frac{C}{\lambda_2} \|\boldsymbol{X}_1 - \boldsymbol{X}_2\|_1 \\
&\leq \frac{C}{\lambda_2} \sqrt{n} \|\boldsymbol{X}_1 - \boldsymbol{X}_2\| \\
&\leq \frac{C}{\lambda_2} \sqrt{n 2(1-\rho)},
\end{aligned}
$$

as shown in Proposition 1. The case $s = 1$ is obtained analogously by setting $\widetilde{\beta}_0 = \widehat{\beta}_0$, $\widetilde{\beta}_1 = \frac{1}{2}(\widehat{\beta}_1 - \widehat{\beta}_2)$, $\widetilde{\beta}_2 = \frac{1}{2}(\widehat{\beta}_2 - \widehat{\beta}_1)$.

A.3 Auxiliary results

The asymptotic analysis relies on a more general theory of constrained M-estimation involving the notion of pointwise convergence in distribution of convex functions.

Definition A. 1. *Let G_n be a sequence of lower semicontinuous convex random functions from $\mathbb{R}^p$ to $\mathbb{R} \cup \{\infty\}$, let G be another such function and let $\mathcal{D}$ be a countable dense set in $\mathbb{R}^p$. Then G_n converges pointwise in distribution to G, in signs $G_n \xrightarrow{D} G$, if for each finite subset $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k\} \subset \mathcal{D}$, $(G_n(\boldsymbol{u}_1), \ldots, G_n(\boldsymbol{u}_k))^\top \xrightarrow{D} (G(\boldsymbol{u}_1), \ldots, G(\boldsymbol{u}_k))^\top$.*

Theorem A. 1. *(Geyer (1996))*
Let G_n, G be as in Definition A. 1 and let $G_n \xrightarrow{D} G$. Define $\widehat{\boldsymbol{u}}_n = \operatorname{argmin} G_n$ and $\widehat{\boldsymbol{u}} = \operatorname{argmin} G$. If G has a unique minimizer , then $\widehat{\boldsymbol{u}}_n \xrightarrow{D} \widehat{\boldsymbol{u}}$.

By means of this preparation, we can prove Theorems 1-4.

A.4 Proof of Theorem 1

Define the random function $V_n(\boldsymbol{u})$ by

$$
\begin{aligned}
V_n(\boldsymbol{u}) = &\sum_{i=1}^{n} (\epsilon_i - \boldsymbol{u}^\top \boldsymbol{x}_i/\sqrt{n})^2 - \epsilon_i^2 \\
&+ \lambda_1^n \sum_{j=1}^{p} |\beta_j^* + u_j/\sqrt{n}| - |\beta_j^*| \\
&+ \lambda_2^n [(\boldsymbol{\beta}^* + \boldsymbol{u}/\sqrt{n})^\top \boldsymbol{\Lambda} (\boldsymbol{\beta}^* + \boldsymbol{u}/\sqrt{n}) - \boldsymbol{\beta}^{*\top} \boldsymbol{\Lambda} \boldsymbol{\beta}^*].
\end{aligned}
$$

Observe that $V_n(\boldsymbol{u})$ is minimized at $\widehat{\boldsymbol{u}}_n = \sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*)$, because with $\boldsymbol{u} = \sqrt{n}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$,

$$
\widehat{\boldsymbol{u}}_n = \operatorname*{argmin}_{\boldsymbol{u}} V_n(\boldsymbol{u}) \Leftrightarrow \widehat{\boldsymbol{\beta}}_n = \operatorname*{argmin}_{\boldsymbol{\beta}} (Q_n(\boldsymbol{\beta}) - Q_n(\boldsymbol{\beta}^*)),
$$

with

$$
Q_n(\boldsymbol{\beta}) = \|\boldsymbol{y}_n - \boldsymbol{X}_n \boldsymbol{\beta}\|^2 + \lambda_1^n \|\boldsymbol{\beta}\|_1 + \lambda_2^n \boldsymbol{\beta}^\top \boldsymbol{\Lambda} \boldsymbol{\beta}. \tag{A.4}
$$

Considering the first term of $V_n(\boldsymbol{u})$, we have

$$
\begin{aligned}
\sum_{i=1}^{n} (\epsilon_i - \boldsymbol{u}^\top \boldsymbol{x}_i/\sqrt{n})^2 - \epsilon_i^2 &= \frac{\boldsymbol{u}^\top \boldsymbol{X}_n^\top \boldsymbol{X}_n \boldsymbol{u}}{n} - 2\boldsymbol{u}^\top \sqrt{n} \frac{\boldsymbol{X}_n^\top (\boldsymbol{y}_n - \boldsymbol{X}_n \boldsymbol{\beta}^*)}{n} \\
&= \frac{\boldsymbol{u}^\top \boldsymbol{X}_n^\top \boldsymbol{X}_n \boldsymbol{u}}{n} - 2\boldsymbol{u}^\top \frac{\boldsymbol{X}_n^\top \boldsymbol{X}_n}{n} \sqrt{n} \left(\left(\frac{\boldsymbol{X}_n^\top \boldsymbol{X}_n}{n} \right)^{-1} \frac{\boldsymbol{X}_n^\top \boldsymbol{y}_n}{n} - \boldsymbol{\beta}^* \right).
\end{aligned}
$$

The first term on the r.h.s. converges to $\boldsymbol{u}^\top \boldsymbol{C}\boldsymbol{u}$ by condition (C.2). The asymptotic result for the ols estimator is

$$\sqrt{n}\left(\left(\frac{\boldsymbol{X}_n^\top \boldsymbol{X}_n}{n}\right)^{-1}\frac{\boldsymbol{X}_n^\top \boldsymbol{y}_n}{n} - \boldsymbol{\beta}^*\right) \xrightarrow{\mathrm{D}} N(\boldsymbol{0}, \sigma^2 \boldsymbol{C}^{-1}),$$

hence the second term of the previous expression converges to $\boldsymbol{w}$ in distribution. Invoking Slutsky's theorem, the first term of V_n converges to $-2\boldsymbol{u}^\top \boldsymbol{w} + \boldsymbol{u}^\top \boldsymbol{C}\boldsymbol{u}$, again in distribution. For the first penalty term in V_n, one has that

$$\begin{aligned}\lambda_1^n \sum_{j=1}^{p} |\beta_j^* + u_j/\sqrt{n}| - |\beta_j^*| = \frac{\lambda_1^n}{\sqrt{n}} \sum_{j=1}^{p} \{\mathrm{sign}(\beta_j^* + u_j/\sqrt{n})(\sqrt{n}\beta_j^* + u_j) - \mathrm{sign}(\beta_j^*)\ \sqrt{n}\beta_j^*\} I(\beta_j^* \neq 0) \\ + \frac{\lambda_1^n}{\sqrt{n}} \sum_{j=1}^{p} |u_j| I(\beta_j^* = 0).\end{aligned}$$

Since $\mathrm{sign}(\beta_j^* + u/\sqrt{n}) \to \mathrm{sign}(\beta_j^*)$ and $\lambda_1^n/\sqrt{n} \to \lambda_1^0$ as $n \to \infty$, the expression converges to the second line in the definition of V. Finally,

$$\lim_{n \to \infty} \lambda_2^n \{(\boldsymbol{\beta}^* + \boldsymbol{u}/\sqrt{n})^\top \boldsymbol{\Lambda}(\boldsymbol{\beta}^* + \boldsymbol{u}/\sqrt{n}) - \boldsymbol{\beta}^{*\top} \boldsymbol{\Lambda}\boldsymbol{\beta}^*\} = 2\lambda_2^0 \boldsymbol{u}^\top \boldsymbol{\Lambda}\boldsymbol{\beta}^*.$$

Applying Slutsky's theorem once again, it holds that $V_n \xrightarrow{\mathrm{D}} V$ in the sense of Definition A. 1. Since V_n is convex and V has a unique minimizer, we conclude from Theorem A. 1 that

$$\operatorname*{argmin}_{\boldsymbol{u}} V_n(\boldsymbol{u}) = \widehat{\boldsymbol{u}}_n \xrightarrow{\mathrm{D}} \operatorname{argmin} V(\boldsymbol{u}) = \sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*).$$

A.5 Proof of Theorem 2

Define $W_n(\boldsymbol{u})$ by

$$W_n(\boldsymbol{u}) = L_n(\boldsymbol{u}) + \lambda_1^n \sum_{j=1}^{p} |\beta_j^* + u_j/\sqrt{n}| - |\beta_j^*| + \lambda_2^n [(\boldsymbol{\beta}^* + \boldsymbol{u}/\sqrt{n})^\top \boldsymbol{\Lambda}(\boldsymbol{\beta}^* + \boldsymbol{u}/\sqrt{n}) - \boldsymbol{\beta}^{*\top} \boldsymbol{\Lambda}\boldsymbol{\beta}^*],$$

$$L_n(\boldsymbol{u}) = 2\phi^{-1} \sum_{i=1}^{n} b(\boldsymbol{x}_i^\top (\boldsymbol{u}/\sqrt{n} + \boldsymbol{\beta}^*)) - b(\boldsymbol{x}_i^\top \boldsymbol{\beta}^*) - y_i \boldsymbol{x}_i^\top \boldsymbol{u}/\sqrt{n}.$$

Observe that $\operatorname{argmin} W(\boldsymbol{u}) = \widehat{\boldsymbol{u}}_n = \sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*)$ with $\widehat{\boldsymbol{\beta}}_n$ as in Theorem 2. We have the following second-order Taylor expansion of $L_n(\boldsymbol{u})$ around $\boldsymbol{u} = \boldsymbol{0}$:

$$\begin{aligned}L_n(\boldsymbol{u}) = 2\phi^{-1} \sum_{i=1}^{n} (y_i - b'(\boldsymbol{x}_i^\top \boldsymbol{\beta}^*)) \frac{\boldsymbol{u}^\top \boldsymbol{x}_i}{\sqrt{n}} \\ + \phi^{-1} \sum_{i=1}^{n} b''(\boldsymbol{x}_i^\top \boldsymbol{\beta}^*) \frac{\boldsymbol{u}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{u}}{n} \\ + R_n(\boldsymbol{u}),\end{aligned}$$

where the remainder is given by

$$R_n(\boldsymbol{u}) = \frac{1}{3n^{3/2}} \phi^{-1} \sum_{i=1}^{n} b'''(\boldsymbol{x}_i^\top \boldsymbol{\xi})(\boldsymbol{x}_i^\top \boldsymbol{u})^3,$$

where $\boldsymbol{\xi}$ is contained in the segment from $\boldsymbol{\beta}^*$ to $\boldsymbol{\beta}^* + \boldsymbol{u}/\sqrt{n}$. Considering the first term of $L_n(\boldsymbol{u})$, standard properties of generalized linear models can be applied (McCullagh and Nelder (1989)):

$$\phi^{-1}\boldsymbol{u}^\top \mathrm{E}[\boldsymbol{x}_i(y_i - b'(\boldsymbol{x}_i^\top\boldsymbol{\beta}^*))] = 0,$$

$$\mathrm{var}[\phi^{-1}(y_i - b'(\boldsymbol{x}_i^\top\boldsymbol{\beta}^*))] = \boldsymbol{u}^\top \frac{\mathrm{E}[\phi^{-1}b''(\boldsymbol{x}_i^\top\boldsymbol{\beta}^*)\boldsymbol{x}_i\boldsymbol{x}_i^\top]}{n}\boldsymbol{u} = \frac{\boldsymbol{u}^\top\mathcal{I}\boldsymbol{u}}{n}.$$

Application of the central limit theorem yields that

$$\phi^{-1}\,\frac{\boldsymbol{u}^\top}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{x}_i(b'(\boldsymbol{x}_i^\top\boldsymbol{\beta}^*) - y_i) \xrightarrow{\mathrm{D}} \boldsymbol{u}^\top\boldsymbol{w}, \quad \boldsymbol{w} \sim N(\boldsymbol{0}, \mathcal{I}).$$

For the second term of $L_n(\boldsymbol{u})$, note that

$$\phi^{-1}\sum_{i=1}^n b''(\boldsymbol{x}_i^\top\boldsymbol{\beta}^*)\frac{\boldsymbol{u}^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{u}}{n} \to \mathcal{I}.$$

Turning to the remainder,

$$3n^{1/2}\phi R_n(\boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^n b'''(\boldsymbol{x}_i^\top\boldsymbol{\xi}) \le \frac{1}{n}\sum_{i=1}^n M(\boldsymbol{x}_i)(\boldsymbol{x}_i^\top\boldsymbol{u})^3 \to \mathrm{E}[M(\mathbb{X})|\boldsymbol{u}^\top\mathbb{X}|^3] < \infty$$

by condition (G.2), concluding that $R_n(\boldsymbol{u}) = O_{\mathrm{P}}(n^{-1/2})$. The limiting behaviour of the regularizer in $W_n(\boldsymbol{u})$ has already been studied in the proof of Theorem 1. As for the latter, Slutsky's theorem and Theorem A. 1 imply that $\operatorname{argmin} W_n(\boldsymbol{u}) \xrightarrow{\mathrm{D}} \operatorname{argmin} W(\boldsymbol{u})$.

A.6 Proof of Theorem 3

Define

$$\begin{aligned}\Psi_n(\boldsymbol{u}) = &\sum_{i=1}^n \left(\epsilon_i - \frac{\lambda_1^n}{n}\boldsymbol{x}_i^\top\boldsymbol{u}\right)^2 - \epsilon_i^2 \\ &+ \lambda_1^n\sum_{j=1}^p \left|\beta_j^* + \frac{\lambda_1^n}{n}\right| - |\beta_j^*| \\ &+ \lambda_2^n\left\{\left(\boldsymbol{\beta}^* + \frac{\lambda_1^n}{n}\boldsymbol{u}\right)^\top\boldsymbol{\Lambda}\left(\boldsymbol{\beta}^* + \frac{\lambda_1^n}{n}\boldsymbol{u}\right) - \boldsymbol{\beta}^{*\top}\boldsymbol{\Lambda}\boldsymbol{\beta}^*\right\},\end{aligned}$$

and $\Xi_n(\boldsymbol{u}) = \Psi_n(\boldsymbol{u}) \cdot (\lambda_1^n)^2/n$. If $\widehat{\boldsymbol{\beta}}_n$ denotes the minimizer of $Q_n(\boldsymbol{\beta})$ in Eq. (A.4), then $\widehat{\boldsymbol{u}}_n = \operatorname{argmin}\Psi_n(\boldsymbol{u}) = \operatorname{argmin}\Xi_n(\boldsymbol{u}) = n/\lambda_1^n\ (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*)$. Next, we consider the termwise limits within $\Xi_n(\boldsymbol{u})$. We have

$$\begin{aligned}\Xi_n(\boldsymbol{u}) = &\frac{1}{n}\boldsymbol{u}^\top\boldsymbol{X}_n^\top\boldsymbol{X}_n\boldsymbol{u} - 2\frac{\boldsymbol{\epsilon}_n^\top\boldsymbol{X}_n}{\lambda_1^n} \\ &+ \sum_{j=1}^p \frac{n}{\lambda_1^n}\left(\left|\beta_j^* + \frac{\lambda_1^n}{n}\right| - |\beta_j^*|\right) \\ &+ \frac{n\lambda_2^n}{(\lambda_1^n)^2}\left\{\left(\boldsymbol{\beta}^* + \frac{\lambda_1^n}{n}\boldsymbol{u}\right)^\top\boldsymbol{\Lambda}\left(\boldsymbol{\beta}^* + \frac{\lambda_1^n}{n}\boldsymbol{u}\right) - \boldsymbol{\beta}^{*\top}\boldsymbol{\Lambda}\boldsymbol{\beta}^*\right\}.\end{aligned}$$

The first term converges to $\boldsymbol{u}^\top \boldsymbol{C}\boldsymbol{u}$. For the second term, we have

$$\frac{\boldsymbol{\epsilon}_n^\top \boldsymbol{X}_n}{\lambda_1^n} = \underbrace{\sqrt{n}\left(\frac{\boldsymbol{\epsilon}_n^\top \boldsymbol{X}_n}{n}\right)}_{=O_{\mathrm{P}}(1)} \frac{\sqrt{n}}{\lambda_1^n} = o_{\mathrm{P}}(1).$$

The third terms converges to

$$P(\boldsymbol{u}) = \sum_{j=1}^{p} \operatorname{sign}(\beta_j^*) I(\beta_j^* \neq 0) + |u_j| I(\beta_j^* = 0).$$

For the last term, we have the limit $2R\boldsymbol{u}^\top \boldsymbol{\Lambda}\boldsymbol{\beta}^*$. Defining $\Xi(\boldsymbol{u}) = \boldsymbol{u}^\top \boldsymbol{C}\boldsymbol{u} + P(\boldsymbol{u}) + 2R\boldsymbol{u}^\top \boldsymbol{\Lambda}\boldsymbol{\beta}^*$, $\Xi_n \xrightarrow{\mathrm{D}} \Xi$ and by Theorem A. 1, $\widehat{\boldsymbol{u}}_n \xrightarrow{\mathrm{P}} \operatorname{argmin} \Xi = \widehat{\boldsymbol{u}}$.
Since we have claimed selection consistency,

$$\mathrm{P}(\widehat{\boldsymbol{\beta}}_{j,n} = 0) \to 1 \text{ for all } j \in A^c,$$

from which it follows that $\widehat{u}_j = 0 \ \forall j \in A^c$. On the other hand, using the partitioning scheme of Eq. (4.3), $\widehat{\boldsymbol{u}}_A$ satisfies the equation

$$2\boldsymbol{C}_A \widehat{\boldsymbol{u}}_A + 2\boldsymbol{C}_{AA^c}\boldsymbol{u}_{A^c} + \boldsymbol{s}_A + 2R\boldsymbol{\Lambda}_A \boldsymbol{\beta}_A^* = \boldsymbol{0}, \quad \boldsymbol{s}_A = (\operatorname{sign}(\beta_j^*), \ \ j \in A)^\top.$$

$$\Rightarrow \ \widehat{\boldsymbol{u}}_A = -\boldsymbol{C}_A^{-1}\left(\boldsymbol{C}_{AA^c}\boldsymbol{u}_{A^c} + \frac{\boldsymbol{s}_A}{2} + R\boldsymbol{\Lambda}_A \boldsymbol{\beta}_A^*\right).$$

Partial optimization of Ξ w.r.t. to $\boldsymbol{u}_{A^c}$ amounts to the minimization of the following expression:

$$\boldsymbol{u}_{A^c}^\top \boldsymbol{C}_{A^c}\boldsymbol{u}_{A^c} + 2\boldsymbol{u}_{A^c}^\top \boldsymbol{C}_{A^c A}\widehat{\boldsymbol{u}}_A + \|\boldsymbol{u}_{A^c}\|_1 + 2R\boldsymbol{u}_{A^c}^\top \boldsymbol{\Lambda}_{A^c A}\boldsymbol{\beta}_A^*.$$

Knowing that $\widehat{\boldsymbol{u}}_{A^c} = \boldsymbol{0}$ and plugging in the expression for $\widehat{\boldsymbol{u}}_A$, the KKT conditions imply that

$$\left|-\boldsymbol{C}_{A^c A}\boldsymbol{C}_A^{-1}(\boldsymbol{s}_A + 2R\boldsymbol{\Lambda}_A\boldsymbol{\beta}_A^*) + 2R\boldsymbol{\Lambda}_{A^c A}\boldsymbol{\beta}_A^*\right| \leq \boldsymbol{1}.$$

A.7 Proof of Theorem 4

Define

$$\begin{aligned} Z_n(\boldsymbol{u}) = &\sum_{i=1}^{n}(\epsilon_i - \boldsymbol{u}^\top \boldsymbol{x}_i/\sqrt{n})^2 - \epsilon_i^2 \\ &+ \lambda_1^n \sum_{j=1}^{p} \omega_j(|\beta_j^* + u_j/\sqrt{n}| - |\beta_j^*|) \\ &+ \lambda_2^n[(\boldsymbol{\beta}^* + \boldsymbol{u}/\sqrt{n})^\top \boldsymbol{\Lambda}(\boldsymbol{\beta}^* + \boldsymbol{u}/\sqrt{n}) - \boldsymbol{\beta}^{*\top}\boldsymbol{\Lambda}\boldsymbol{\beta}^*], \end{aligned}$$

which is minimized at $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n^{\text{adaptive}} - \boldsymbol{\beta}^*)$. From the proof of Theorem 1, we know that the first line in Z_n converges in distribution to $-2\boldsymbol{u}^\top \boldsymbol{w} + \boldsymbol{u}^\top \boldsymbol{C}\boldsymbol{u}$, $\boldsymbol{w} \sim N(0, \sigma^2 \boldsymbol{C})$. For the second line, one has to distinguish two cases.
Case 1: $\beta_j^* \neq 0$. Then $(|\beta_j^* + u_j/\sqrt{n}| - |\beta_j^*|) \to u_j \operatorname{sign}(\beta_j^*)$. Moreover, from the definition of ω_j, the assumptions made for $\widehat{\beta}_j^{\text{init}}$ and the continuous mapping theorem, we have $\omega_j \xrightarrow{\mathrm{P}} |\beta_j^*|^{-\gamma}$. Since $\lambda_1^n/\sqrt{n} \to 0$ by assumption, the whole term vanishes.
Case 2: $\beta_j^* = 0$. Then $\sqrt{n}(|\beta_j^* + u_j/\sqrt{n}| - |\beta_j^*|) \to |u_j|$, $n^{-1/2}\lambda_1^n \omega_j = n^{-1/2}\lambda_1^n r_n^\gamma |r_n \widehat{\beta}_j^{\text{init}}|^{-\gamma} \to$

∞, noting that $|r_n\widehat{\beta}_j^{\text{init}}| = O_{\text{P}}(1)$ by assumption. Overall, if $u_j \neq 0$, the whole term tends to infinity as $n \to \infty$. For the third line in $Z_n(\boldsymbol{u})$, one obtains the limit $2\lambda_2^0 \boldsymbol{u}^\top \boldsymbol{\Lambda}\boldsymbol{\beta}^*$. Putting all together, we have for all $\boldsymbol{u}$ that

$$Z_n(\boldsymbol{u}) \xrightarrow{\text{D}} Z(\boldsymbol{u}) = \begin{cases} -2\boldsymbol{u}_A^\top \boldsymbol{w}_A + \boldsymbol{u}_A^\top \boldsymbol{C}_A \boldsymbol{u}_A + 2\lambda_2^0 \boldsymbol{u}_A^\top \boldsymbol{\Lambda}_A \boldsymbol{\beta}_A^* & \text{if } \boldsymbol{u}_{A^c} = \boldsymbol{0}, \\ \infty, & \text{otherwise,} \end{cases}$$

and that $\operatorname{argmin} Z_n \xrightarrow{\text{D}} \operatorname{argmin} Z = \widehat{\boldsymbol{u}}$ by Theorem A. 1. We have $\widehat{\boldsymbol{u}}_{A^c} = \boldsymbol{0}$, and, by differentiation

$$\widehat{\boldsymbol{u}}_A = \boldsymbol{C}_A^{-1}(\boldsymbol{w}_A - \lambda_2^0 \boldsymbol{\Lambda}_A \boldsymbol{\beta}_A^*), \quad \boldsymbol{w}_A \sim N(0, \sigma^2 \boldsymbol{C}_A),$$

i.e.

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{A,n}^{\text{adaptive}} - \boldsymbol{\beta}_A^*) \xrightarrow{\text{D}} N(-\lambda_2^0 \boldsymbol{C}_A^{-1} \boldsymbol{\Lambda}_A \boldsymbol{\beta}_A^*, \boldsymbol{C}_A^{-1}),$$

and consequently

$$\lim_{n \to \infty} \text{P}(\exists j \in A : \ \widehat{\beta}_{j,n}^{\text{adaptive}} = 0) = 0.$$

On the other hand, take $j \in A^c$ and assume that $\lim_{n \to \infty} \widehat{\beta}_{j,n}^{\text{adaptive}} \neq 0$. Then one can differentiate the adaptive structured elastic net criterion w.r.t. β_j, yielding the equation

$$2\boldsymbol{X}_{j,n}^\top (\boldsymbol{y}_n - \boldsymbol{X}_n \widehat{\boldsymbol{\beta}}_n^{\text{adaptive}}) - 2\lambda_2^n l_{jj} \widehat{\beta}_{j,n}^{\text{adaptive}} - 2\lambda_2^n \sum_{\substack{r \\ r \neq j}} l_{jr} \widehat{\beta}_{r,n}^{\text{adaptive}} = \lambda_1^n \operatorname{sign}(\widehat{\beta}_{j,n}^{\text{adaptive}}) \omega_j.$$

Dividing both sides of the previous equation by $\sqrt{n}$, the left hand side is $O_{\text{P}}(1)$, while the expression on the right hand side tends to infinity. Hence, the probability that the equation is fulfilled tends to zero, which contradicts the assumption that $\widehat{\beta}_{j,n}^{\text{adaptive}} \neq 0$.

B Analytical solution for two predictors

First, assume that $\widehat{\beta}_1, \widehat{\beta}_2 \neq 0$. From the KKT conditions, we have

$$\widehat{\beta}_1 = \frac{\overbrace{\boldsymbol{X}_1^\top (\boldsymbol{y} - \boldsymbol{X}_2 \widehat{\beta}_2) + \lambda_2 l_{12} \widehat{\beta}_2}^{\widetilde{\beta}_1} - \frac{\lambda_1}{2} \operatorname{sign}(\widetilde{\beta}_1)}{1 + \lambda_2 l_{11}},$$

$$\widehat{\beta}_2 = \frac{\overbrace{\boldsymbol{X}_2^\top (\boldsymbol{y} - \boldsymbol{X}_1 \widehat{\beta}_1) + \lambda_2 l_{12} \widehat{\beta}_1}^{\widetilde{\beta}_2} - \frac{\lambda_1}{2} \operatorname{sign}(\widetilde{\beta}_2)}{1 + \lambda_2 l_{22}},$$

i.e.

$$\begin{aligned} \widehat{\beta}_1 &= (1 + \lambda_2 l_{11})^{-1} \operatorname{sign}(\widetilde{\beta}_1)(|\widetilde{\beta}_1| - \lambda_1/2) \\ &= (1 + \lambda_2 l_{11})^{-1} \operatorname{sign}(\widetilde{\beta}_1) \left(\frac{\widetilde{\beta}_1| - |\widetilde{\beta}_2|}{2} - \gamma \right), \quad \gamma = \frac{|\widetilde{\beta}_1| + |\widetilde{\beta}_2| - \lambda_1}{2} > 0. \end{aligned}$$

Likewise, we obtain

$$\widetilde{\beta}_2 = (1 + \lambda_2 l_{22})^{-1} \operatorname{sign}(\widetilde{\beta}_2) \left(\frac{\widetilde{\beta}_2| - |\widetilde{\beta}_1|}{2} - \gamma \right).$$

Otherwise, assume that $\widetilde{\beta}_1 - \lambda_1/2 > 0$ and that

$$\left[\frac{|\widetilde{\beta}_2| - |\widetilde{\beta}_1|}{2} + \gamma\right]_+ = 0,$$

where $[z]_+ = \max(z, 0)$. Then, it follows that $\widehat{\beta}_2 = 0$. The case $\widehat{\beta}_1 = 0$ is derived analogously. Overall, this gives the following recipe for determining $\widehat{\beta}_1, \widehat{\beta}_2$: one starts with

$$\begin{bmatrix} \widetilde{\beta}_1 \\ \widetilde{\beta}_2 \end{bmatrix} = \begin{bmatrix} 1 + \lambda_2 l_{11} & \boldsymbol{X}_1^\top \boldsymbol{X}_2 + \lambda_2 l_{12} \\ \boldsymbol{X}_1^\top \boldsymbol{X}_2 + \lambda_2 l_{12} & 1 + \lambda_2 l_{22} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{X}_1^\top \boldsymbol{y} \\ \boldsymbol{X}_2^\top \boldsymbol{y} \end{bmatrix} \begin{bmatrix} 1 + \lambda_2 l_{11} & 0 \\ 0 & 1 + \lambda_2 l_{22} \end{bmatrix},$$

computes γ, and determines

$$\widehat{\beta}_1 = \left[\frac{|\widetilde{\beta}_1| - |\widetilde{\beta}_2|}{2} + \gamma\right]_+ (1 + \lambda_2 l_{11})^{-1},$$
$$\widehat{\beta}_2 = \left[\frac{|\widetilde{\beta}_2| - |\widetilde{\beta}_1|}{2} + \gamma\right]_+ (1 + \lambda_2 l_{22})^{-1}.$$

References

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society Series B 36*, 192–236.

Chung, F. (1997). *Spectral Graph Theory*. AMS Publications.

Daumer, M., K. Thaler, E. Kruis, W. Feneberg, G. Staude, and M. Scholz (2007). Steps towards a miniaturized, robust and autonomous measurement device for the long-term monitoring of patient activity: ActiBelt. *Biomed Tech 52*, 149–155.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least Angle Regression (with discussion). *The Annals of Statistics 32*, 407–499.

Efron, B. and R. Tibshirani (2007). On testing the significance of sets of genes. *The Annals of Applied Statistics 1*, 107–129.

Eilers, P. and B. Marx (1996). Flexible smoothing with B-Splines and penalties (with discussion). *Statistical Science 11*, 89–121.

Eilers, P. and B. Marx (1999). Generalized linear regression on sampled signals and curves: a P-spline approach. *Technometrics 41*, 1–13.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 97*, 210–221.

Frank, I. and J. Friedman (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics 35*, 109–148.

Friedman, J., T. Hastie, H. Hoefling, and R. Tibshirani (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics 2*, 302–332.

Genkin, A., D. Lewis, and D. Madigan (2007). Large-scale Bayesian logistic regression for text categorization. *Technometrics 49*, 589–616.

Geyer, C. (1996). On convex stochastic optimization. Technical report, Department of Statistics, University of Minnesota.

Goeman, J. (2007). An efficient algorithm for ℓ^1-penalized estimation. Technical report, Department of Medical Statistics and Bioinformatics, University of Leiden.

Golub, T., D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science 286*, 531–537.

Hastie, T., A. Buja, and R. Tibshirani (1995). Penalized Discriminant Analysis. *The Annals of Statistics 23*, 73–102.

Hoerl, A. and R. Kennard (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics 8*, 27–51.

Kanehisa, M. and S. Goto (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research 28*, 27–30.

Knight, K. and W. Fu (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics 28*, 1356–1378.

Li, C. and H. Li (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics 24*, 1175–1182.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models*. Chapman and Hall, London.

Park, T. and G. Casella (2008). The Bayesian Lasso. *Journal of the American Statistical Association 103*, 681–686.

Rosset, S., J. Zhu, and T. Hastie (2004). Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research 5*, 941–973.

Rue, H. and L. Held (2001). *Gaussian Markov Random Fields*. Chapman and Hall/CRC, Boca Raton.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics 9*, 1135–1151.

Le Cun, Y., B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation 2*, 541–551.

Tibshirani, R. (1996). Regression shrinkage and variable selection via the lasso. *Journal of the Royal Statistical Society Series B 58*, 671–686.

Tibshirani, R., T. Hastie, B. Narasimhan, and G. Chu. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statistical Science 18*, 104–117.

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

Tseng, P. (2001). Convergence of a block coordinate descent method for nondiffferentiable minimization. *Journal of Optimization Theory and Applications 109*, 475–494.

Tutz, G. and J. Gertheiss (2009). Feature extraction in signal regression: a boosting technique for functional data regression. *Journal of Computational and Graphical Statistics, to appear*.

Wang, L., J. Zhu, and H. Zhou (2006). The doubly regularized support vector machine. *Statistica Sinica 16*, 589–616.

Wu, T. and K. Lange (2008). Coordinate descent procedures for lasso penalized regression. *The Annals of Applied Statistics 1*, 224–244.

Zhao, P. and B. Yu (2006). On model selection consistency of the lasso. *Journal of Machine Learning Research 7*, 2541–2567.

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

Zhou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B 67*, 301–320.