Stat Comput (2012) 22:153–168
DOI 10.1007/s11222-010-9214-z

The structured elastic net for quantile regression and support vector classification

Martin Slawski

Received: 10 May 2010 / Accepted: 20 October 2010 / Published online: 6 November 2010

Abstract In view of its ongoing importance for a variety of practical applications, feature selection via ℓ_1-regularization methods like the lasso has been subject to extensive theoretical as well empirical investigations. Despite its popularity, mere ℓ_1-regularization has been criticized for being inadequate or ineffective, notably in situations in which additional structural knowledge about the predictors should be taken into account. This has stimulated the development of either systematically different regularization methods or double regularization approaches which combine ℓ_1-regularization with a second kind of regularization designed to capture additional problem-specific structure. One instance thereof is the 'structured elastic net', a generalization of the proposal in Zou and Hastie (J. R. Stat. Soc. Ser. B 67:301–320, 2005), studied in Slawski et al. (Ann. Appl. Stat. 4(2):1056–1080, 2010) for the class of generalized linear models.

In this paper, we elaborate on the structured elastic net regularizer in conjunction with two important loss functions, the check loss of quantile regression and the hinge loss of support vector classification. Solution paths algorithms are developed which compute the whole range of solutions as one regularization parameter varies and the second one is kept fixed.

The methodology and practical performance of our approach is illustrated by means of case studies from image classification and climate science.

Keywords Double regularization · Elastic net · Quantile regression · Solution path · Support vector classification

M. Slawski (✉)
Machine Learning Group, Department of Computer Science, Saarland University, Saarbrücken, Germany
e-mail: ms@cs.uni-sb.de

1 Introduction

Adopting the framework in Slawski et al. (2010), we let $\mathbb{X} = (X_1, \ldots, X_p)^\top$ be a random vector of real-valued predictor variables and let Y be a random response variable taking values in a set $\mathcal{Y}$. The aim is to model a functional $\psi[Y|\mathbb{X}]$ of the conditional distribution of $Y|\mathbb{X}$ via a linear predictor $f(\mathbb{X}; \beta_0^*, \boldsymbol{\beta}^*) = \beta_0^* + \mathbb{X}^\top \boldsymbol{\beta}^*$ and a function $\zeta : \mathbb{R} \to \mathcal{Y}$ such that $\psi[Y|\mathbb{X}] = \zeta(f(\mathbb{X}; \beta_0^*, \boldsymbol{\beta}^*))$. Given an i.i.d. sample $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ are realizations of $\mathbb{X}$ and Y, respectively, we try to infer $\beta_0^*, \boldsymbol{\beta}^*$ using regularized risk estimation, i.e. we determine estimators $\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}$ as minimizers

$$(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) = \operatorname*{argmin}_{(\beta_0, \boldsymbol{\beta})} \sum_{i=1}^n L_\psi(y_i, f(\boldsymbol{x}_i; \beta_0, \boldsymbol{\beta})) + \lambda \Omega(\boldsymbol{\beta}),$$
$$\lambda > 0, \tag{1.1}$$

where $L_\psi : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_0^+$ is a loss function related to ψ and $\Omega : \mathbb{R}^p \to \mathbb{R}_0^+$ is a regularizer. In this paper, we treat quantile regression (QR) and support vector classification (SVC), for which the relevant quantities are listed in Table 1.

For quantile regression, the quantity $\inf\{y : P(Y \leq y|\mathbb{X} = \boldsymbol{x}) \geq \tau\}$ is called the τ-quantile of the conditional distribution of Y given $\mathbb{X}$. Varying τ from 0 to 1 allows one to characterize that conditional distribution as a whole, which may be more informative than the usual approach of modeling the conditional expectation only. Choosing $\tau = 0.5$ yields least absolute deviation regression. For detailed information about quantile regression and its applications, we refer to Koenker (2005) and the references therein.

Turning to SVC, the loss is known as hinge loss or soft margin loss in the literature (Bennett and Mangasarian 1993). Like the traditionally used logistic loss, the hinge loss

Table 1 Key quantities for quantile regression (QR) and support vector classification (SVC). For any $z \in \mathbb{R}$, we define $[z]_+ = \max(0, z)$

	$\mathcal{Y}$	ζ	ψ	L_ψ
QR	$\mathbb{R}$	$\mathrm{id}_{\mathbb{R}}$	$\inf\{y : P(Y \le y \mid \mathbb{X} = \boldsymbol{x}) \ge \tau\},\ \tau \in (0,1)$	$\tau[y - f]_+ + (1-\tau)[f - y]_+$
SVC	$\{-1, 1\}$	sign	$\mathrm{sign}(P(Y = 1 \mid \mathbb{X} = \boldsymbol{x}) - 1/2)$	$[1 - yf]_+$

is a convex surrogate to the 0-1 loss in the sense of Bartlett et al. (2006). The result that the hinge loss satisfies

$$\begin{aligned}&\operatorname*{argmin}_f E_{Y|\mathbb{X}=\boldsymbol{x}}[[1 - Yf(\boldsymbol{x})]_+] \\ &\quad = \mathrm{sign}(P(Y = 1|\mathbb{X} = \boldsymbol{x}) - 1/2),\end{aligned}$$

i.e. the population minimizer is the Bayes rule, was proved in Lin (2002). This property is not shared by the logistic loss which targets the log-odds. Estimating the log-odds is little effective for regions far distant from the decision boundary, which is attributed full attention by the hinge loss. Its decision boundary is determined by usually a small fraction of the sample, the so-called support vectors. If the estimation of conditional class probabilities is secondary, these facts make SVC more attractive for classification than, e.g., logistic regression.

The most common choice for the regularizer $\Omega(\boldsymbol{\beta})$ in (1.1) is $\Omega(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2$ which in combination which the hinge loss yields the standard formulation of the soft-margin SVC and its generalizations to reproducing kernel Hilbert spaces, see the monographs of Christianini and Shawe-Taylor (2000), Schölkopf and Smola (2002) and Steinwart and Christmann (2008). QR in reproducing kernel Hilbert spaces is studied in Takeuchi et al. (2006) and Li et al. (2007). While reproducing kernel Hilbert space methods aim at more flexibility by implicitly mapping the predictor variables into a space of higher dimension, the goal of the ℓ_1- or lasso (Tibshirani 1996) regularizer $\Omega(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$ is dimension reduction by feature selection, yielding a minimizer $\widehat{\boldsymbol{\beta}}$ typically containing only a few non-zero coefficients, thus being a computationally tractable alternative to best subset selection if the number of predictors p is large. In conjunction with the hinge loss, the resulting classifier is known as linear programming machine (Bradley and Mangasarian 1998). In this paper, we equip QR and SVC with the structured elastic net, a combined ℓ_1–ℓ_2 regularizer originally introduced and studied with a focus on simple exponential family models for the response in Slawski et al. (2010).

Additional strong motivation for studying QR and SVC is given by the applicability of piecewise linear regularized solution path algorithms (Rosset and Zhu 2007), which tremendously facilitate model selection. This issue is studied in detail in Sect. 3.

The rest of this paper is organized as follows: Sect. 2 recalls the essential ideas of the structured elastic net regularizer and illustrates the key methodological concepts of the paper by means of an application to feature extraction in image classification. Section 4 contains two further data examples. We conclude with a short discussion in Sect. 5.

2 The structured elastic net regularizer

Apart from few modifications, the material of Sect. 2.1 can be found in Slawski et al. (2010); it is included here to achieve a self-contained treatment.

2.1 Double regularization

While there exists a large body of work about optimality properties of the lasso, several modifications have been suggested with the purpose to make ℓ_1-regularization behave more effectively in certain scenarios. For example, one important task in gene expression analysis is to identify a set of genes predictive for some phenotype. Concerning this problem, Zou and Hastie (2005) argue that the 'elastic net'-regularizer $\Omega(\boldsymbol{\beta}) = \alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\boldsymbol{\beta}\|_2^2$, $\alpha \in (0,1)$ possesses a crucial advantage over the lasso: the combined ℓ_1–ℓ_2 regularizer allows one to select whole groups of strongly correlated predictors, thereby enabling the scientist to reveal potentially interacting genes in the context of biological pathways. In a similar spirit, Bondell and Reich (2008) complement ℓ_1-regularization with a second regularizer which additionally clusters the set of predictors on the basis of their estimated regression coefficients. A second line of research has aimed at the explicit inclusion of structural knowledge about the predictors. One simple form of such structural knowledge is an order relation $X_j \prec X_{j'} \Leftrightarrow j < j'$, which is satisfied, e.g., when the predictors represent a signal sampled at different time points. This situation is addressed by the 'fused lasso' procedure of Tibshirani et al. (2005) who propagate

$$\Omega(\boldsymbol{\beta}) = \alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\boldsymbol{D}\boldsymbol{\beta}\|_1, \quad \alpha \in (0,1),$$

where

$$\begin{aligned}\boldsymbol{D}: \quad & \mathbb{R}^p \to \mathbb{R}^{p-1} \\ & (\beta_1, \ldots, \beta_p)^\top \mapsto ([\beta_2 - \beta_1], \ldots, [\beta_p - \beta_{p-1}])^\top\end{aligned} \tag{2.1}$$

is the first forward difference operator. Adding the total variation $\|\boldsymbol{D}\boldsymbol{\beta}\|_1$ of $\boldsymbol{\beta}$ as second regularizer yields a sequence

of estimators $\widehat{\beta}_1, \ldots, \widehat{\beta}_p$, which is piecewise constant. Although the fused lasso simplifies interpretation, because it automatically clusters the set of predictors, one can argue that it yields a sequence $\widehat{\beta}_1, \ldots, \widehat{\beta}_p$, which is too rough in the sense that large deviations $|\beta_j - \beta_{j-1}|$, $j = 2, \ldots, p$, are penalized less severely as in the case one would take $\|\boldsymbol{D}\boldsymbol{\beta}\|_2^2$ instead of $\|\boldsymbol{D}\boldsymbol{\beta}\|_1$. This prompts the introduction of the structured elastic net regularizer

$$\Omega(\boldsymbol{\beta}) = \alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\boldsymbol{\beta}^\top \boldsymbol{\Lambda}\boldsymbol{\beta}, \quad \alpha \in (0,1), \tag{2.2}$$

where $\boldsymbol{\Lambda}$ is a $p \times p$ symmetric and positive semidefinite matrix. Setting $\boldsymbol{\Lambda} = \boldsymbol{D}^\top \boldsymbol{D}$ yields the sum of the squared forward differences as second regularizer. In general, the two ingredients of the structured elastic net regularizer are differences of pairs of coefficients and a weight function $w : \{1, \ldots, p\}^2 \to \mathbb{R}$ satisfying $w(j, j) = 0$ and $w(j, j') = w(j', j)$ for all j, j'. The generic form of the regularizer is then given by

$$\begin{aligned} &\frac{1}{2}\sum_{j=1}^{p}\sum_{j'=1}^{p} |w(j,j')|(\beta_j - \operatorname{sign}\{w(j,j')\}\beta_{j'})^2 \\ &\quad = \boldsymbol{\beta}^\top \boldsymbol{\Lambda}\boldsymbol{\beta}, \end{aligned} \tag{2.3}$$

where $\boldsymbol{\Lambda}$ has entries

$$l_{jj'} = \begin{cases} \sum_{k=1}^{p} |w(j,k)| & \text{if } j = j', \\ -w(j,j') & \text{otherwise,} \end{cases} \qquad j, j' = 1, \ldots, p. \tag{2.4}$$

One may think of $|w(j, j')|$ as the strength of the prior association of the predictors j, j' such that $|w(j, j')| \to \infty$ enforces $\beta_j = \operatorname{sign}(w(j, j'))\beta_{j'}$. If $w(j, j') \geq 0$ for all j, j', it is not hard to verify that constant vectors are contained in the null space of $\boldsymbol{\Lambda}$. The regularizer (2.3) is versatile since it can handle any kind of neighborhood relation, in particular spatial or temporal ones. Combined with an ℓ_1 regularizer, we aim at the selection of relevant substructures. The structured elastic net regularizer covers the graph-constrained estimation procedure of Li and Li (2010) as well as the correlation-based penalization method of Tutz and Ulbricht (2009), see also El Anbari and Mkhadri (2008), as special cases. These references also point to interesting applications in which, as opposed to those in this paper, $\boldsymbol{\Lambda}$ is not constructed from temporal or spatial neighborhood information.

We would like to clarify that the form of regularization employed in this paper is fundamentally different from the group lasso (Yuan and Lin 2006) or the composite absolute penalty family (Zhao et al. 2009) in which groups are fixed prior to estimation, whereas the structured elastic net only encourages grouping of predictors according to specific structural prior knowledge.

2.2 Illustration

This subsection is intended to illustrate the main concepts of the paper by means of simulated data mimicking a functional magnetic resonance imaging (FMRI) experiment in a simplistic way. FMRI is a technology widely used in human brain mapping, an area of research where scientists try to associate each brain region with the primary task it is responsible for. This can be achieved by identifying those brain regions which become activated as reaction to specific experimental stimuli. The output of FMRI is a signal sampled on a three-dimensional grid of voxels, each corresponding to one small section of the brain. The fourth dimension of this kind of data is time, since periods with and without stimuli alternate. In the artificial example presented here, the problem is only two-dimensional, because only one time point and one slice of the brain are considered. Suppose we are given ten different scans of the same brain region, five corresponding to an active state, at which the experimental stimulus is present, and the remaining five to an inactive state. The task is to construct a system discriminating between scans of the two states and extracting brain regions with high discriminatory power, thereby potentially locating the regions showing a reaction to the experimental stimulus. With the notation of Sect. 1, we have $\mathcal{Y} = \{-1, 1\}$ with 1 corresponding to the active state and -1 to the inactive state, respectively. Each scan be seen as a realization of a random vector $\mathbb{X} = (X_{\boldsymbol{t}})_{\boldsymbol{t}\in D}$, where D is a finite set of points contained in a bounded, connected subset of $\mathbb{R}^2$. The data displayed in Fig. 1 are generated according to the model

$$\begin{aligned} &\mathbb{X}|Y = y \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Sigma} + \tau^2 \boldsymbol{I}), \\ &\boldsymbol{\mu}_{-1} = (0, \ldots, 0)^\top, \qquad \boldsymbol{\mu}_1 = (\mu_{\boldsymbol{t}})_{\boldsymbol{t}\in D}, \\ &\mu_{\boldsymbol{t}} = \begin{cases} c_{\boldsymbol{t}} & \boldsymbol{t} \in \mathcal{A}, \\ 0 & \text{otherwise.} \end{cases} \qquad \boldsymbol{t} \in D, \end{aligned}$$

where the $c_{\boldsymbol{t}} > 0$ and $\mathcal{A} \subset D$ is the set of points in D belonging to the brain region responding to the experimental stimulus and the symmetric positive definite matrix $\boldsymbol{\Sigma}$ has entries

$$\begin{aligned} &\gamma(\|\boldsymbol{t} - \boldsymbol{t}'\|), \quad \boldsymbol{t}, \boldsymbol{t}' \in D, \\ &\gamma(h) = \frac{1}{2^{\kappa-1}\Gamma(\kappa)}\left(\frac{h}{\phi}\right)^{\kappa} K_\kappa\left(\frac{h}{\phi}\right). \end{aligned}$$

The function γ is the covariance function of the Matérn family depending on a smoothness parameter $\kappa > 0$ and a range parameter $\phi > 0$ (Stein 1999). K_κ denotes the modified Bessel function of the third kind of order κ. To be specific, the covariance parameters are chosen as $\tau^2 = 1$, $\kappa = \frac{3}{2}$ and $\phi = \sqrt{2}$. The shape of the grid D and the activation region $\mathcal{A}$ are taken from the dataset `brain` contained

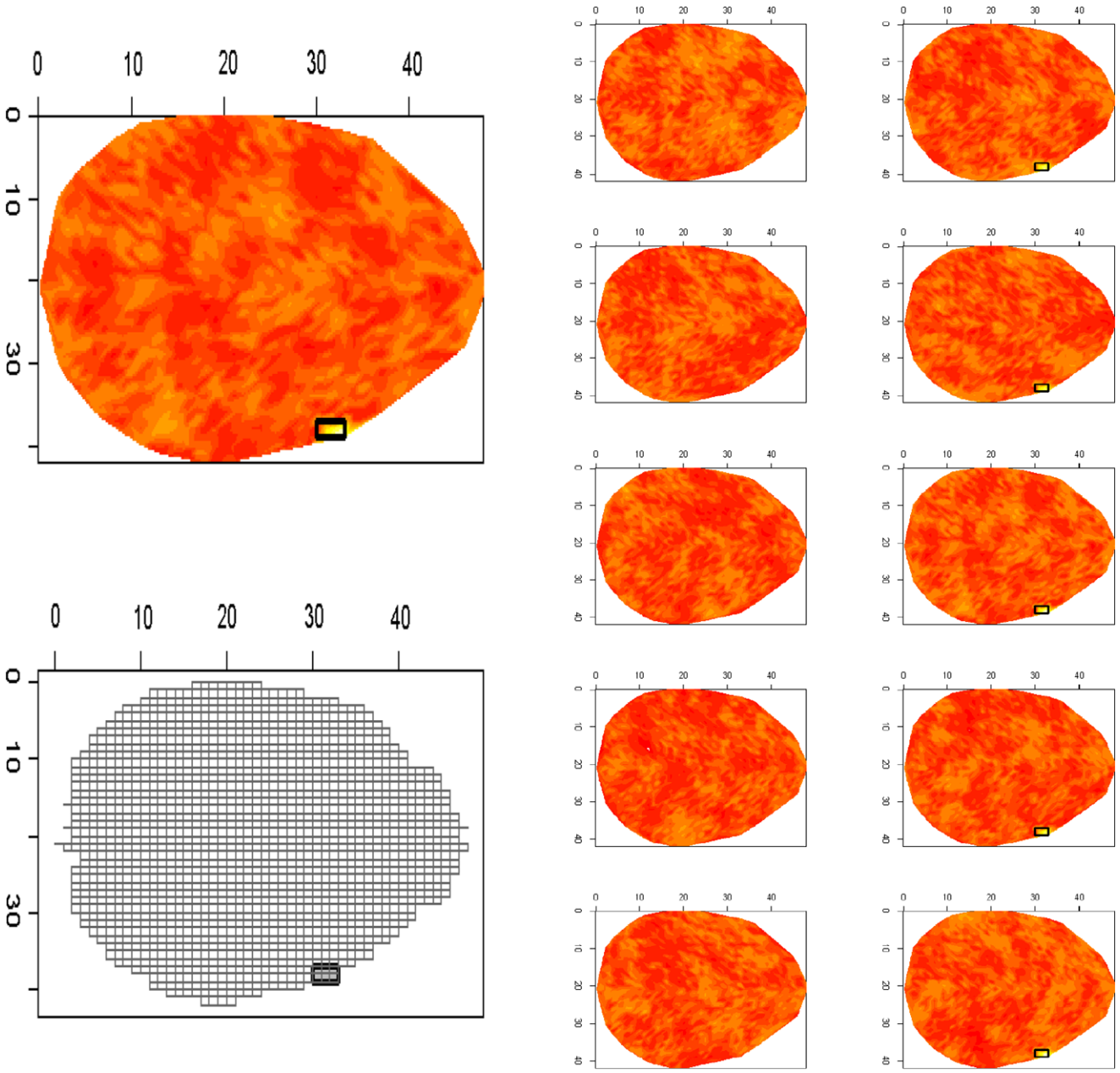

Fig. 1 The artificial FMRI data generated according to the model described in the text. The ten scans are displayed on the *right half* of the *figure*. The five scans with stimulus are shown in the *rightmost column*, and the scan of the *top right column* is zoomed at in the *upper left half* of the *figure*. It is characterized by one eminent *bright spot*, which is the region of activation $\mathcal{A}$. The shape of the grid D is depicted in the *lower left half* of the *figure*, in which $\mathcal{A}$ is marked by a *rectangle*

in the `R` package `gamair` (Wood 2006). These data were originally analyzed in Landau et al. (2003). We address the joint classification-feature selection problem using SVC endowed with the structured elastic net regularizer based on the weight function

$$w(\boldsymbol{t},\boldsymbol{t}') = \begin{cases} 1 & \text{if } t_1 = t_1' \text{ and } t_2' \in \{t_2 - 1, t_2 + 1\} \\ & \text{or } t_2 = t_2' \text{ and } t_1' \in \{t_1 - 1, t_1 + 1\}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.5}$$

The resulting neighborhood graph is depicted in the lower left panel of Fig. 1. The regression coefficients $\widehat{\beta}_{\boldsymbol{t}}, \widehat{\beta}_{\boldsymbol{t}'}$ of adjacent pixels ($w(\boldsymbol{t},\boldsymbol{t}') = 1$) are hence enforced to be similar, thus incorporating a prior assumption of spatial smoothness. The ℓ_1 part of the structured elastic net takes the sparse nature of the problem into account, since it is known that the stimulus only affects certain regions of the brain.

The combination of two regularizers comes at the expense of the difficulty to choose two tuning parameters

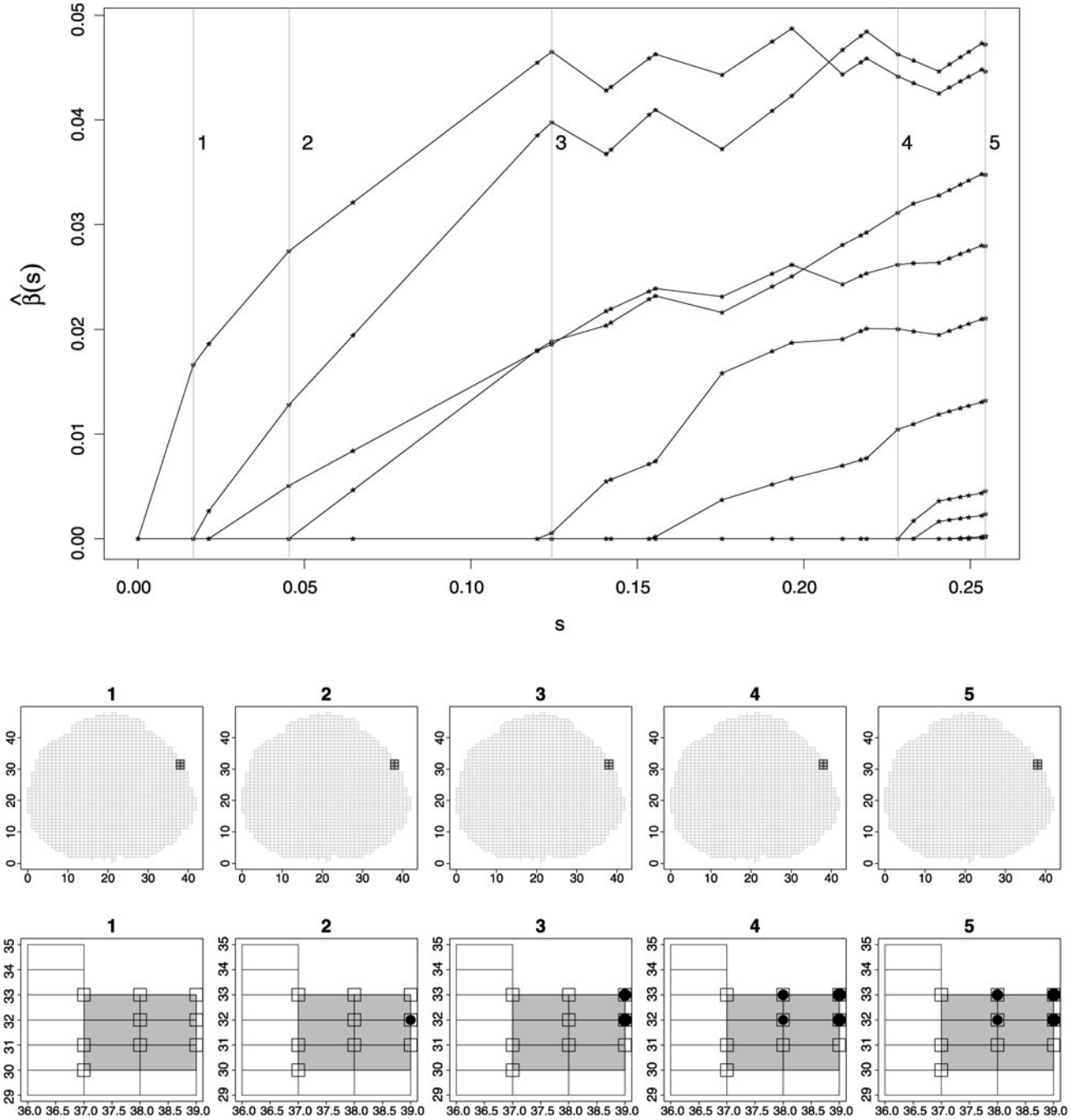

Fig. 2 Solution path and coefficient surfaces $\{(\boldsymbol{t}, \widehat{\beta}_t), \boldsymbol{t} \in D\}$ for the artificial FMRI dataset when the amount of regularization in $\boldsymbol{\beta}^\top \boldsymbol{\Lambda} \boldsymbol{\beta}$ is kept fixed and the amount of regularization in $\|\boldsymbol{\beta}\|_1$ decreases from *left* to *right*. The piecewise linear path is in $s = \|\boldsymbol{\beta}\|_1$. Each breakpoint of the piecewise linear path is marked by a *star*. The five numbered *vertical lines* indicate the choices of s for which a coefficient surface is displayed at two different levels of resolution. The size of the dots is proportional to the size of the coefficient at the respective location. While the *upper row* displays the whole grid D, the *lower row* zooms at the region $\mathcal{A}$ colored in *grey*. The places where $\boldsymbol{\mu}_1$ is different from zero are emphasized by *white squares*

which in general entails a two-dimensional grid search. However, when using one of the two loss functions considered in this paper, a significant simplification arises from the fact that for one of the two tuning parameters kept fixed, the whole range of solutions can be obtained by tracking a piecewise linear solution path (Rosset and Zhu 2007). As a consequence, a reduction to a one-dimensional grid search is possible. More specifically, for k-fold cross-validation, one specifies a grid of values to be explored for one of the two tuning parameters; in an outer loop, one runs through

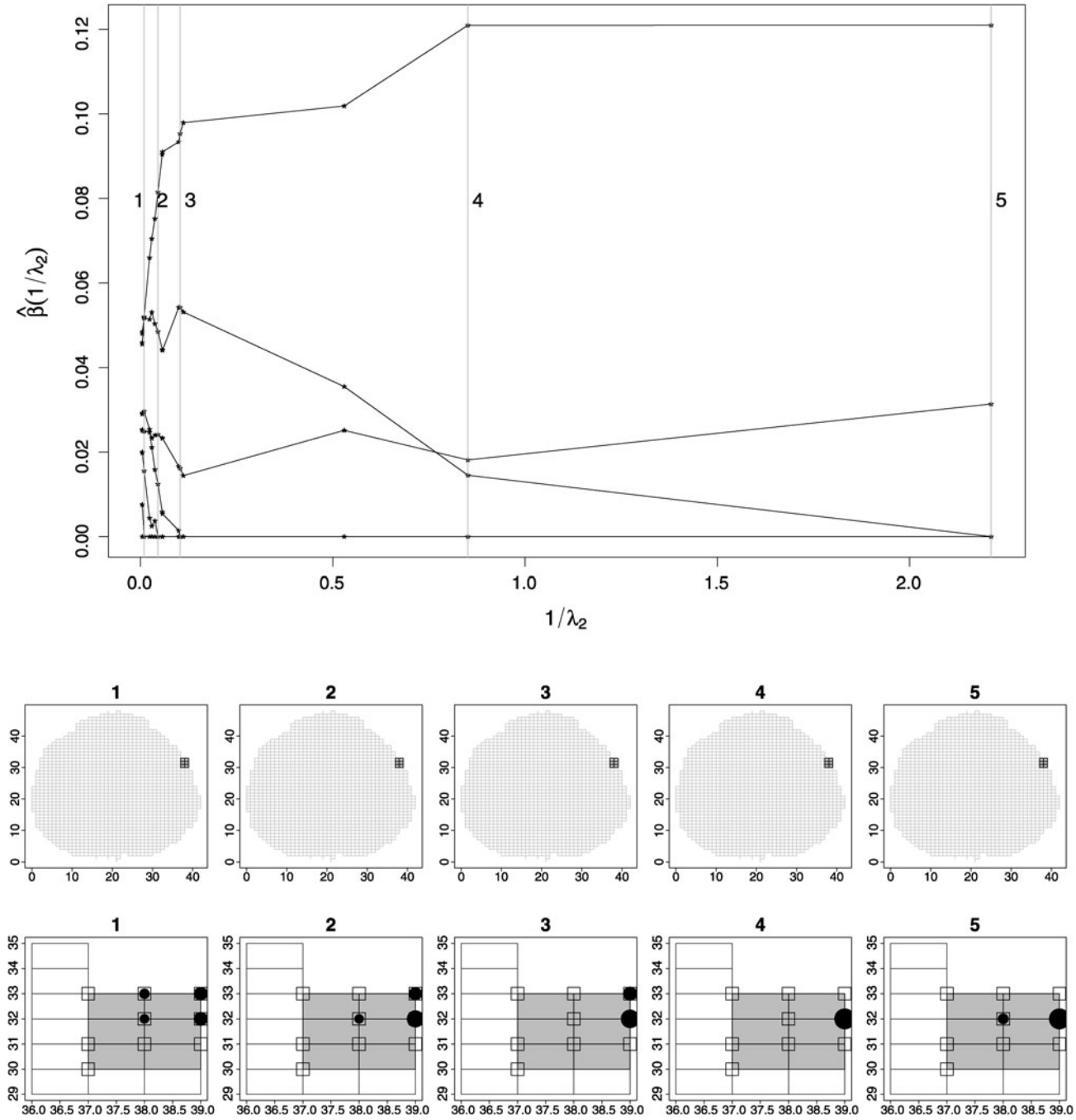

Fig. 3 Solution path and coefficient surfaces for the artificial FMRI dataset when the amount of regularization in $\|\boldsymbol{\beta}\|_1$ is kept fixed and the amount of regularization in $\boldsymbol{\beta}^\top \boldsymbol{\Lambda} \boldsymbol{\beta}$ decreases from *left* to *right*. The piecewise linear path is in $1/\lambda_2 = 1/(\lambda(1-\alpha))$. The structure of the *figure* is identical to that of Fig. 2

the different folds, fixing a training and a test sample (the sample hold out) each time. In an inner loop, one runs through the grid points. Having one of the two parameters fixed within each inner iteration, a solution path of the second parameter is computed. Figures 2 and 3 show two selected solution paths and the corresponding coefficient surfaces for the artificial FMRI dataset. The two figures clearly demonstrate the influence as well as the interplay of the two regularizers. The ℓ_1-regularizer keeps most coefficients down at zero, thereby highlighting locations in the region that matters. Figure 3 shows that when reducing the amount of quadratic regularization, one obtains a stronger concentration on single locations, which reveals that the quadratic regularizer is indeed beneficial

when the selection of whole neighborhoods of pixels is desired.

3 Solution path algorithms

The two path-following algorithms presented in the next two subsections rely on methodology developed in a series of papers discussing solution paths for the case that either an ℓ_1-regularizer or a quadratic regularizer is employed. More specifically, solution paths for ℓ_1-regularized SVC and QR are described in Zhu et al. (2003) and Li and Zhu (2008), respectively, while the second alternative is studied in Hastie et al. (2004) and Li et al. (2007), respectively. The approach of this paper is closely related to the work of Wang et al. (2006) who treat SVC with the elastic net regularizer. The authors prove that as long as the amount of ℓ_1- or ℓ_2-regularization is kept fixed, the construction of piecewise linear solution paths is possible, which coincides with the approach pursued in this paper.

In view of the fact that the solution paths for SVC and QR parallel each other, we unify their description for the sake of brevity, at the cost of a slightly more involved notation. For better guidance, Fig. 4 displays most of it at a glance. Throughout this section, we use the following notational conventions. For a vector $\boldsymbol{a} = (a_j)_{j=1,\ldots,N}$, a matrix $\boldsymbol{A} = (a_{jj'})_{\substack{j=1,\ldots,N,\\ j'=1,\ldots,N'}}$, index sets $J \subseteq \{1,\ldots,N\}$ and $J' \subseteq \{1,\ldots,N'\}$ we define

$$\boldsymbol{a}_J = (a_j)_{j\in J}, \qquad \boldsymbol{A}_J = (a_{jj'})_{j\in J},$$
$$\boldsymbol{A}^{J'} = (a_{jj'})_{j'\in J'}, \qquad \boldsymbol{A}_J^{J'} = (a_{jj'})_{j\in J, j'\in J'}.$$

3.1 Solution path in $\|\boldsymbol{\beta}\|_1$

In this subsection, we will show that the function

$$(\widehat{\beta}_0(s), \widehat{\boldsymbol{\beta}}(s)) = \operatorname*{argmin}_{\beta_0,\boldsymbol{\beta}} L_\psi(y_i, f(\boldsymbol{x}_i;\beta_0,\boldsymbol{\beta})) + \frac{\lambda_2}{2}\boldsymbol{\beta}^\top\boldsymbol{\Lambda}\boldsymbol{\beta}, \quad \lambda_2 > 0 \tag{3.1}$$
$$\text{subject to } \|\boldsymbol{\beta}\|_1 \le s$$

is well-defined and piecewise linear if L_ψ is chosen as one of the loss functions of Table 1 and regularity Assumption 1 given below holds. Note that above formulation is equivalent to a structured elastic net problem in the sense that for each s in (3.1) there exist a corresponding Lagrangian multiplier $\lambda_1(s)$, which will be shown to depend in a piecewise constant manner on s, such that

$$\lambda(s) = \lambda_1(s) + \frac{\lambda_2}{2}, \qquad \alpha(s) = \frac{\lambda_1(s)}{\lambda_1(s) + \frac{\lambda_2}{2}}.$$

In the sequel, we tend to suppress dependence on s whenever it is clear from the context.

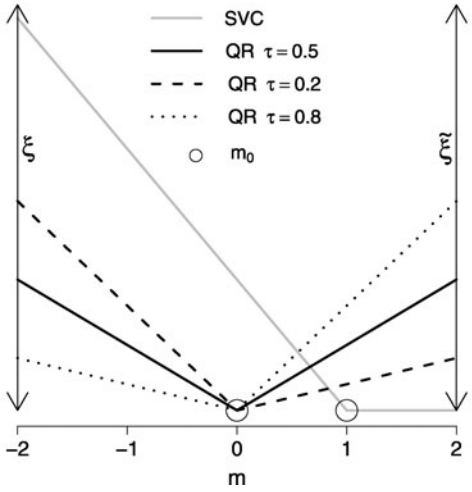

Fig. 4 Graphical overview on the notation used in Sects. 3.1 and 3.2. The *figure* sketches the hinge loss of SVC and the check loss of QR for three different choices of τ as functions of the margin m

3.1.1 KKT conditions

Before developing the algorithm, we derive the Karush-Kuhn-Tucker (KKT) conditions of the convex optimization problem which is reformulated as the following quadratic program:

$$\text{Minimize } (1-\tau)\sum_{i=1}^n \xi_i + \tau\sum_{i=1}^n \widetilde{\xi}_i + \frac{\lambda_2}{2}\boldsymbol{\beta}^\top\boldsymbol{\Lambda}\boldsymbol{\beta}$$
$$\text{subject to}$$
$$\text{(C.1)}: \|\boldsymbol{\beta}\|_1 \le s, \tag{3.2}$$
$$\text{(C.2)}: m_i + \xi_i \ge m_0, \quad i = 1,\ldots,n,$$
$$\text{(C.3)}: m_i - \widetilde{\xi}_i \le m_0, \quad i = 1,\ldots,n,$$
$$\text{(C.4)}: \xi_i,\ \widetilde{\xi}_i \ge 0, \quad i = 1,\ldots,n,$$

where for QR, we have

$$m_i = y_i - f(\boldsymbol{x}_i;\beta_0,\boldsymbol{\beta}), \qquad m_0 = 0,$$
$$\xi_i = [f(\boldsymbol{x}_i;\beta_0,\boldsymbol{\beta}) - y_i]_+,$$
$$\widetilde{\xi}_i = [y_i - f(\boldsymbol{x}_i;\beta_0,\boldsymbol{\beta})]_+, \quad i = 1,\ldots,n.$$

On the other hand, for SVC, the relevant quantities are given by

$$m_i = m_C(y_i, f(\boldsymbol{x}_i;\beta_0,\boldsymbol{\beta})) = y_i f(\boldsymbol{x}_i;\beta_0,\boldsymbol{\beta}),$$
$$m_0 = 1, \qquad \xi_i = [1-m_i]_+, \quad i = 1,\ldots,n,$$

and we require further that $\tau = 0$ such that the set of auxiliary variables $\{\widetilde{\xi}_i\}_{i=1}^n$ and 'constraint' (C.3) are in fact not needed.

The primal Lagrangian of the program (3.2) has the form

$$\begin{aligned}&(1-\tau)\sum_{i=1}^{n}\xi_i+\tau\sum_{i=1}^{n}\widetilde{\xi}_i\\&+\lambda_1(\|\boldsymbol{\beta}\|_1-s)+\frac{\lambda_2}{2}\boldsymbol{\beta}^\top\boldsymbol{\Lambda}\boldsymbol{\beta}\\&+\sum_{i=1}^{n}\gamma_i(m_i-m_0-\widetilde{\xi}_i)\\&-\sum_{i=1}^{n}\eta_i(m_i-m_0+\xi_i)\\&-\sum_{i=1}^{n}\mu_i\xi_i-\sum_{i=1}^{n}\widetilde{\mu}_i\widetilde{\xi}_i,\end{aligned}\tag{3.3}$$

where $\{\eta_i\}_{i=1}^n$, $\{\gamma_i\}_{i=1}^n$, $\{\mu_i\}_{i=1}^n$, $\{\widetilde{\mu}_i\}_{i=1}^n$ and λ_1 are non-negative Lagrangian multipliers. Denote the active set by $\mathcal{V}=\{j:\widehat{\beta}_j\neq 0\}$. Differentiating the primal Lagrangian (3.3) with respect to β_0, $\boldsymbol{\beta}_{\mathcal{V}}$, $\boldsymbol{\xi}=(\xi_i)_{i=1,\ldots,n}$, $\widetilde{\boldsymbol{\xi}}=(\widetilde{\xi}_i)_{i=1,\ldots,n}$ and setting the result equal to zero, we obtain the following KKT conditions:

$$\begin{aligned}&\lambda_2\boldsymbol{\Lambda}_{\mathcal{V}}^{\mathcal{V}}\widehat{\boldsymbol{\beta}}_{\mathcal{V}}+\lambda_1\boldsymbol{\sigma}(\mathcal{V})-[\boldsymbol{X}^{\mathcal{V}}]^\top\boldsymbol{\theta}=\mathbf{0}_{|\mathcal{V}|},\\&\boldsymbol{\theta}^\top\mathbf{1}_n=0,\\&(1-\tau)\mathbf{1}_n-\boldsymbol{\eta}-\boldsymbol{\mu}=\mathbf{0}_n,\\&\tau\mathbf{1}_n-\boldsymbol{\gamma}-\widetilde{\boldsymbol{\mu}}=\mathbf{0}_n,\end{aligned}\tag{3.4}$$

$$\begin{aligned}&\boldsymbol{X}=((\boldsymbol{x}_i)_j)_{i=1,\ldots,n,\ j=1,\ldots,p},\\&\boldsymbol{\sigma}(\mathcal{V})=(\operatorname{sign}(\widehat{\beta}_j))_{j\in\mathcal{V}},\\&\boldsymbol{\eta}=(\eta_i)_{i=1,\ldots,n},\\&\boldsymbol{\gamma}=(\gamma_i)_{i=1,\ldots,n},\end{aligned}\tag{3.5}$$

and

$$\boldsymbol{\theta}=(\theta_i)_{i=1,\ldots,n}=\begin{cases}\boldsymbol{\gamma}-\boldsymbol{\eta} & \text{for QR},\\ \operatorname{diag}(y_1,\ldots,y_n)\boldsymbol{\eta} & \text{for SVC},\end{cases}$$

noting that $\tau=0$ implies that $\boldsymbol{\gamma}=\widetilde{\boldsymbol{\mu}}=\mathbf{0}_n$. The KKT conditions additionally include the constraints

$$\begin{aligned}&\gamma_i(m_i-m_0-\widetilde{\xi}_i)=0,\\&\eta_i(m_i-m_0+\xi_i)=0,\\&\mu_i\xi_i=0,\\&\widetilde{\mu}_i\widetilde{\xi}_i=0,\quad i=1,\ldots,n.\end{aligned}\tag{3.6}$$

These constraints allow for a division of the elements of the sample S into three subsets. Combining information in (3.2), (3.4) and (3.6), we have the following implications.

$$\begin{aligned}&(1)\quad m_i<m_0\Rightarrow\xi_i>0,\widetilde{\xi}_i=0\Rightarrow\mu_i=0,\\&\qquad\gamma_i=0\Rightarrow\theta_i=\begin{cases}-(1-\tau) & \text{for QR},\\ \pm1 & \text{for SVC},\end{cases}\\&(2)\quad m_i=m_0\Rightarrow\xi_i=\widetilde{\xi}_i=0\Rightarrow\mu_i,\\&\qquad\widetilde{\mu}_i\geq0\Rightarrow\eta_i\leq1-\tau,\\&\qquad\gamma\leq\tau\Rightarrow\begin{cases}-(1-\tau)\leq\theta_i\leq\tau & \text{for QR},\\ -1\leq\theta_i\leq1 & \text{for SVC}.\end{cases}\\&(3)\quad m_i>m_0\Rightarrow\widetilde{\xi}_i>0,\ \xi_i=0\Rightarrow\widetilde{\mu}_i=0,\\&\qquad\eta_i=0\Rightarrow\theta_i=\tau,\ i=1,\ldots,n.\end{aligned}\tag{3.7}$$

We introduce sets $\mathcal{L}=\{i:m_i<m_0\}$, $\mathcal{E}=\{i:m_i=m_0\}$, $\mathcal{R}=\{i:m_i>m_0\}$ and define

$$\Theta(\mathcal{L})=\begin{cases}-(1-\tau) & \text{for QR},\\ \pm1 & \text{for SVC},\end{cases}\qquad\Theta(\mathcal{R})=\tau.\tag{3.8}$$

As conclusion which turns out to be essential for the next two subsections, we obtain that for elements in S belonging to the set $\mathcal{L}$ and $\mathcal{R}$, we know the values of the dual variables θ_i, while for the elements in $\mathcal{E}$, the θ_i are between $\Theta(\mathcal{L})$ and $\Theta(\mathcal{R})$ and further that their margins m_i are all equal to the value m_0, which corresponds to the only non-smooth point of L_ψ as a function of f. The main principle of the algorithm is to identify changes in the sets $\mathcal{E}$ and $\mathcal{V}$, which correspond to the breakpoints of the piecewise linear path. Once a breakpoint is encountered, a new direction and a stepsize for the solution path is calculated such that the KKT conditions continue to hold.

3.1.2 Regularity conditions

For the algorithm to work as formulated in Sects. 3.1.3 and 3.1.4, we require the following conditions to hold.

Assumption 1

(i) At every step of the algorithm, the matrix $\boldsymbol{X}_{\mathcal{E}}^{\mathcal{V}}$ has rank $|\mathcal{E}|$, where $\boldsymbol{X}$ is defined in (3.5).
(ii) The matrix $\boldsymbol{\Lambda}$ is positive definite.
(iii) At every step of the algorithm, precisely one of the sets $\mathcal{E}$ and $\mathcal{V}$ changes by precisely one element.

Condition (i) and (ii) guarantee that the solution of the linear system (3.12) below is unique. Note that condition (i) claims in particular that $|\mathcal{E}|\leq|\mathcal{V}|$. While we have required $\boldsymbol{\Lambda}$ to be positive semidefinite in Sect. 2, positive definiteness is achieved by adding some small value to the diagonal. The third condition is related to the 'one-at-a-time-condition' in

Efron et al. (2004). We conjecture that a relaxation of condition (iii) is possible. However, practical experience has shown that multiple changes in $\mathcal{E}$ are typically irreconcilable with condition (i).

3.1.3 Initialization

The initalization ($s = 0$) is actually the most critical part of the algorithm since in general $\widehat{\beta}_0(0)$ or the next direction the algorithm proceeds into are not determined uniquely. For QR, as pointed out by Li and Zhu (2008), $\widehat{\boldsymbol{\beta}}(0) = \mathbf{0}_p$ and $\widehat{\beta}_0(0)$ equals the sample τ-quantile of the $\{y_i\}_{i=1}^n$. The latter is unique and attained by one of the $\{y_i\}_{i=1}^n$ if and only if $n\tau$ is not an integer. Let $\{\pi(1), \ldots, \pi(n)\}$ be an index permutation of $\{1, \ldots, n\}$ such that $y_{\pi(1)} \leq \cdots \leq y_{\pi(n)}$, then we set

$$\widehat{\beta}_0(0) = y_{\widehat{i}}, \quad \widehat{i} = \pi(\lfloor n\tau \rfloor + 1) \;\Rightarrow\; \mathcal{E} = \{\widehat{i}\},$$
$$\mathcal{L} = \{i : y_i < \widehat{\beta}_0(0)\}, \qquad \mathcal{R} = \{i : y_i > \widehat{\beta}_0(0)\},$$

where we exclude ties in $\{y_i\}_{i=1}^n$. Otherwise,

$$\widehat{\beta}_0(0) \in [y_{\widehat{l}}, y_{\widehat{r}}], \quad \widehat{l} = \pi(\lfloor n\tau \rfloor), \; \widehat{r} = \pi(\lfloor n\tau \rfloor + 1).$$

We circumvent this issue by setting $\widehat{\beta}_0(0)$ equal to the left endpoint $y_{\widehat{l}}$, likewise $\mathcal{E} = \{\widehat{l}\}$ and accordingly $\mathcal{L}$ and $\mathcal{R}$. Equations (3.8) and (3.4) allow us to compute the $\{\theta_i\}_{i=1}^n$. The latter are in turn used to compute the *generalized correlations* $\{c_j\}_{j=1}^p$, the Lagrangian multiplier λ_1 and the active set $\mathcal{V}$ according to the first equation in the system (3.4) as

$$\begin{aligned} c_j &= \sum_{i=1}^n \theta_i (\boldsymbol{x}_i)_j, \quad j = 1, \ldots, p, \\ \lambda_1 &= \max_{j=1,\ldots,p} |c_j|, \qquad \mathcal{V} = \{j : |c_j| = \lambda_1\}. \end{aligned} \tag{3.9}$$

For the same reasons as for condition (iii) in Assumption 1, we assume that $\mathcal{V}$ consists of precisely one element.

For support vector classification, the situation is even more difficult (Zhu et al. 2003). Denoting by $I_+ = \{i : y_i = 1\}$ and by $I_- = \{i : y_i = -1\}$ the indices of the samples belonging to the positive and negative class respectively, then $\widehat{\beta}_0(0) = 1$ if $|I_+| > |I_-|$ and $\widehat{\beta}_0(0) = -1$ if $|I_-| > |I_+|$. Consequently, $\mathcal{E} = I_+$ or $\mathcal{E} = I_-$, which leads to a violation of condition (i) in Assumption 1. We resort to a workaround communicated by Ji Zhu, who suggests to add a small amount of jitter to the y_i of the majority class, generating slightly perturbed responses $\{\widetilde{y}_i\}_{i=1}^n$. Plugging these $\{\widetilde{y}_i\}_{i=1}^n$ into the optimization problem (3.2) yields a solution $\widehat{\beta}_0(0)$ such that the set $\mathcal{E}$ consists of precisely one element. We then proceed as in (3.9) to obtain λ_1 and $\mathcal{V}$. A closely related workaround is employed in Wang and Shen (2005).

3.1.4 Main algorithm

The principle of the algorithm can be summarized as follows. Given an initial solution $\boldsymbol{q}(0) = (\widehat{\beta}_0(0), \boldsymbol{\theta}(0)^\top, \lambda_1(0))^\top$ together with an initial configuration of the sets $\mathcal{L}$, $\mathcal{E}$, $\mathcal{R}$ and $\mathcal{V}$, we determine the smallest δ_s such that the structure of the KKT conditions (3.4) and (3.6) belonging to optimization problem (3.2) with a specific choice of s is different from the same problem with $s + \delta_s$. We call these optimization problems different in the structure if one of the following conditions holds:

$$\begin{aligned} &(1) \quad \mathcal{V}(s) \neq \mathcal{V}(s + \delta_s), \\ &(2) \quad \mathcal{E}(s) \neq \mathcal{E}(s + \delta_s). \end{aligned} \tag{3.10}$$

Note that, by continuity, a change in $\mathcal{E}$ precedes every transition of an index from $\mathcal{L}$ to $\mathcal{R}$. Adopting the terminology of Zhu et al. (2003) and Wang et al. (2006), changes (1) and (2) are called *events*. The next event is predicted by computing the right derivative

$$\frac{\Delta \boldsymbol{q}(s)}{\Delta s} = \lim_{\Delta s \to 0} \frac{\boldsymbol{q}(s + \Delta s) - \boldsymbol{q}(s)}{\Delta s},$$
$$\boldsymbol{q}(s) = (\widehat{\beta}_0(s), \boldsymbol{\theta}(s)^\top, \lambda_1(s))^\top,$$

from the KKT conditions and one then updates $\boldsymbol{q}(s + \delta_s) \leftarrow \boldsymbol{q}(s) + \delta_s \frac{\Delta \boldsymbol{q}(s)}{\Delta s}$. This is repeated from event to event, until a stopping criterion is met (see below). A key observation for the computation of the right derivative is that, according to the observation made in (3.7),

$$i \in \mathcal{L} \cup \mathcal{R} \quad \Rightarrow \quad \frac{\Delta \theta_i}{\Delta s} = 0.$$

The right derivative can hence be computed by solving the linear system

$$\begin{aligned} &\lambda_2 \boldsymbol{\Lambda}_{\mathcal{V}}^{\mathcal{V}} \frac{\Delta \widehat{\boldsymbol{\beta}}_{\mathcal{V}}}{\Delta s} - \left[\boldsymbol{X}_{\mathcal{E}}^{\mathcal{V}}\right]^\top \frac{\Delta \boldsymbol{\theta}_{\mathcal{E}}}{\Delta s} + \frac{\Delta \lambda_1}{\Delta s} \boldsymbol{\sigma}(\mathcal{V}) = \mathbf{0}_{|\mathcal{V}|}, \\ &\frac{\Delta \boldsymbol{\theta}_{\mathcal{E}}^\top \mathbf{1}}{\Delta s} = 0, \\ &\frac{\Delta \widehat{\beta}_0}{\Delta s} + \boldsymbol{X}_{\mathcal{E}}^{\mathcal{V}} \frac{\Delta \widehat{\boldsymbol{\beta}}_{\mathcal{V}}}{\Delta s} = \mathbf{0}_{|\mathcal{E}|}, \end{aligned} \tag{3.11}$$

where the third equation follows from $i \in \mathcal{E} \Leftrightarrow m_i = m_0$, $i = 1, \ldots, n$. Furthermore, $\frac{\Delta \lambda_1}{\Delta s}$ is a negative scalar (negativity follows from the fact that $\lambda_1(s)$ is decreasing in s), therefore we may set $\frac{\Delta \lambda_1}{\Delta s} = -1$. The linear system (3.11) consists of $|\mathcal{E}| + |\mathcal{V}| + 1$ equations and the same number of unknowns. It can be reduced to a system in $|\mathcal{E}| + 1$ unknowns by solving the first equation for $\frac{\Delta \widehat{\boldsymbol{\beta}}_{\mathcal{V}}}{\Delta s}$ and plugging the result into the third equation. This yields the following

linear system to be solved.

$$\begin{bmatrix} \lambda_2 \mathbf{1}_{|\mathcal{E}|} & \boldsymbol{G}(\mathcal{E},\mathcal{V}) \\ 0 & \mathbf{1}_{|\mathcal{E}|}^\top \end{bmatrix} \begin{bmatrix} \frac{\Delta\widehat{\beta}_0}{\Delta s} \\ \frac{\Delta\boldsymbol{\theta}_\mathcal{E}}{\Delta s} \end{bmatrix} = \begin{bmatrix} -\boldsymbol{X}_\mathcal{E}^\mathcal{V} \left[\boldsymbol{\Lambda}_\mathcal{V}^\mathcal{V}\right]^{-1} \boldsymbol{\sigma}(\mathcal{V}) \\ 0 \end{bmatrix}, \tag{3.12}$$

where

$$\boldsymbol{G}(\mathcal{E},\mathcal{V}) = \boldsymbol{X}_\mathcal{E}^\mathcal{V} \left[\boldsymbol{\Lambda}_\mathcal{V}^\mathcal{V}\right]^{-1} \left[\boldsymbol{X}_\mathcal{E}^\mathcal{V}\right]^\top. \tag{3.13}$$

Given $\frac{\Delta\boldsymbol{\theta}_\mathcal{E}}{\Delta s}$, $\frac{\Delta\widehat{\beta}_0}{\Delta s}$, we now compute the resulting derivatives for the remaining quantities:

$$\begin{aligned} \frac{\Delta\widehat{\boldsymbol{\beta}}_\mathcal{V}}{\Delta s} &= [\boldsymbol{\Lambda}_\mathcal{V}^\mathcal{V}]^{-1} \frac{[\boldsymbol{X}_\mathcal{E}^\mathcal{V}]^\top \frac{\Delta\boldsymbol{\theta}_\mathcal{E}}{\Delta s} + \boldsymbol{\sigma}(\mathcal{V})}{\lambda_2}, \\ \frac{\Delta\boldsymbol{c}_{\mathcal{V}^c}}{\Delta s} &= [\boldsymbol{X}_\mathcal{E}^{\mathcal{V}^c}]^\top \frac{\Delta\boldsymbol{\theta}_\mathcal{E}}{\Delta s} - \lambda_2 \boldsymbol{\Lambda}_{\mathcal{V}^c}^\mathcal{V} \frac{\Delta\widehat{\boldsymbol{\beta}}_\mathcal{V}}{\Delta s}, \\ \frac{\Delta\boldsymbol{r}_{\mathcal{E}^c}}{\Delta s} &= -\frac{\Delta\widehat{\beta}_0}{\Delta s} - \boldsymbol{X}_{\mathcal{E}^c}^\mathcal{V} \frac{\Delta\widehat{\boldsymbol{\beta}}_\mathcal{V}}{\Delta s}, \quad \boldsymbol{r} = \boldsymbol{\xi} + \widetilde{\boldsymbol{\xi}}. \end{aligned} \tag{3.14}$$

Note that $\xi_i\widetilde{\xi}_i = 0,\ i = 1,\ldots,n$. Together with the solution of the linear system (3.12), the right derivatives (3.14) are used to calculate the stepsize $\delta = \delta_s / \|\frac{\Delta\widehat{\boldsymbol{\beta}}_\mathcal{V}}{\Delta s}\|_1$ and to update the status sets $\mathcal{V}$, $\mathcal{E}$, $\mathcal{L}$ and $\mathcal{R}$. Defining

$$\begin{aligned} \delta_{\boldsymbol{\theta}} &= \inf\Big\{ d > 0 : \exists i \in \mathcal{E} : \\ &\qquad \theta_i + d\frac{\Delta\theta_i}{\Delta s} \in \Theta(\mathcal{L}) \cup \Theta(\mathcal{R}) \Big\}, \\ \delta_{\boldsymbol{r}} &= \inf\Big\{ d > 0 : \exists i \in \mathcal{E}^c : r_i + d\frac{\Delta r_i}{\Delta s} = 0 \Big\}, \\ \delta_{\widehat{\boldsymbol{\beta}}} &= \inf\Big\{ d > 0 :\ \exists j \in \mathcal{V} :\ \widehat{\beta}_j + d\frac{\Delta\widehat{\beta}_j}{\Delta s} = 0 \Big\}, \\ \delta_{\boldsymbol{c}} &= \inf\Big\{ d > 0 : \exists j \in \mathcal{V}^c : \\ &\qquad \Big| c_j + d\frac{\Delta c_j}{\Delta s} \Big| = \lambda_1 - d \Big\}, \end{aligned} \tag{3.15}$$

we have that $\delta = \min\{\delta_{\boldsymbol{\theta}}, \delta_{\boldsymbol{r}}, \delta_{\widehat{\boldsymbol{\beta}}}, \delta_{\boldsymbol{c}}, \lambda_1(s)\}$. Including $\lambda_1(s)$ guarantees that $\lambda_1(s+\delta_s) \geq 0$. If $\lambda_1(s+\delta_s) = 0$, the algorithm terminates. A further stopping criterion is that the loss in the sample drops to zero, i.e. $\forall i\, m_i(s+\delta_s) \geq m_0$ for SVC and $\forall i\ m_i(s+\delta_s) = m_0$ for QR, respectively.

3.2 Solution path in $\boldsymbol{\beta}^\top \boldsymbol{\Lambda} \boldsymbol{\beta}$

In order to examine how the solution varies for a fixed amout of ℓ_1-regularization, we proceed as follows. We first fix $\lambda_2 = \lambda_2^{\text{init}}$, say, and compute the entire solution path $(\widehat{\beta}_0(s), \widehat{\boldsymbol{\beta}}(s))$ according to the previous subsection. For each s, we then know the optimum values $\boldsymbol{\theta}(s)$, $\widehat{\boldsymbol{\beta}}(s)$, $\lambda_1(s)$ as well as the status sets $\mathcal{E}(s)$ and $\mathcal{V}(s)$. For any choice of s, we may subsequently trace the solution path $(\widehat{\beta}_0(\lambda_2), \widehat{\boldsymbol{\beta}}(\lambda_2))$, $0 \leq \lambda_2 \leq \lambda_2^{\text{init}}$. From the first equation of the KKT conditions (3.4), one obtains

$$\lambda_2 \widehat{\boldsymbol{\beta}}_\mathcal{V} = [\boldsymbol{\Lambda}_\mathcal{V}^\mathcal{V}]^{-1} \left([\boldsymbol{X}^\mathcal{V}]^\top \boldsymbol{\theta} - \lambda_1 \boldsymbol{\sigma}(\mathcal{V}) \right).$$

Next, we use that $m_i = m_0$ for all $i \in \mathcal{E}$, which, after premultiplying λ_2, can be written as

$$\begin{aligned} &\theta_0 \mathbf{1}_{|\mathcal{E}|} + \boldsymbol{X}_\mathcal{E}^\mathcal{V} [\boldsymbol{\Lambda}_\mathcal{V}^\mathcal{V}]^{-1} \left([\boldsymbol{X}_\mathcal{E}^\mathcal{V}]^\top \boldsymbol{\theta}_\mathcal{E} - \lambda_1 \boldsymbol{\sigma}(\mathcal{V}) \right) = \lambda_2 \boldsymbol{y}_\mathcal{E}, \\ &\theta_0 = \lambda_2 \widehat{\beta}_0. \end{aligned}$$

Differentiating both sides w.r.t. λ_2 yields

$$\frac{\Delta\theta_0}{\Delta\lambda_2} \mathbf{1}_{|\mathcal{E}|} + \boldsymbol{G}(\mathcal{E},\mathcal{V}) \frac{\Delta\boldsymbol{\theta}_\mathcal{E}}{\Delta\lambda_2} = \boldsymbol{y}_\mathcal{E},$$

where the matrix $\boldsymbol{G}(\mathcal{E},\mathcal{V})$ is defined in (3.13). We additionally know that (cf. (3.11))

$$\sum_{i \in \mathcal{E}} \frac{\Delta\theta_i}{\Delta\lambda_2} = 0.$$

Altogether, $\frac{\Delta\theta_0}{\Delta\lambda_2}$, $\frac{\Delta\boldsymbol{\theta}_\mathcal{E}}{\Delta\lambda_2}$ can be computed by solving the linear system

$$\begin{bmatrix} \mathbf{1}_{|\mathcal{E}|} & \boldsymbol{G}(\mathcal{E},\mathcal{V}) \\ 0 & \mathbf{1}_{|\mathcal{E}|}^\top \end{bmatrix} \begin{bmatrix} \frac{\Delta\theta_0}{\Delta\lambda_2} \\ \frac{\Delta\boldsymbol{\theta}_\mathcal{E}}{\Delta\lambda_2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_\mathcal{E} \\ 0 \end{bmatrix}.$$

It remains to compute the stepsize δ_{λ_2} such that $\lambda_2^{\text{new}} = \lambda_2 - \delta_{\lambda_2}$, where $\lambda_2^{\text{new}} < \lambda_2$ corresponds to the next breakpoint of the solution path. As in Sect. 3.1, the stepsize δ_{λ_2} is computed as the smallest value such that the next event occurs. Analogously to (3.15), we compute

$$\begin{aligned} \delta'_{\boldsymbol{\theta}} &= \inf\left\{ d > 0 : \exists i \in \mathcal{E} : \theta_i - d\frac{\Delta\theta_i}{\Delta\lambda_2} \in \Theta(\mathcal{L}) \cup \Theta(\mathcal{R}) \right\}, \\ \delta'_{\boldsymbol{r}} &= \inf{}^+\left\{ \frac{\theta_0 + \boldsymbol{x}_i^\top [\boldsymbol{\Lambda}_\mathcal{V}^\mathcal{V}]^{-1} ([\boldsymbol{X}^\mathcal{V}]^\top \boldsymbol{\theta} - \lambda_1 \boldsymbol{\sigma}(\mathcal{V})) - \lambda_2 y_i}{\frac{\Delta\theta_0}{\Delta\lambda_2} + \boldsymbol{x}_i^\top [\boldsymbol{\Lambda}_\mathcal{V}^\mathcal{V}]^{-1} [\boldsymbol{X}_\mathcal{E}^\mathcal{V}]^\top \frac{\Delta\boldsymbol{\theta}_\mathcal{E}}{\Delta\lambda_2} - y_i}, \right. \\ &\qquad \left. i \in \mathcal{E}^c \right\}, \\ \delta'_{\widehat{\boldsymbol{\beta}}} &= \inf{}^+\left\{ \frac{([\boldsymbol{\Lambda}_\mathcal{V}^\mathcal{V}]^{-1} [\boldsymbol{X}^\mathcal{V}]^\top \boldsymbol{\theta})_j - \lambda_1 (\boldsymbol{\sigma}(\mathcal{V}))_j}{([\boldsymbol{\Lambda}_\mathcal{V}^\mathcal{V}]^{-1} [\boldsymbol{X}_\mathcal{E}^\mathcal{V}]^\top \frac{\Delta\boldsymbol{\theta}_\mathcal{E}}{\Delta\lambda_2})_j},\ j \in \mathcal{V} \right\}, \\ \delta'_{\boldsymbol{c}} &= \inf{}^+\left\{ \frac{\pm\lambda_1 + (\boldsymbol{X}_{\mathcal{V}^c}^\top \boldsymbol{\theta} - \lambda_2 \boldsymbol{\Lambda}_{\mathcal{V}^c}^\mathcal{V} \widehat{\boldsymbol{\beta}}_\mathcal{V})_j}{([\boldsymbol{X}_\mathcal{E}^{\mathcal{V}^c}]^\top \frac{\Delta\boldsymbol{\theta}_\mathcal{E}}{\Delta\lambda_2} - \boldsymbol{\Lambda}_{\mathcal{V}^c}^\mathcal{V} [\boldsymbol{\Lambda}_\mathcal{V}^\mathcal{V}]^{-1} [\boldsymbol{X}_\mathcal{E}^\mathcal{V}]^\top \frac{\Delta\boldsymbol{\theta}_\mathcal{E}}{\Delta\lambda_2})_j}, \right. \\ &\qquad \left. j \in \mathcal{V}^c \right\}, \end{aligned}$$

where $\inf^+$ indicates that the infimum is taken over indices for which the corresponding expression within the curly brackets is positive. The stepsize is determined as $\delta_{\lambda_2} = \min\{\delta'_{\boldsymbol{\theta}}, \delta'_{\boldsymbol{r}}, \delta'_{\widehat{\boldsymbol{\beta}}}, \delta'_{\widehat{\boldsymbol{c}}}, \lambda_2\}$. The stopping criteria of the algorithm are the same as for the solution path in $\|\boldsymbol{\beta}\|_1$.

4 Data analysis

The following section presents applications of the structured elastic net in one regression and one classification problems taken from climate science and image processing, respectively.

4.1 Canadian weather data

In Chap. 15 of Ramsay and Silverman (2006), the authors illustrate their functional linear model approach by predicting the logarithm of the total annual precipitation at 35 Canadian weather stations from the temperatures measured at 365 days of a year. The data were averaged over 35 annual reports starting in 1960 and ending in 1994. In our analysis, we aim at the prediction of the 0.25-, 0.5-, and 0.75-quantile of the total annual precipitation by a linear regression on the temperature pattern, i.e. we set

$$\widehat{y}_{i,\tau} = \widehat{\beta}_0^\tau + \sum_{j=1}^{p} \widehat{\beta}_j^\tau x_{ij},$$

where $\widehat{y}_{i,\tau}$ is the prediction of the τ-quantile, $\tau \in \{0.25, 0.5, 0.75\}$ for weather station i, $i = 1, \ldots, 35$, and x_{ij} is the temperature for weather station i at day j, $j = 1, \ldots, p$, $p = 365$. The use of the structured elastic net regularizer can be motivated from the following considerations. At first, the predictors are ordered temporally such that their influence on the response as quantified by the regression coefficients is expected to be similar. Secondly, interpretation is greatly simplified if predictors with low or no influence on the response are removed. On the other hand, mere ℓ_1-regularization is not fully adequate for this purpose, since it tends to select single days. Intuitively, selection on the basis of days is an unreliable procedure with respect to the prediction of future observations. It therefore seems to be beneficial to identify relevant sequences of days, e.g. weeks or (parts of) months. One might argue that one could alternatively coarsen the data by averaging over days. However, this would be a rather wasteful treatment of the information one has at hand. The spirit of our approach is closely related to the ideas underlying the analyses of the same dataset in James et al. (2008) and Tutz and Gertheiss (2010). Both use squared loss and computationally different approaches to incorporate temporal structure and to perform variable selection. We compare the performance of ℓ_1, elastic net and the structured elastic net regularizer, where the structure matrix $\boldsymbol{\Lambda}$ is composed of squared first forward differences for adjacent days. We also compute the generalized ridge solution which only uses the quadratic part of the structured elastic net regularizer. All four approaches have in common that model selection is simplified by the availability of piecewise linear solution path algorithms. For the evaluation of the performance, we compute 50 random splits of the dataset into learning and test sample consisting of 30 and 5 observations, respectively. Hyperparameters are determined via tenfold cross-validation on the learning sets. Table 2 displays the mean loss (averages over the 50 splits) on the test sample, where 'loss' refers to the check loss (cf. Table 1) and hence depends on τ. It is seen that mere ℓ_1-regularization shows a poor performance, confirming the argumentation above. The structured elastic net performs best for $\tau \in \{0.25, 0.5\}$, but is beaten by the ordinary elastic net for $\tau = 0.75$.

Table 2 Mean loss on the random test samples, averaged over 50 iterations, for the Canadian weather dataset. Standard errors are given in parentheses. The term 'e.net' abbreviates 'elastic net'

Method	$\tau = 0.25$ loss (test sets)	$\tau = 0.5$ loss (test sets)	$\tau = 0.75$ loss (test sets)
Generalized ridge	0.276 (0.018)	0.300 (0.020)	0.290 (0.021)
Lasso	0.285 (0.029)	0.361 (0.024)	0.326 (0.022)
e.net	0.261 (0.014)	0.295 (0.014)	0.219 (0.012)
Structured e.net	0.222 (0.013)	0.265 (0.014)	0.229 (0.014)

The upper panel of Fig. 5 displays the temperature profiles of the 35 weather stations and the median (over the 50 splits) of the estimated coefficients $\widehat{\beta}_j^\tau$, $j = 1, \ldots, 365$ of the structured elastic net. According to the lower panel, the days in November carry much potential to predict the outcome, while the days in June up to October seem to be irrelevant. The latter is not surprising when looking at the upper panel: in the summer months, there is considerable overlap of the temperature profiles of weather stations with rather different annual precipitations. Too a large extent, the coefficient profile for the conditional median ($\tau = 0.5$) agrees with those displayed in James et al. (2008) and Tutz and Gertheiss (2010), whose approaches target the conditional mean. In particular, the eminent bump at the end as well as the flat in the middle of the year is a common feature.

Figure 6 displays an assessment of the fit, which appears to be decent except for few data. A particularly strong misfit results for the weather station in Kamloops, an observation already made in the analysis of Ramsay and Silverman

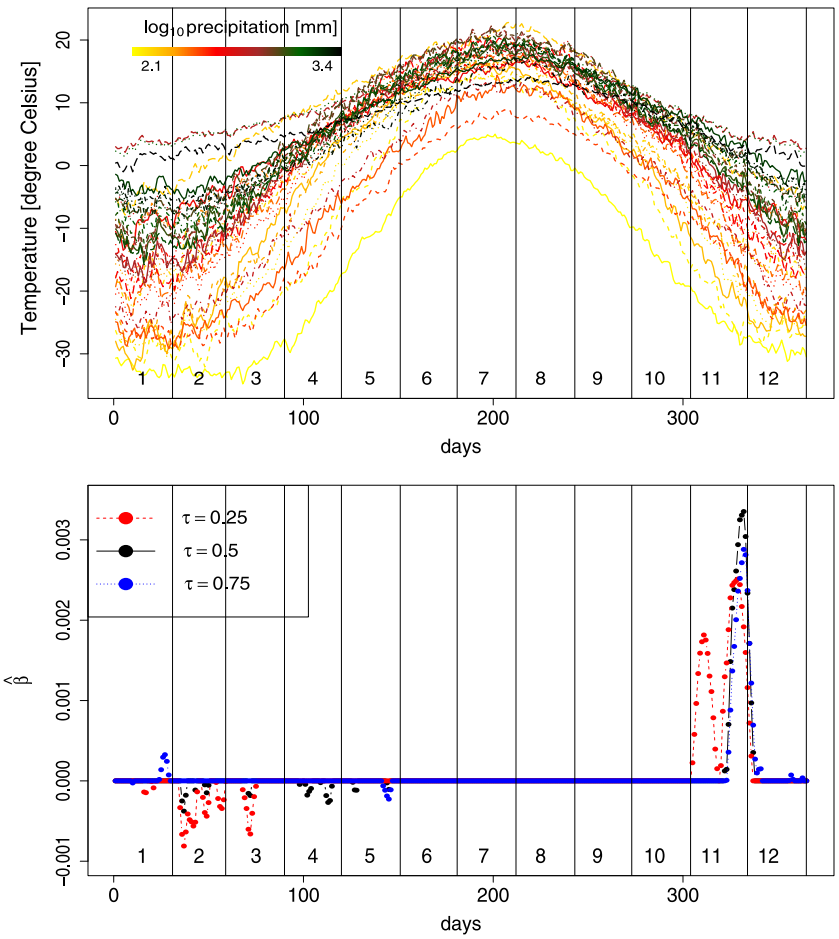

Fig. 5 (Color online) *Upper panel*: temperatures of 35 Canadian weather stations. Their annual ($\log_{10}$) precipitations can be read off from the *colour bar*. *Lower panel*: Regression coefficients of the structured elastic net for median, lower and upper quartile. In both *panels*, the *vertical lines* divide the 365 days into 12 months

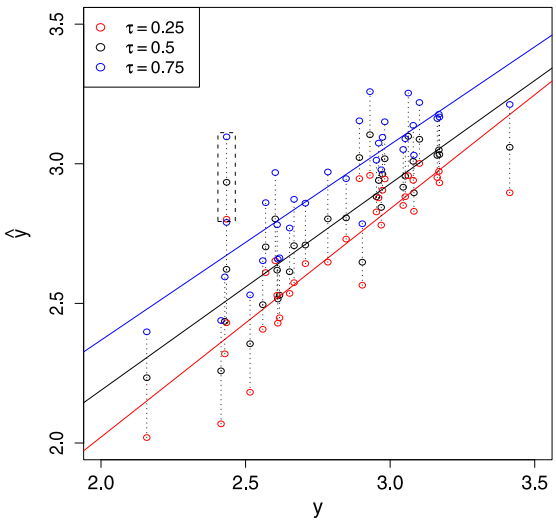

Fig. 6 Predicted quantiles ($\widehat{y}$) vs. observed values (y). The *straight lines* correspond to linear regressions of the fitted values on the observed values, separately for each $\tau \in \{0.25, 0.5, 0.75\}$. The $(y, \widehat{y})$-pairs for the weather station in Kamloops are emphasized by means of a frame in *dashed lines*

(2006), p. 269, where a plausible meteorological explanation for the poor fit is provided.

4.2 Handwritten digit recognition

We analyze a subset of the USPS database described in Le Cun et al. (1989) and Hastie et al. (2001). The complete database consists of more than 9,000 greyscale images of handwritten digits at a resolution of 16×16 pixels, with a predefined division into learning and test sample. We extract all instances of the digits '3', '5', '6', '8', '9', which—in view of their composition of arcs and circular shapes—we regard to be suitable for an effective visualization (see Fig. 7). We consider all resulting pairwise classification problems, in which each of the 256 pixels is used as input of a linear classification rule of the form $\widehat{y} = \text{sign}(\widehat{\beta}_0 + x_{j,k}\widehat{\beta}_{j,k})$, where pixel j, k is denoted by $x_{j,k}$, and $\widehat{\beta}_{j,k}$ is the corresponding coefficient, $j, k = 1, \ldots, 16$. We here state clearly that linear classifiers are well-known to be of inferior performance for this dataset. Instead, our attention focuses on a comparison of the lasso, elastic net and structured elastic net regularization. The ridge regularizer corresponds to standard linear SVC, and is hence included as baseline. For the structured elastic net, the matrix

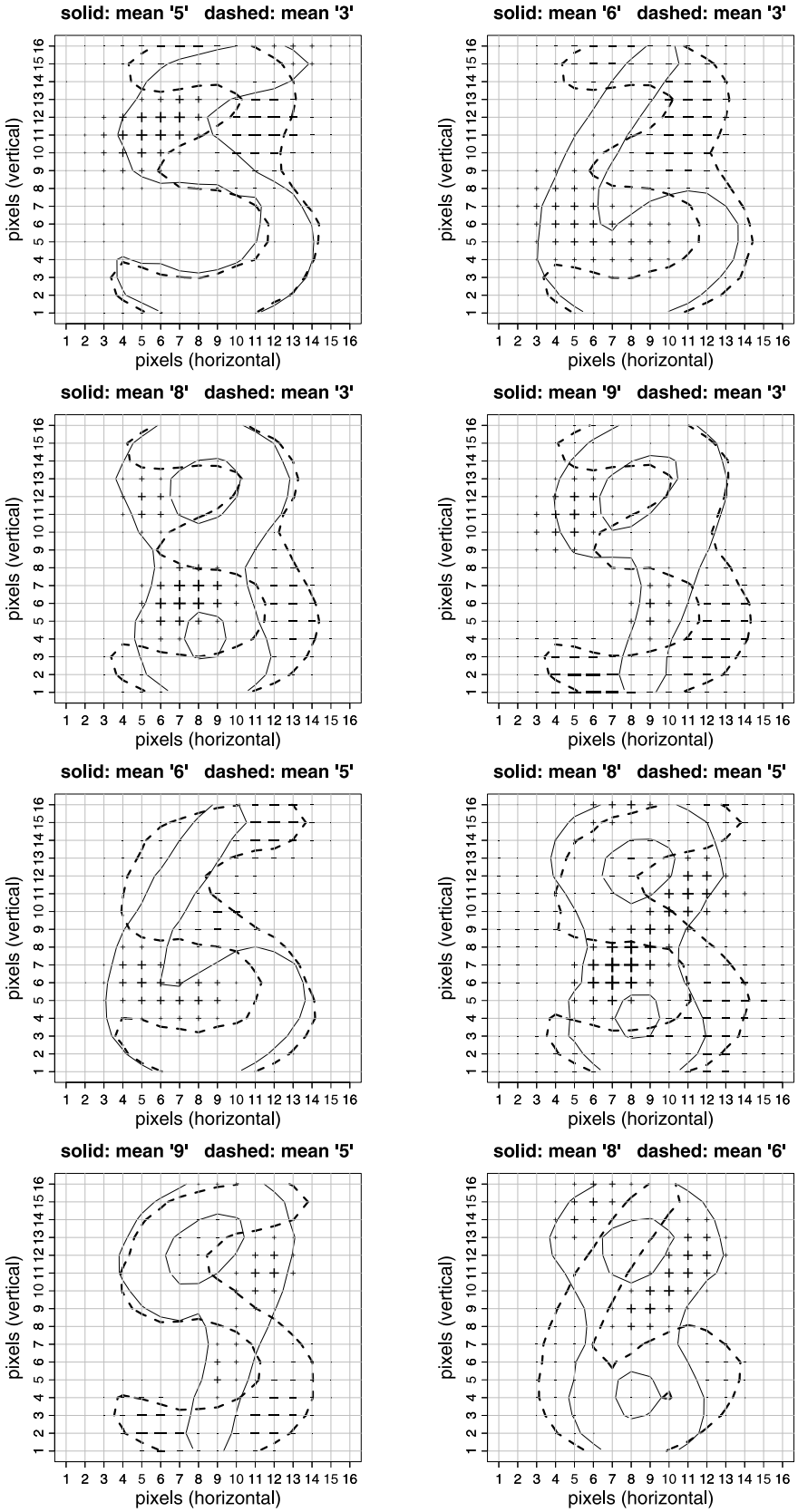

Fig. 7 Schematic representation of the coefficient profiles $\{\widehat{\beta}_{j,k}, 1 \le j,k \le 16\}$ of the ten CV-optimal structured elastic net models. The symbols '+' and '−' indicate the sign of the coefficients (with the convention that the digit which is numerically the larger of the two is labeled +1), and the symbol size their absolute magnitude. For better guidance, the average shapes of the digits within the USPS database have been added

Fig. 7 (*Continued.*)

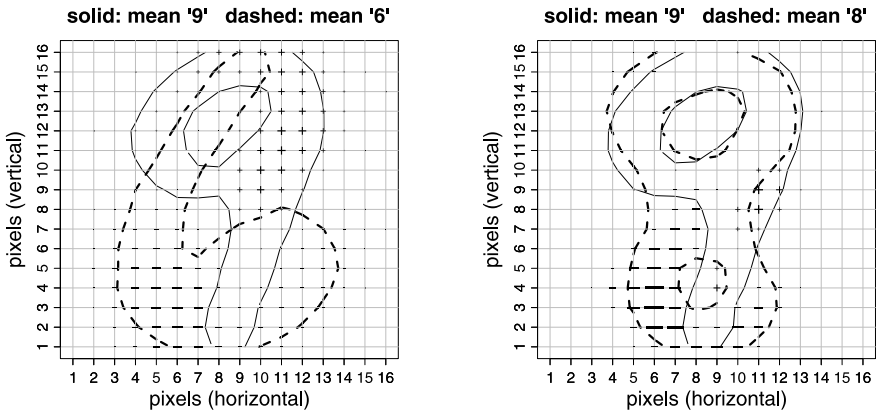

Table 3 Misclassification rates on the test for the ten selected handwritten digit recognition problems. The table also displays the number of nonzero coefficients ('nonzero.coef') of the CV-optimal models

Problem		Method			
		Ridge	Lasso	e.net	Structured e.net
'3' vs. '5'	test error	39/326	48/326	38/326	36/326
	nonzero.coef		29	226	189
'3' vs. '6'	test error	10/336	24/336	6/336	6/336
	nonzero.coef		33	241	208
'3' vs. '8'	test error	26/332	38/332	31/332	30/332
	nonzero.coef		30	125	104
'3' vs. '9'	test error	8/343	11/343	8/343	9/343
	nonzero.coef		26	256	201
'5' vs. '6'	test error	13/330	25/330	17/330	16/330
	nonzero.coef		25	158	116
'5' vs. '8'	test error	21/326	36/326	21/326	21/326
	nonzero.coef		38	256	252
'5' vs. '9'	test error	12/337	22/337	12/337	13/337
	nonzero.coef		30	174	99
'6' vs. '8'	test error	9/336	26/336	12/336	12/336
	nonzero.coef		28	121	123
'6' vs. '9'	test error	5/347	11/347	1/347	3/347
	nonzero.coef		36	256	190
'8' vs. '9'	test error	15/343	39/343	20/343	18/343
	nonzero.coef		26	124	84
Total	test error	158/3356	280/3356	166/3356	164/3356

$\boldsymbol{\Lambda}$ is chosen such that

$$\boldsymbol{\beta}^\top \boldsymbol{\Lambda} \boldsymbol{\beta} = \sum_j \sum_k (\beta_{j,k} - \beta_{j-1,k})^2 + (\beta_{j+1,k} - \beta_{j,k})^2 + (\beta_{j,k} - \beta_{j,k-1})^2 + (\beta_{j,k} - \beta_{j,k+1})^2, \tag{4.1}$$

the usual discretization of the Laplacian acting on functions defined on $\mathbb{R}^2$. Note that the regularizer (4.1) results from the weight specification (2.5), identifying a grid point $\boldsymbol{t}$ with a pair of indices (j, k). Hyperparameters are determined such that the numbers of misclassifications in tenfold cross-validation on the learning sample is minimized. The results are reported in Table 3, showing that mere ℓ_1-regularization performs poorly. The elastic net and the structured elastic net select far more pixels, and the latter is typically a good deal more sparse than the former, while being equally good

in terms of classification performance. However, both are beaten by standard linear SVC. It is instructive to have a look at the coefficient surfaces of the structured elastic net, depicted in Fig. 7. The results are in accordance to what one would pick as predictive regions by visual inspection: large coefficients are predominantly placed in regions where the two digits show little or no overlap.

5 Discussion

In this paper, we have described the structured elastic net regularizer for quantile regression and support vector classification tailored to the grouped selection of variables on the basis of a known association structure. By means of artificial and real world datasets, we have demonstrated in what way the structured elastic net differs from its predecessors, and we have shown a series of scenarios in which it can be useful. A drawback of the structured elastic net is its dependence on two tuning parameters. The loss functions of quantile regression and support vector classification admit a computational shortcut by tracking a sequence of piecewise linear solution paths with one of the two parameters kept fixed. While this is already a considerable computational simplification, there is further room for improvement. Cross-validation could be replaced by the computation of a suitable model selection criterion, as done in previous work (Li and Zhu 2008). For quantile regression, Rosset (2008) proposes a bi-level solution path algorithm for varying regularization parameter and varying quantile τ, thus elegantly avoiding the problem of *quantile crossing* (Koenker 2005). An extension of the path algorithm of Sect. 3 into this direction would be of much practical use.

The approach pursued in this paper focuses on how to compute estimators as minimizers of convex optimization problems, but no advice on how do further inference, notably how to obtain standard errors, is given—a problem that seems to be inherent in frequentist approaches to variable selection based on ℓ_1-regularization. A Bayesian framework could provide an error assessment of all parameters of interest, also covering more complicated cases where the matrix $\mathbf{\Lambda}$ depends on additional unknown parameters, which is out of the scope of the present paper. In this context, it is worth mentioning that Bayesian counterparts of the two loss functions employed in this paper (Sollich 2002; Lancaster and Jun 2009) as well as the ℓ_1-regularizer (Park and Casella 2008; Hans 2009, 2010) have been developed.

Availability An `R` implementation of the solution path algorithms is available from the author's homepage at http://www.ml.uni-saarland.de/people/slawski/mspublications.shtml.

Acknowledgements The author is greatly indebted to Ji Zhu (University of Michigan) for providing him example code for the solution path of ℓ_1-regularized support vector classification, which turned out to be helpful for the development and implementation of the algorithm described in Sect. 3.

The author thanks the two reviewers for their constructive comments, leading to increased clarity in presentation and an improved discussion.

References

Bartlett, P., Jordan, M., McAuliffe, J.: Convexity, classification, and risk bounds. J. Am. Stat. Assoc. **101**, 138–156 (2006)

Bennett, K., Mangasarian, O.: Multicategory separation via linear programming. Optim. Methods Softw. **3**, 27–39 (1993)

Bondell, H., Reich, B.: Simultaneous regression shrinkage, variable selection and clustering of predictors with OSCAR. Biometrics **64**, 115–123 (2008)

Bradley, P., Mangasarian, O.: Feature selection via concave minimization and support vector machines. In: International Conference on Machine Learning (1998)

Christianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines. Cambridge University Press, Cambridge (2000)

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression (with discussion). Ann. Stat. **32**, 407–499 (2004)

El Anbari, M., Mkhadri, A.: Penalized regression combing the L_1 norm and a correlation based penalty. Technical report, Université Paris Sud 11 (2008)

Hans, C.: Bayesian lasso regression. Biometrika **96**, 221–229 (2009)

Hans, C.: Model uncertainty and variable selection in Bayesian lasso regression. Stat. Comput. **20**, 221–229 (2010)

Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, New York (2001)

Hastie, T., Rosset, S., Tibshirani, R., Zhu, J.: The entire regularization path for the support vector machine. J. Mach. Learn. Res. **5**, 1391–1415 (2004)

James, G., Wang, J., Zhu, J.: Functional linear regression that's interpretable. Ann. Stat. **37**, 2083–2108 (2008)

Koenker, R.: Quantile Regression. Cambridge University Press, Cambridge (2005)

Lancaster, T., Jun, S.J.: Bayesian quantile regression methods. J. Appl. Econom. **25**, 287–307 (2009)

Landau, S., Ellison-Wright, I., Bullmore, E.: Tests for a difference in timing of physiological response between two brain regions measured by using functional magnetic resonance imaging. Appl. Stat. **63–82**, 53 (2003)

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

Li, C., Li, H.: Variable selection and regression analysis for graph-structured covariates with an application to genomics. Ann. Appl. Stat. **4**(3), 1498–1516 (2010)

Li, Y., Zhu, J.: L_1-norm Quantile regression. J. Comput. Graph. Stat. **17**, 163–185 (2008)

Li, Y., Liu, Y., Zhu, J.: Quantile regression in reproducing kernel Hilbert spaces. J. Am. Stat. Assoc. **102**, 255–268 (2007)

Lin, Y.: Support vector machines and the Bayes rule in classification. Data Min. Knowl. Discov. **6**, 259–275 (2002)

Park, T., Casella, G.: The Bayesian lasso. J. Am. Stat. Assoc. **103**, 681–686 (2008)

Ramsay, J., Silverman, B.: Functional Data Analysis. Springer, New York (2006)

Rosset, S.: Bi-level path following for cross validated solution of kernel quantile regression. In: International Conference on Machine Learning (2008)

Rosset, S., Zhu, J.: Piecewise linear regularized solution paths. Ann. Stat. **35**, 1012–1030 (2007)
Schölkopf, B., Smola, A.: Learning with Kernels. MIT Press, Cambridge (2002)
Slawski, M., zu Castell, W., Tutz, G.: Feature selection guided by structural information. Ann. Appl. Stat. **4**(2), 1056–1080 (2010)
Sollich, P.: Bayesian methods for support vector machines: evidence and predictive class probabilities. Mach. Learn. **46**, 21–52 (2002)
Stein, M.: Interpolation of Spatial Data. Springer, New York (1999)
Steinwart, I., Christmann, A.: Support Vector Machines. Springer, Berlin (2008)
Takeuchi, I., Le, Q., Sears, T., Smola, A.: Nonparametric quantile regression. J. Mach. Learn. Res. **7**, 1231–1264 (2006)
Tibshirani, R.: Regression shrinkage and variable selection via the lasso. J. R. Stat. Soc. Ser. B **58**, 671–686 (1996)
Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K.: Sparsity and smoothness via the fused lasso. J. R. Stat. Soc. Ser. B **67**, 91–108 (2005)
Tutz, G., Gertheiss, J.: Feature extraction in signal regression: a boosting technique for functional data regression. J. Comput. Graph. Stat. **19**, 154–174 (2010)
Tutz, G., Ulbricht, J.: Penalized regression with correlation based penalty. Stat. Comput. **19**, 239–253 (2009)
Wang, L., Shen, X.: Multi-category support vector machines, feature selection, and solution path. Stat. Sin. **16**, 617–634 (2005)
Wang, L., Zhu, J., Zou, H.: The doubly regularized support vector machine. Stat. Sin. **16**, 589–616 (2006)
Wood, S.: R package gamair: Data for "GAMs: An Introduction with R", Version 0.0-4. Available from www.r-project.org (2006)
Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B **68**, 49–67 (2006)
Zhao, P., Rocha, G., Yu, B.: The composite absolute penalties family for grouped and hierarchical variable selection. Ann. Stat. **37**, 3468–3497 (2009)
Zhu, J., Rosset, S., Hastie, T., Tibshirani, R.: L_1 norm support vector machine. Adv. Neural Inf. Process. Syst. **16**, 55–63 (2003)
Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B **67**, 301–320 (2005)