The Annals of Applied Statistics
2010, Vol. 4, No. 2, 1056–1080
DOI: 10.1214/09-AOAS302

FEATURE SELECTION GUIDED BY STRUCTURAL INFORMATION

BY MARTIN SLAWSKI[1,2], WOLFGANG ZU CASTELL AND GERHARD TUTZ

Saarland University, Helmholtz Zentrum München and University of Munich

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. Deficiencies of the lasso in certain scenarios, notably strongly correlated design, were unmasked when Zou and Hastie [*J. Roy. Statist. Soc. Ser. B* **67** (2005) 301–320] introduced the elastic net. In this paper we propose to extend the elastic net by admitting general nonnegative quadratic constraints as a second form of regularization. The generalized ridge-type constraint will typically make use of the known association structure of features, for example, 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. 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, with a focus on 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 $\mathcal{Y}$. Given a realization $\mathbf{x} = (x_1, \ldots, x_p)^\top$ of $\mathbb{X}$, a prediction $\widehat{y}$ for a specific functional of the distribution of $Y|\mathbb{X} = \mathbf{x}$ is obtained via a linear predictor

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

and a function $\zeta : \mathbb{R} \to \mathcal{Y}$ such that $\widehat{y} = \zeta(f(\mathbf{x}))$. Given an i.i.d. sample $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ 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

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

Received May 2009; revised October 2009.

[1]A large fraction of this work was done while the author was at the Department of Statistics, University of Munich and the Sylvia Lawry Centre for Multiple Sclerosis Research, Munich.

[2]Supported in part by the Porticus Foundation in the context of the International School for Clinical Medicine and Bioinformatics.

Key words and phrases. Generalized linear model, regularization, sparsity, $p \gg n$, lasso, elastic net, random fields, model selection, signal regression.

where $L : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_0^+$ is a convex loss function. 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) 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) 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 equation (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, that is, one maximizes the loss subject to $\Omega(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1 < s, s > 0$. The latter is particularly attractive if one is interested in feature selection, since one obtains estimates $\widehat{\beta}_j$, $j \in \{1, \ldots, p\}$, 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, Lewis and Madigan (2007)] for determining the coefficients are further virtues of the lasso. Its limitations have recently been revealed by several researchers. Zou and Hastie (2005) pointed out that the lasso need not be unique in the $p \gg n$ setting, where the lasso is able to select at most n features [Rosset, Zhu and Hastie (2004)]. Furthermore, Zou 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 $\Omega(\boldsymbol{\beta}) = \alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\boldsymbol{\beta}\|^2, \alpha \in (0, 1)$ 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 $\Omega(\boldsymbol{\beta}) = \alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\mathbf{D}\boldsymbol{\beta}\|_1$, where

$$\begin{aligned} \mathbf{D}: \quad & \mathbb{R}^p \to \mathbb{R}^{p-1}, \\ & (\beta_1, \ldots, \beta_p)^\top \mapsto ([\beta_2 - \beta_1], \ldots, [\beta_p - \beta_{p-1}])^\top \end{aligned} \tag{2}$$

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:

$$\Omega(\boldsymbol{\beta}) = \alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\boldsymbol{\beta}^\top \boldsymbol{\Lambda}\boldsymbol{\beta}, \tag{3}$$

where $\boldsymbol{\Lambda} = (l_{jj'})_{1 \le j, j' \le p}$ is assumed to be symmetric and positive semidefinite. Setting $\boldsymbol{\Lambda} = \mathbf{I}$ yields the elastic net. 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. Therefore, expression (3) will be referred to as the structured elastic net regularizer. The structured elastic net estimator is defined as

$$
\begin{aligned}
&(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) = \underset{(\beta_0, \boldsymbol{\beta})}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i; \beta_0, \boldsymbol{\beta})) \\
&\text{subject to} \qquad \alpha \|\boldsymbol{\beta}\|_1 + (1-\alpha)\boldsymbol{\beta}^\top \boldsymbol{\Lambda} \boldsymbol{\beta} \le s, \qquad \alpha \in (0,1), s > 0,
\end{aligned} \tag{4}
$$

which is equivalent to the Lagrangian formulation

$$
\begin{aligned}
&(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) = \underset{(\beta_0, \boldsymbol{\beta})}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i; \beta_0, \boldsymbol{\beta})) + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \boldsymbol{\beta}^\top \boldsymbol{\Lambda} \boldsymbol{\beta}, \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \lambda_1, \lambda_2 > 0.
\end{aligned} \tag{5}
$$

The rest of the paper is organized as follows: in Section 2 we discuss the choice of the matrix $\boldsymbol{\Lambda}$, followed by an analysis of some important properties of our proposal (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 an algorithmic solution to compute the minimizers (5) in the generalized linear model family. The practical performance of the structured elastic net is contained in Section 6. Section 7 concludes with a discussion and an outlook. All proofs can be found in the online supplement supporting this article [Slawski, zu Castell and Tutz (2010)].

2. Structured features.

2.1. *Motivation*. A considerable fraction of contemporary regression problems are characterized by a large number of features, which are 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 generally, 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, the 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,

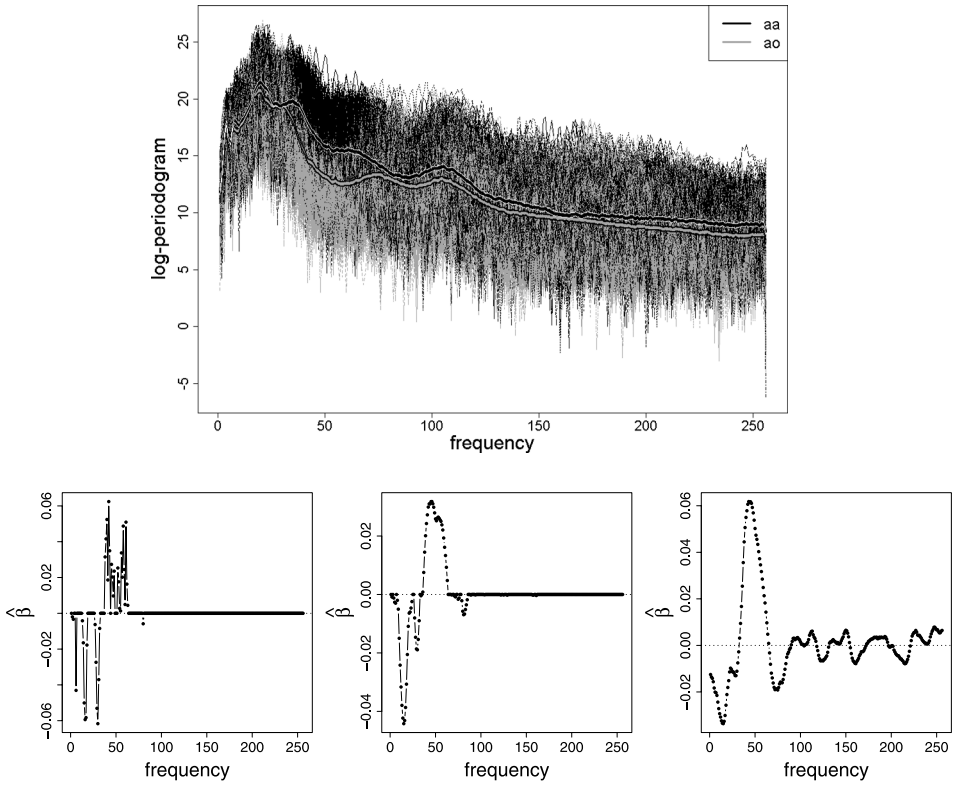

FIG. 1. *Phoneme data* [*Hastie, Buja and Tibshirani* (1995)]. *The upper panel shows several thousand log-periodogramms of the speech frames for the phonemes "aa"* (*as occuring in "dark"*) *and "ao"* (*as occuring in "water"*). *The classwise means are given by thick lines. We use linear logistic regression 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*).

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.

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, that is, $(\beta_j, \beta_j) \notin E$ for all j. The edges may additionally be weighted by a function $w : E \to \mathbb{R}$, $w((\beta_j, \beta_{j'})) =$

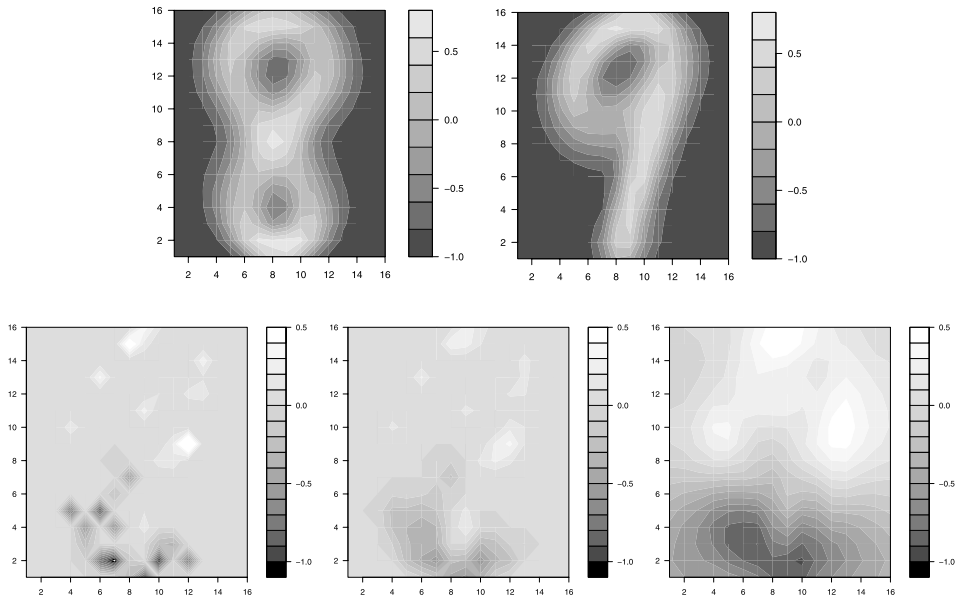

FIG. 2. *Handwritten digit recognition data set* [*Le Cun et al.* (1989)]. *One observation is given by a greyscale image composed of* 16×16 *pixels. The upper panel shows the contour of the pixel-wise means for the numerals "*8*" and "*9*." We use a training set of* 1500 *observations of eights and 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*).

$w((\beta_{j'}, \beta_j))$ 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 the Gauss–Markov random fields [Besag (1974); Rue and Held (2001)]. In our setup, the pairwise Markov property reads

$$\neg \beta_j \sim \beta_{j'} \quad \Leftrightarrow \quad \beta_j \perp\!\!\!\perp \beta_{j'} | V \setminus \{\beta_j, \beta_{j'}\}, \tag{6}$$

with $\perp\!\!\!\perp$ denoting conditional independence. Property (6) is conformed to the following choice for the precision matrix $\mathbf{\Lambda} = (l_{jj'})_{1 \le j, j' \le p}$:

$$l_{jj'} = \begin{cases} \displaystyle\sum_{k=1}^{p} |w((\beta_j, \beta_k))|, & \text{if } j = j', \\ -w((\beta_j, \beta_{j'})), & \text{if } j \neq j', \end{cases} \tag{7}$$

which is singular in general. If $\operatorname{sign}\{w((\beta_j, \beta_{j'}))\} \geq 0$ for all $(\beta_j, \beta_{j'})$ in E, then $\mathbf{\Lambda}$ as given in equation (7) is known as the combinatorial graph Laplacian in the

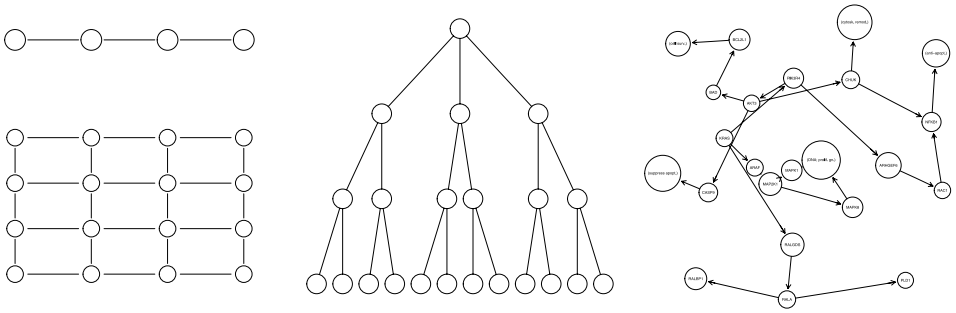

FIG. 3. *A collection of some graphs. A path and a grid* (*left panel*), *a rooted tree* (*middle panel*), *and an irregular graph describing a part of the so-called MAPK signaling pathway* (*right panel*).

spectral graph theory [Chung (1997)]. It is straightforward to verify the following properties:

-

$$\boldsymbol{\beta}^\top \boldsymbol{\Lambda} \boldsymbol{\beta} = \sum_{\beta_j \sim \beta_{j'}} |w(\beta_j, \beta_{j'})| \big(\beta_j - \operatorname{sign}\{w((\beta_j, \beta_{j'}))\}\beta_{j'}\big)^2 \geq 0, \tag{8}$$

  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 $\operatorname{sign}\{w((\beta_j, \beta_{j'}))\} \geq 0$ for all $(\beta_j, \beta_{j'})$ in E, the null space 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 (8) equals the summed squared forward differences

$$\sum_{j=2}^{p} (\beta_j - \beta_{j-1})^2 = \|\mathbf{D}\boldsymbol{\beta}\|^2 = \boldsymbol{\beta}^\top \mathbf{D}^\top \mathbf{D} \boldsymbol{\beta}, \tag{9}$$

where $\mathbf{D}$ is defined in equation (2). More complex graphical structures can be generated from simple ones using the notion of Cartesian products of graphs [Chung (1997), page 37]. For instance (as displayed in the left panel of Figure 3), the Cartesian product of a p-path and a p'-path equals a $p \times p'$ regular grid, in which case the standard choice of $\boldsymbol{\Lambda}$ is seen to be a 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, Buja and Tibshirani (1995)] and spline smoothing [Eilers and Marx (1996)].

2.3. *Connection to manifold regularization.* As pointed out by one of the referees, regularizers of the form (8) are applicable to a variety of learning problems in which data are supposed to be generated according to a probability measure supported on a compact, smooth manifold $M \subset \mathbb{R}^p$. The canonical regularization operator acting on smooth functions on M is the Laplace–Beltrami operator Δ_M [Rosenberg (1997)], which generalizes the Laplacian for Euclidean domains. As suggested, for example, in Belkin, Niyogi and Sindwhani (2006), given a set of data points in $\mathbb{R}^p$, a discrete proxy for a potential manifold structure can be obtained by computing a (possibly weighted) neighborhood graph of the points, and, in turn, a proxy for Δ_M is obtained by a discrete Laplacian of the form (7) resulting from the neighborhood graph.

Relating these ideas to our framework, one might think of settings where each of the $\{\mathbf{x}_i\}_{i=1}^n$ represents a collection of p points sampled on a compact, smooth manifold M. This is a natural extension of the introductory examples in Section 2.1, where the corresponding M would be given by an interval and a rectangle, respectively. Assuming a linear relationship between scalar responses $\{y_i\}_{i=1}^n$ and the predictors $\{\mathbf{x}_i\}_{i=1}^n$, we expect the corresponding coefficient vector to be both sparse and smooth with respect to the manifold structure. Without going into detail, the approach might be useful for predictors with geographical information. The idea is illustrated in Figure 4 where M is chosen as a sphere embedded in $\mathbb{R}^3$.

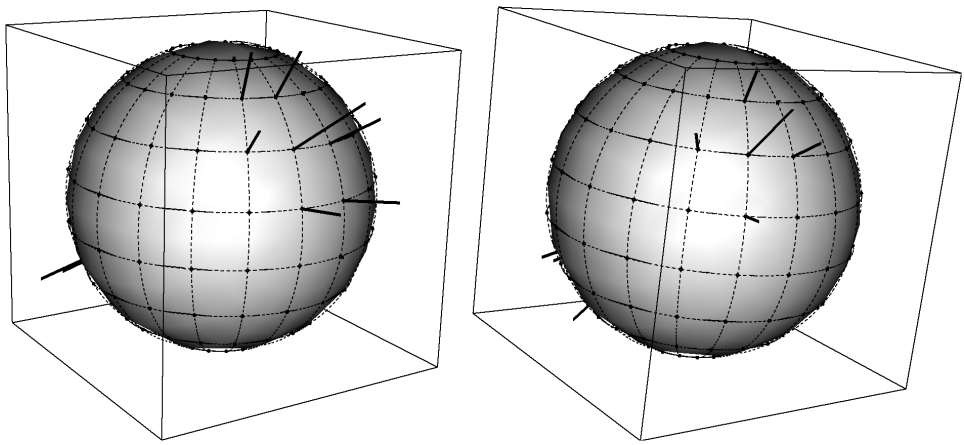

FIG. 4. *A manifold setting suitable to our regularizer. The black dots represent points at which the random variables X_j, $j = 1, \ldots, p$, are realized, and the spikes normal to the surface indicate the size of the corresponding β_j^*, $j = 1, \ldots, p$. Except for the two groups highlighted in the left and right panel, respectively, the coefficients equal zero. The dashed lines represent the neighborhood graph obtained by connecting each dot with its four nearest neighbors with respect to the geodesic distance on the sphere.*

3. Properties.

3.1. *Bayesian and geometric interpretation.* In the setup of Section 1, consider the regularizer

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

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

$$L(y, f(\mathbf{x}; \beta_0, \boldsymbol{\beta})) = \phi^{-1}\big(b(f(\mathbf{x})) - y f(\mathbf{x})\big) + c(y, \phi), \tag{10}$$

that is, 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 [cf. McCullagh and Nelder (1989)]

$$\begin{aligned} &Y|\mathbb{X} = \mathbf{x} \sim \text{simple exponential family}, \\ &\widehat{y} = \mathrm{E}[Y|\mathbb{X} = \mathbf{x}] = \mu = \frac{d}{df} b(f(\mathbf{x})), \\ &\mathrm{var}[Y|\mathbb{X} = \mathbf{x}] = \phi \frac{d^2}{df^2} b(f(\mathbf{x})). \end{aligned} \tag{11}$$

The form (10) 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 (10), the regularizer $\Omega(\boldsymbol{\beta})$ can be interpreted as the combined Laplace (double exponential)-Gaussian prior $p(\boldsymbol{\beta}) \propto \exp(-\Omega(\boldsymbol{\beta}))$, for which the structured elastic net estimator (5), provided $p(\beta_0) \propto 1$, is the maximum posterior (MAP) estimator given the sample S. It is instructive to consider two predictors, that is, $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top$. Figure 5 gives a geometric interpretation for the basic choices

$$\boldsymbol{\Lambda} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Lambda} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix},$$

corresponding 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 $\boldsymbol{\Lambda}$ determines the orientation of the ellipsoidal part.

3.2. *A grouping property.* For the elastic net, Zou 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 Zou 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 $\mathbf{X} = (x_{ij})_{\substack{1 \le i \le n \\ 1 \le j \le p}}$ and denote by $\mathbf{X}_j = (x_{1j}, \ldots, x_{nj})^\top$ the realizations of predictor j in S, and the response vector

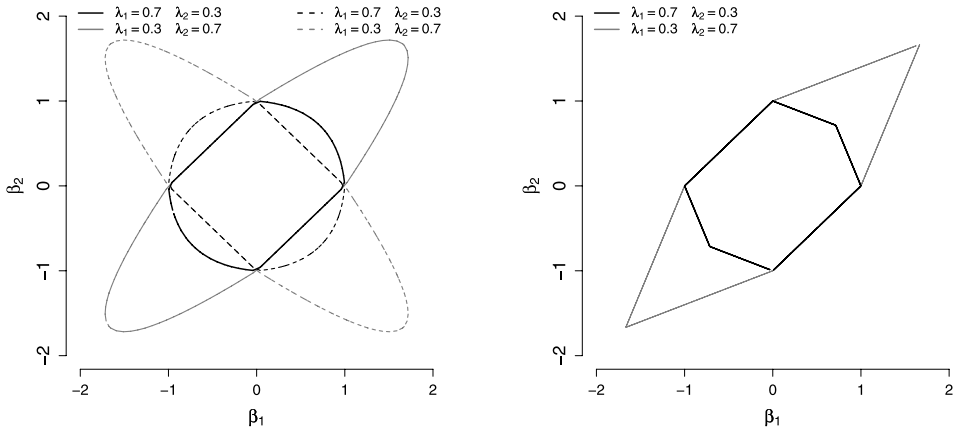

FIG. 5. *Level sets* $\{(\beta_1, \beta_2) : \lambda_1(|\beta_1| + |\beta_2|) + \lambda_2(\beta_1 - \beta_2)^2 = 1\}$ (*left panel*, *solid lines*) *and* $\{(\beta_1, \beta_2) : \lambda_1(|\beta_1| + |\beta_2|) + \lambda_2(\beta_1 + \beta_2)^2 = 1\}$ (*left panel*, *dashed lines*) *of the structured elastic net regularizer and* $\{(\beta_1, \beta_2) : \lambda_1(|\beta_1| + |\beta_2|) + \lambda_2|\beta_1 - \beta_2| = 1\}$ *of the fused lasso*.

is defined by $\mathbf{y} = (y_1, \ldots, y_n)^\top$. 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, that is,

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} x_{ij} = 0, \qquad \sum_{i=1}^{n} x_{ij}^2 = 1, \qquad j = 1, \ldots, p. \tag{12}$$

PROPOSITION 1. *Letting* $p = 2$, *let the loss function be of the form* (10), *let* $\rho = \mathbf{X}_1^\top \mathbf{X}_2$ *denote the sample correlation of* $\mathbf{X}_1$ *and* $\mathbf{X}_2$, *and let* $\mathbf{\Lambda} = \frac{1}{2}\left(\begin{smallmatrix} 1 & s \\ s & 1 \end{smallmatrix}\right)$, $s \in \{-1, 1\}$. *If* $-s\widehat{\beta}_1\widehat{\beta}_2 > 0$, *then*

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

In particular, in the setting of Proposition 1, we have the implication that if $\mathbf{X}_1 = -s\mathbf{X}_2$, then $\widehat{\beta}_1 = -s\widehat{\beta}_2$.

3.3. *Decorrelation.* Let us now consider the important special case

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

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

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \operatorname*{argmin}_{\boldsymbol{\beta}} -2\mathbf{y}^\top \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top [\mathbf{C} + \lambda_2 \mathbf{\Lambda}]\boldsymbol{\beta} + \lambda_1 \|\boldsymbol{\beta}\|_1, \qquad \mathbf{C} = \mathbf{X}^\top \mathbf{X}, \\ &= \operatorname*{argmin}_{\boldsymbol{\beta}} -2\mathbf{y}^\top \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \widetilde{\mathbf{C}}\boldsymbol{\beta} + \lambda_1 \|\boldsymbol{\beta}\|_1, \qquad \widetilde{\mathbf{C}} = \mathbf{X}^\top \mathbf{X} + \lambda_2 \mathbf{\Lambda}. \end{aligned} \tag{13}$$

Note that for standardized predictors, $\mathbf{C}$ equals the matrix of sample correlations $\rho_{jj'} = \mathbf{X}_j^\top \mathbf{X}_{j'}$, $j, j' = 1, \ldots, p$. With a large number of predictors or elements $\rho_{jj'}$ with large $|\rho_{jj'}|$, $\mathbf{C}$ 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 $\mathbf{X}_j$ and $\mathbf{X}_{j'}$, which in turn yield a strongly negative correlation between $\widehat{\beta}_j^{\text{ols}}$ and $\widehat{\beta}_{j'}^{\text{ols}}$ and, as a consequence, high variances $\text{var}[\widehat{\beta}_j^{\text{ols}}]$ and $\text{var}[\widehat{\beta}_{j'}^{\text{ols}}]$. In the prevalence of high correlations, performance of the lasso may degrade as well. For example, Donoho, Elad and Temlyakov (2006) showed that the lower the *mutual coherence* $\max_{\substack{j,j' \\ j \neq j'}} |\rho_{j,j'}|$, the more stable is lasso estimation. The modified matrix $\widetilde{\mathbf{C}}$ can be written as $\widetilde{\mathbf{C}} = \mathbf{V}_{\boldsymbol{\Lambda}}^{1/2} \mathbf{R}_{\boldsymbol{\Lambda}} \mathbf{V}_{\boldsymbol{\Lambda}}^{1/2}$, $\mathbf{V}_{\boldsymbol{\Lambda}} = \text{diag}(1 + \lambda_2 \sum_{k=1}^p |l_{1k}|, \ldots, 1 + \lambda_2 \sum_{k=1}^p |l_{pk}|)$, and the modified correlation matrix $\mathbf{R}_{\boldsymbol{\Lambda}}$ has entries

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

In light of Section 2, the entries of $\mathbf{R}_{\boldsymbol{\Lambda}}$ combine sample- and prior correlations. Decorrelation occurs if $\rho_{jj'} \approx -\lambda_2 l_{jj'}$.

4. Consistency. The asymptotic analysis presented in this section closely follows the ideas of Knight and Fu (2000) and Zou (2006). Both have studied asymptotics for the lasso in linear regression for a fixed number of predictors under conditions ensuring $\sqrt{n}$-consistency and asymptotic normality of the ordinary least squares estimator. Knight and Fu (2000) proved that the lasso estimator $\widehat{\boldsymbol{\beta}}^{\text{lasso}}$ is $\sqrt{n}$-consistent for the true coefficient vector $\boldsymbol{\beta}^*$ provided $\lambda_1^n = O(\sqrt{n})$. Zou (2006) has shown that while this choice of λ_1^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 be selection consistent if

$$\lim_{n \to \infty} \text{P}(\widehat{\beta}_{j,n}^\delta \neq 0) = 1 \qquad \text{for } j \in A,$$

$$\lim_{n \to \infty} \text{P}(\widehat{\beta}_{j,n}^\delta = 0) = 1 \qquad \text{for } j \in A^c,$$

where here and in the following, the sub- or superscript n indicates that the corresponding quantity depends on the sample size n. Moreover, Zou (2006) and Zhao and Yu (2006) have shown that if $\lambda_1^n = o(n)$ 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. Zou (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. *Define*

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

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*

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

Then, under conditions (C.1)–(C.3) *in the online supplement,* $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*) \xrightarrow{D} \operatorname{argmin} V(\mathbf{u})$.

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. equation (11)]. Let the true model be defined by

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

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

$$\widehat{\boldsymbol{\beta}}_n = \operatorname*{argmin}_{\boldsymbol{\beta}} 2\phi^{-1}\sum_{i=1}^{n} b(f(\mathbf{x}_i; \boldsymbol{\beta})) - y_i f(\mathbf{x}_i; \boldsymbol{\beta}) + \lambda_1^n\|\boldsymbol{\beta}\|_1 + \lambda_2^n\boldsymbol{\beta}^\top\boldsymbol{\Lambda}\boldsymbol{\beta}. \tag{14}$$

THEOREM 2. *For the estimator* (14), *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*

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

Then under conditions (G.1) *and* (G.2) *in the online supplement,* $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*) \xrightarrow{D} \operatorname{argmin} W(\mathbf{u})$.

Now let us turn to the question of selection consistency. In the setup of Theorem 1, if λ_1^n and λ_2^n both are $O(\sqrt{n})$, then, using arguments similar to those in

Knight and Fu (2000) and Zou (2006), $\widehat{\boldsymbol{\beta}}_n$ is shown not to be selection consistent. Selection consistency can be achieved if one lets λ_1^n, λ_2^n grow more strongly and if the quantities $\mathbf{C}$, $\boldsymbol{\Lambda}$, and $\boldsymbol{\beta}^*$ jointly fulfill a nontrivial condition, which can be seen as analog to the *irrepresentable condition* of the lasso [Zou (2006), Zhao and Yu (2006)].

THEOREM 3. *In the situation of Theorem 1, let* $\lambda_1^n/n \to 0$, $\lambda_1^n/\sqrt{n} \to \infty$, $\lambda_2^n/\lambda_1^n \to R, 0 < R < \infty$ *and consider the partitioning scheme*

$$\boldsymbol{\beta}^* = \begin{pmatrix} \boldsymbol{\beta}_A^* \\ \boldsymbol{\beta}_{A^c}^* \end{pmatrix}, \qquad \mathbf{C} = \begin{pmatrix} \mathbf{C}_A & \mathbf{C}_{AA^c} \\ \mathbf{C}_{A^cA} & \mathbf{C}_{A^c} \end{pmatrix} \quad \text{and} \tag{15}$$
$$\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_A & \boldsymbol{\Lambda}_{AA^c} \\ \boldsymbol{\Lambda}_{A^cA} & \boldsymbol{\Lambda}_{A^c} \end{pmatrix},$$

so that here and in the following, the subscripts A and A^c *refer to active and inactive set, respectively. Then, if selection consistency holds, the following condition must be fulfilled: there exists a sign vector* $\mathbf{s}_A$ *such that*

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

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 Zou (2006) under the name adaptive lasso, which replaces ℓ^1-regularization uniform in $\beta_j, j = 1, \ldots, p$, 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 a function of an "initial estimator" $\widehat{\boldsymbol{\beta}}^{\text{init}}$:

$$\omega_j = |\widehat{\beta}_j^{\text{init}}|^{-\gamma}, \qquad \gamma > 0, j = 1, \ldots, p. \tag{16}$$

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

THEOREM 4. *In the situation of Theorem 1, define*

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

where the weights are as in equation (16), and suppose that the initial estimator satisfies

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

Furthermore, suppose that

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

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

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

(2) $\lim_{n\to\infty} \mathrm{P}(\widehat{\boldsymbol{\beta}}^{\text{adaptive}}_{A^c,n} = \mathbf{0}) = 1$.

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 null space of $\boldsymbol{\Lambda}_A$. Interestingly, if $\boldsymbol{\Lambda}$ equals the combinatorial graph Laplacian (cf. Section 2.2), 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. Last, we remark that while Theorem 4 applies to linear regression, it can be extended to hold for generalized linear models, similarly as we have extended 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 (10).

5.1. *Data augmentation*. From the discussions in Section 3.3, it follows that the structured elastic net for squared loss, assuming centered data, can be recast as the lasso on augmented data

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

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 $\widehat{\boldsymbol{\beta}}(\lambda_1; \lambda_2)$. This approach is parallel to that proposed by Zou and Hastie (2005) for the elastic net. In addition, the augmented data representation is helpful when addressing uniqueness of the structured elastic net in the $p \gg n$ setting: if $\operatorname{rank}(\mathbf{X}) + \operatorname{rank}(\lambda_2^{1/2}\mathbf{Q}) \geq p$ and the rows of $\mathbf{X}$ combined with the rows of $\lambda_2^{1/2}\mathbf{Q}$ form a linearly independent set, $\widetilde{\mathbf{C}}$ as defined in equation (13) is of full rank and, hence, the structured elastic net is unique. Moreover, this shows that even for $p \gg n$, in principle, all features can be selected.

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:

$$
\begin{aligned}
\begin{pmatrix} \widehat{\beta}_0^{(k+1)} \\ \widehat{\boldsymbol{\beta}}^{(k+1)} \end{pmatrix} &= \left([\mathbf{1} \quad \mathbf{X}]^\top \mathbf{W}^{(k)} [\mathbf{1} \quad \mathbf{X}]\right)^{-1} [\mathbf{1} \quad \mathbf{X}]^\top \mathbf{W}^{(k)} \mathbf{z}^{(k)}, \\
\mathbf{z}^{(k)} &= \mathbf{f}^{(k)} + \left[\mathbf{W}^{(k)}\right]^{-1} (\mathbf{y} - \boldsymbol{\mu}^{(k)}), \\
\mathbf{f}^{(k)} &= \left(f_1^{(k)}, \ldots, f_n^{(k)}\right)^\top, \qquad f_i^{(k)} = \widehat{\beta}_0^{(k)} + \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}^{(k)}, i = 1, \ldots, n, \\
\boldsymbol{\mu}^{(k)} &= \left(\mu_1^{(k)}, \ldots, \mu_n^{(k)}\right)^\top, \qquad \mu_i^{(k)} = b'\left(f_i^{(k)}\right), i = 1, \ldots, n, \\
\mathbf{W}^{(k)} &= \operatorname{diag}\left(w_1^{(k)}, \ldots, w_n^{(k)}\right), \qquad w_i^{(k)} = \phi^{-1} b''\left(f_i^{(k)}\right), i = 1, \ldots, n.
\end{aligned} \tag{17}
$$

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

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

with

$$
\begin{aligned}
\widetilde{w}_i^{(k)} &= w_i^{(k)}, \qquad i = 1, \ldots, n, \qquad \text{as in equation (17)}, \\
\widetilde{w}_i^{(k)} &= 1, \qquad i = (n+1), \ldots, (n+p), \\
\widetilde{z}_i^{(k)} &= z_i^{(k)}, \qquad i = 1, \ldots, n, \qquad \text{as in equation (17)}, \\
\widetilde{z}_i^{(k)} &= 0, \qquad i = (n+1), \ldots, (n+p), \\
\widetilde{\mathbf{x}}_i &= (1 \quad \mathbf{x}_i^\top)^\top, \qquad i = 1, \ldots, n, \\
\widetilde{\mathbf{x}}_i &= (0 \quad \sqrt{\lambda_2} \mathbf{q}_i^\top)^\top, \qquad i = (n+1), \ldots, (n+p),
\end{aligned}
$$

with $\mathbf{q}_i^\top$ denoting the ith row of $\mathbf{Q}$.

Alternatives to augmented data representation include cyclical coordinate descent in the spirit of Friedman et al. (2007) and a direct modification of Goeman's algorithm [Goeman (2007)]. Descriptions can be a found in the full technical report underlying this article [Slawski, zu Castell and Tutz (2009), available online].

6. Data analysis.

6.1. *One-dimensional signal regression.* In one-dimensional signal regression, as described, for example, in Frank and Friedman (1993), one aims at the prediction of a response given a sampled signal $\mathbf{x}^\top = (x(t))_{t=1}^T$, 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

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

6.1.1. *Simulation study.* Similarly to Tutz and Gertheiss (2010), we simulate signals $x(t), t = 1, \ldots, T$, $T = 100$, according to

$$\{x(t)\} \sim \sum_{k=1}^{5} b_k \sin\bigl(t\pi(5 - b_k)/50 - m_k\bigr) + \tau(t),$$
$$\{b_k\} \sim U(0, 5), \qquad \{m_k\} \sim U(0, 2\pi), \qquad \{\tau(t)\} \sim N(0, 0.25),$$

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 the "bump setting," we use

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

In the second case, referred to as the "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.

For both settings, data are simulated according to

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

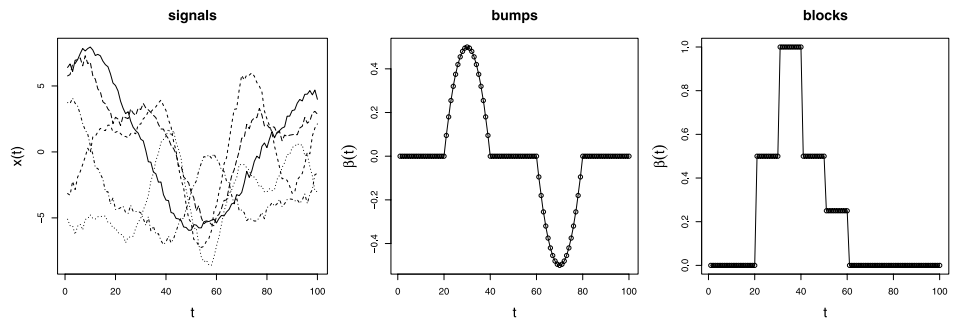

FIG. 6. *The setting of the simulation study. A collection of five signals (left panel), the coefficient functions for "bump"—(middle panel) and "block" setting (right panel), respectively.*

For 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 $L^1(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1$ 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. 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.

For the bump setting, Figure 7 shows that the double-regularized procedures employing decorrelation clearly outperform a visibly unstable lasso. Due to a favorable signal-to-noise ratio, even simplistic approaches such as ridge- or generalized ridge regression show competitive performance with respect to prediction

TABLE 1
Results for the bump setting, averaged over 50 *simulations*

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

For annotation, see Table 2.

TABLE 2

Results for the block 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*

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

of future observations. In pure numbers, the estimation of $\beta^*(t)$ 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, but it lacks smoothness. Its numerical inferiority to ridge regression results from double shrinkage as discussed in Zou and Hastie (2005). 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

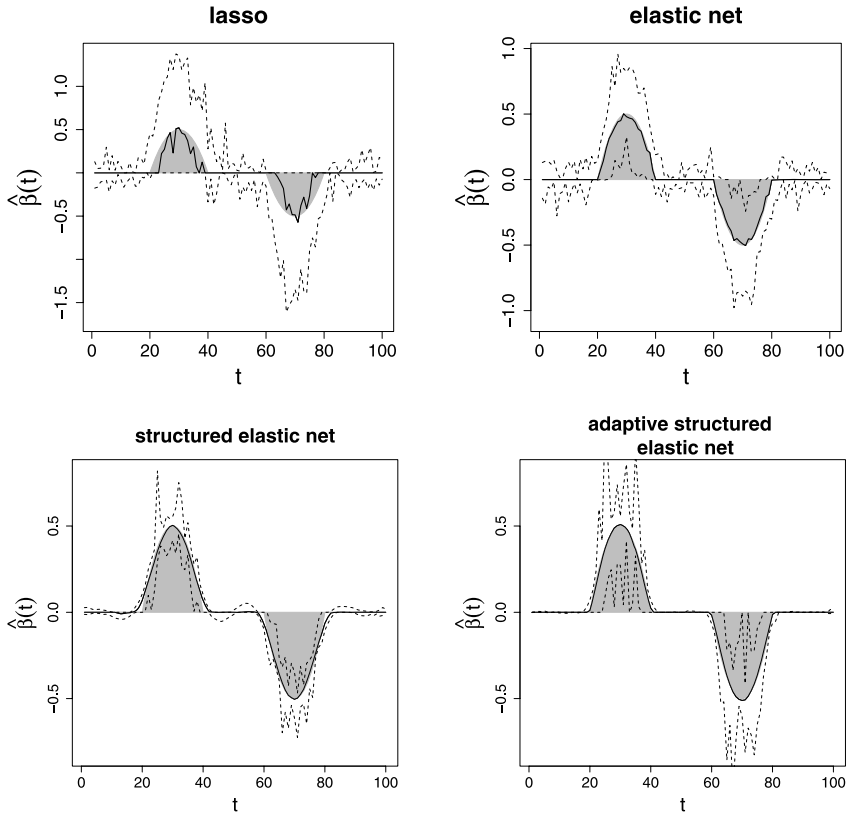

FIG. 7. *Estimated coefficient functions for the bump setting. The pointwise median curve over* 50 *iterations is represented by a solid line, pointwise* 0.05- *and* 0.95*-quantiles are drawn in dashed lines.*

net with respect to prediction, though the structure part is seen to be not fully appropriate in the block setting (cf. Figure 8). As opposed to the bump setting, fitting the block function seems to be much more difficult to accomplish in general.

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 the 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

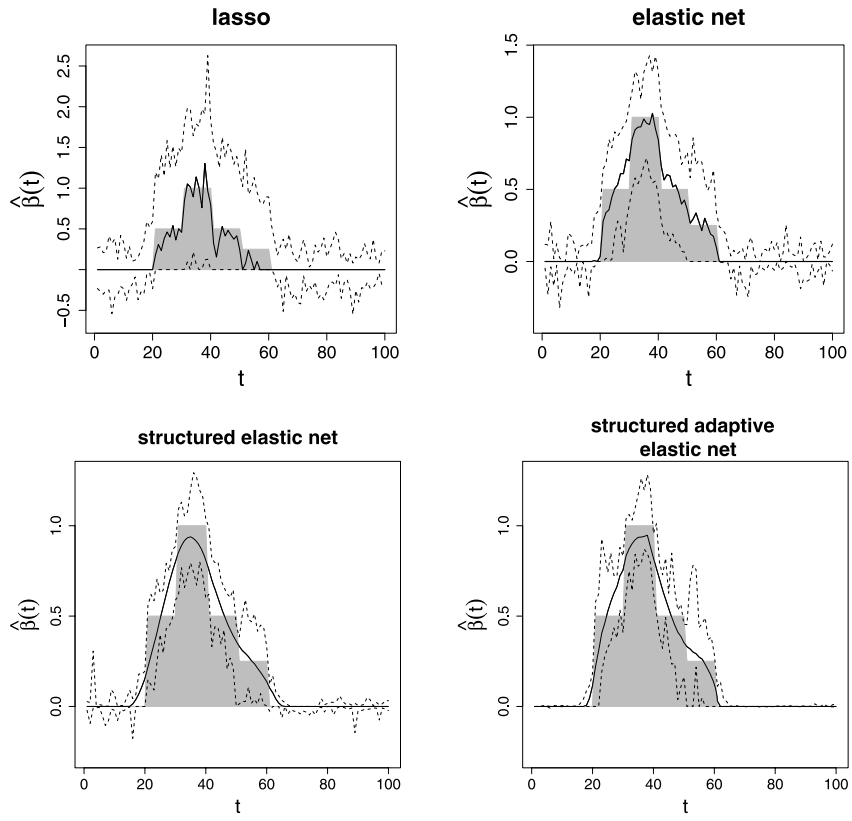

FIG. 8. *Estimated coefficient functions for the block setting.*

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 a natural unit. As a 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 $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where each $\mathbf{x}_i = (x_i(t)), t = 1, \ldots, T = 102$, 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

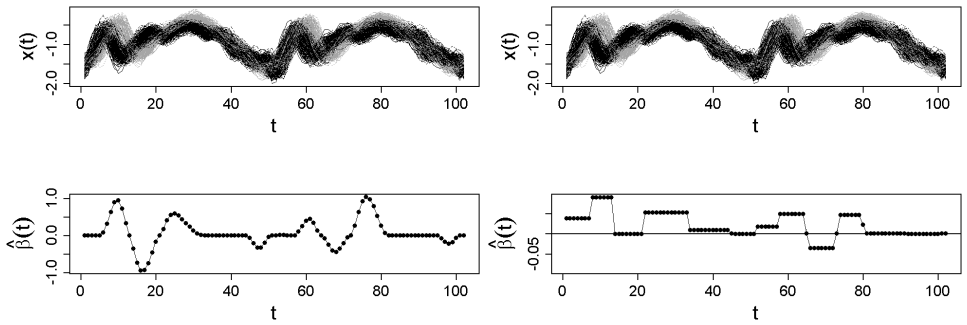

FIG. 9. *Coefficient functions for structured elastic net-regularized logistic regression* (*left panel*) *and the fused lasso support vector machine* (*right panel*). *Within each panel, the upper panel displays the overlayed double step patterns of the complete sample* (406 *double steps*). *The colors of the curves refer to the two persons.*

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 Figure 9 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.

6.2. *Surface fitting*. Figure 10 depicts the surface to be fitted on a 20×20 grid. The surface can be represented by a discrete function $\beta^*(t, u), t, u = 1, \ldots, 20$. It consists of three nonoverlapping truncated Gaussians of different shape and one

TABLE 3

Results of step classification for the fused lasso support vector machine and structured elastic net-regularized logistic regression. The bound imposed on the 1*-norm of* $\boldsymbol{\beta}$ *corresponding to* λ_1 *is denoted by* t_1, *while* t_2 *corresponding to* λ_2 *denotes the bound imposed on the absolute differences* $\sum_{t=2}^{T} |\beta(t) - \beta(t-1)|$ *for the fused lasso and the squared differences* $\sum_{t=2}^{T} (\beta(t) - \beta(t-1))^2$ *for the structured elastic net, respectively. Concerning the degrees of freedom of the two procedures, we take the number of nonzero blocks for the fused lasso. For the structured elastic net, we make use of a heuristic due to Tibshirani* (1996) *that rewrites the lasso fit as the weighted ridge fit; see Slawski, zu Castell and Tutz* (2009) *for details*

t_1	t_2	Test error	Degrees of freedom	# nonzero coefficients
			Fused lasso	
2.5	0.5	0	9	46
			Structured elastic net	
23	2	1	7.85	61

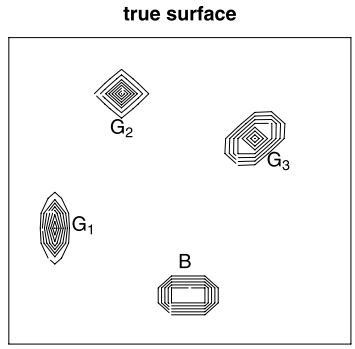

FIG. 10. *Contours of the surface according to equation* (18).

plateau function. We have

$$
\begin{aligned}
&\beta^*(t,u) = B(t,u) + G_1(t,u) + G_2(t,u) + G_3(t,u),\\
&B(t,u) = \frac{1}{2} I(t \in \{10,11,12\}, u \in \{3,4\}),\\
&G_1(t,u) = \max\left\{0, \exp\left(-(t-3u-8)\begin{pmatrix} 3 & 0 \\ 0 & 0.25 \end{pmatrix}\begin{pmatrix} t-3 \\ u-8 \end{pmatrix}\right) - 0.2\right\},\\
&G_2(t,u) = \max\left\{0, \exp\left(-(t-7u-17)\right.\right.\\
&\qquad\qquad \left.\left.\times \begin{pmatrix} 0.75 & 0 \\ 0 & 0.75 \end{pmatrix}\begin{pmatrix} t-7 \\ u-17 \end{pmatrix}\right) - 0.2\right\},\\
&G_3(t,u) = \max\left\{0, \exp\left(-(t-15u-14)\right.\right.\\
&\qquad\qquad \left.\left.\times \begin{pmatrix} 0.5 & -0.25 \\ -0.25 & 0.5 \end{pmatrix}\begin{pmatrix} t-15 \\ u-14 \end{pmatrix}\right) - 0.2\right\}.
\end{aligned} \tag{18}
$$

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

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

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 11 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.

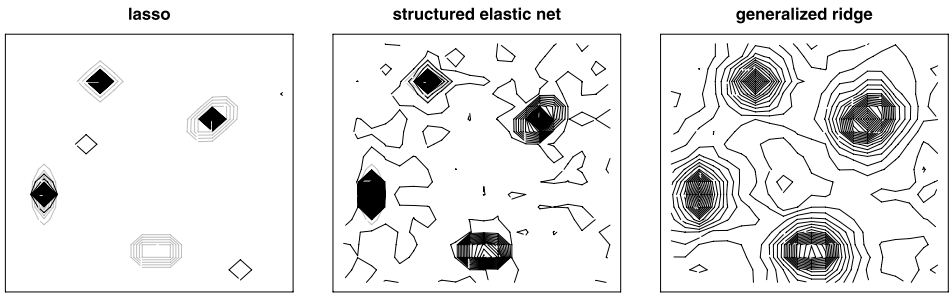

FIG. 11. *Contours of the estimated surfaces for three selected methods, averaged pointwise over* 50 *runs*.

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 nondifferentiability of the structure part of the fused lasso, computation may be intractable even for moderately sized graphs.

TABLE 4

Results of the simulation, averaged over 50 *iterations* (*standard errors in parentheses*). *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* 2. *The prediction error has been rescaled by* 100

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

Turning to the drawbacks of the structured elastic net, it is obvious that model selection and computation of standard errors and, in turn, the quantification of uncertainty, are 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 possible directions of future research, we will consider studying the structured elastic net in combination with other loss functions, for example, the hinge loss of support vector machines or the check loss for quantile regression. 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. We thank the Sylvia Lawry Centre Munich e.V. for its support with the accelerometer data example, in particular, Martin Daumer for numerous discussions, and 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.

We thank two reviewers, an associate editor, and an area editor for their constructive comments and suggestions, which helped us to improve on an earlier draft.

SUPPLEMENTARY MATERIAL

Supplement to "Feature Selection guided by Structural Information" (DOI: 10.1214/09-AOAS302SUPP; .pdf). The supplement contains proof of all statements of the main article.

REFERENCES

BELKIN, M., NIYOGI, P. and SINDWHANI, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.* **7** 2399–2434. MR2274444

BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc. Ser. B* **36** 192–236. MR0373208

CHUNG, F. (1997). *Spectral Graph Theory*. AMS Publications. MR1421568

DAUMER, M., THALER, K., KRUIS, E., FENEBERG, W., STAUDE, G. and SCHOLZ, M. (2007). Steps towards a miniaturized, robust and autonomous measurement device for the long-term monitoring of patient activity: ActiBelt. *Biomed. Tech.* **52** 149–155.

DONOHO, D., ELAD, M. and TEMLYAKOV, V. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory* **52** 6–18. MR2237332

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression (with discussion). *Ann. Statist.* **32** 407–499. MR2060166

EILERS, P. and MARX, B. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statist. Sci.* **11** 89–121. MR1435485

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

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

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

FRIEDMAN, J., HASTIE, T., HOEFLING, H. and TIBSHIRANI, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Statist.* **2** 302–332. MR2415737

GENKIN, A., LEWIS, D. and MADIGAN, D. (2007). Large-scale Bayesian logistic regression for text categorization. *Technometrics* **49** 589–616. MR2408634

GOEMAN, J. (2007). An efficient algorithm for ℓ^1-penalized estimation. Technical report, Dept. Medical Statistics and Bioinformatics, Univ. Leiden.

HASTIE, T., BUJA, A. and TIBSHIRANI, R. (1995). Penalized discriminant analysis. *Ann. Statist.* **23** 73–102. MR1331657

HOERL, A. and KENNARD, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **8** 27–51.

KNIGHT, K. and FU, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28** 1356–1378. MR1805787

LE CUN, Y., BOSER, B., DENKER, J., HENDERSON, D., HOWARD, R., HUBBARD, W. and JACKEL, L. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **2** 541–551.

MCCULLAGH, P. and NELDER, J. (1989). *Generalized Linear Models*. Chapman & Hall, London. MR0727836

PARK, T. and CASELLA, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.* **103** 681–686.

ROSENBERG, S. (1997). *The Laplacian on a Riemannian Manifold*. Cambridge Univ. Press, Cambridge. MR1462892

ROSSET, S., ZHU, J. and HASTIE, T. (2004). Boosting as a regularized path to a maximum margin classifier. *J. Mach. Learn. Res.* **5** 941–973. MR2248005

RUE, H. and HELD, L. (2001). *Gaussian Markov Random Fields*. Chapman & Hall/CRC, Boca Raton. MR2130347

SLAWSKI, M., ZU CASTELL, W. and TUTZ, G. (2009). Feature selection guided by structural Information. Technical report, Dept. Statistics, Univ. Munich. Available at http://epub.ub.uni-muenchen.de/10251/.

SLAWSKI, M., ZU CASTELL, W. and TUTZ, G. (2010). Supplement to "Feature selection guided by structural information." DOI: 10.1214/09-AOAS302SUPP.

TIBSHIRANI, R. (1996). Regression shrinkage and variable selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 671–686. MR1379242

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. Roy. Statist. Soc. Ser. B* **67** 91–108. MR2136641

TUTZ, G. and GERTHEISS, J. (2010). Feature extraction in signal regression: A boosting technique for functional data regression. *J. Computat. Graph. Statist.* **19** 154–174.

ZHAO, P. and YU, B. (2006). On model selection consistency of the lasso. *J. Mach. Learn. Res.* **7** 2541–2567. MR2274449

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B* **67** 301–320. MR2137327

M. SLAWSKI
DEPARTMENT OF COMPUTER SCIENCE
SAARLAND UNIVERSITY
SAARBRÜCKEN
GERMANY
E-MAIL: ms@cs.uni-sb.de

W. ZU CASTELL
INSTITUTE OF BIOMATHEMATICS AND BIOMETRY
HELMHOLTZ ZENTRUM MÜNCHEN
NEUHERBERG
GERMANY
E-MAIL: castell@helmholtz-muenchen.de

G. TUTZ
DEPARTMENT OF STATISTICS
UNIVERSITY OF MUNICH
MUNICH
GERMANY
E-MAIL: tutz@stat.uni-muenchen.de