Penalized likelihood regression for generalized linear models with non-quadratic penalties

Anestis Antoniadis · Irène Gijbels · Mila Nikolova

Abstract   One of the popular method for fitting a regression function is regularization: minimizing an objective function which enforces a roughness penalty in addition to coherence with the data. This is the case when formulating penalized likelihood regression for exponential families. Most of the smoothing methods employ quadratic penalties, leading to linear estimates, and are in general incapable of recovering discontinuities or other important attributes in the regression function. In contrast, nonlinear estimates are generally more accurate. In this paper, we focus on non-parametric penalized likelihood regression methods using splines and a variety of *non-quadratic* penalties, pointing out common basic principles. We present an asymptotic analysis of convergence rates that justifies the approach. We report on a simulation study including comparisons between our method and some existing ones. We illustrate our approach with an application to Poisson non-parametric regression modeling of frequency counts of reported acquired immune deficiency syndrome (AIDS) cases in the UK.

A. Antoniadis
Laboratoire Jean Kuntzmann, Department de Statistique, Université Joseph Fourier,
Tour IRMA, B.P. 53, 38041 Grenoble Cedex 9, France

I. Gijbels (✉)
Department of Mathematics, Leuven Statistics Research Centre (LStat),
Katholieke Universiteit Leuven, Celestijnenlaan 200B, Box 2400, 3001 Leuven, Belgium
e-mail: irene.gijbels@wis.kuleuven.be

M. Nikolova
Centre de Mathématiques et de Leurs Applications, CNRS-ENS de Cachan,
PRES UniverSud, 61 av. du Président Wilson, 94235 Cachan Cedex, France

1 Introduction

In many statistical applications, non-parametric modeling can provide insight into the features of a dataset that are not obtainable by other means. One successful approach involves the use of (univariate or multivariate) spline spaces. As a class, these methods have inherited much from classical tools for parametric modeling. Smoothing splines, in particular, are appealing because they provide computationally efficient estimation and often do a good job of smoothing noisy data. Two shortcomings of smoothing splines, however, are the need to choose a global smoothness parameter and the inability of linear procedures (conditional on the choice of smoothness) to adapt to spatially heterogeneous functions. This has led to investigations of curve-fitting using free-knot splines, that is, splines in which the number of knots and their locations are determined from the data (Eilers and Marx 1996; Denison et al. 1998; Lindstrom 1999; Zhou and Shen 2001 and DiMatteo et al. 2001, among others). Such procedures are strongly connected to variable and model selection methods and may be seen as a particular case of non-quadratic regularization, which is the point of view adopted in this paper.

Our approach to smoothing and recovering eventual discontinuities or other important attributes in a regression function is based on methods for non-regular penalization in the context of generalized regression, which includes logistic regression, probit regression, and Poisson regression as special cases. For global smoothing, penalized likelihood regression with responses from exponential family distributions traces back to O'Sullivan et al. (1986); see also Green and Yandell (1985) and Eilers and Marx (1996). The asymptotic convergence rates of the penalized likelihood regression estimates have been studied by Cox and O'Sullivan (1990) and Gu and Qiu (1994). Other related more recent ideas for smoothing of non-normal data are discussed by Biller (2000), Klinger (2001), DiMatteo et al. (2001), where smoothing is seen as a variable selection problem. However, there has not been much previous work about the design of appropriate penalty functions that ensure the preservation of eventual discontinuities. Exceptions are the papers by Ruppert and Carroll (2000), Antoniadis and Fan (2001) and Fan and Li (2001), although the issue of inference on eventual change points is not given much emphasis. In the present paper, we are concerned with noise reduction or smoothing of functions where there is evidence for smooth regions from the data and the problem is not to smooth where there is evidence for breaks or boundaries.

We represent the regression function as a linear combination of a large number of basis functions which also have the capability of catching some sharp changes in the regression relationship. Given the large number of basis functions that might be used, most of the inferential procedures from generalized linear models cannot be used directly and penalization procedures that are strongly connected to variable and model selection methods and which may be seen as a particular case of non-quadratic

regularization are specifically designed. There exists some recent works that are closely related to the topics discussed in this paper. After reviewing the literature, putting as such all in a unifying framework, we develop some theoretical results concerning the bias and variance of our estimators and their asymptotic properties and highlight some new approaches for the determination of regularization parameters involved in our estimation procedures.

The structure of the paper is as follows. In the next section we formulate the class of non-parametric regression problems that we study within the framework of non-parametric regression for exponential families. Using a spline basis, approximating the regression by its projection onto a finite dimensional spline space and introducing an appropriate penalty to the log-likelihood, we are able to perform a discontinuity-preserving regularization. In Sect. 3, we provide some further details about choices of bases and penalties, presenting all in a unifying framework. Details of the corresponding optimization problems are provided in Sect. 4. The asymptotic analysis is conducted in Sect. 5, where the sampling properties of the penalized likelihood estimators are established. Section 6 discusses data-driven choices for the regularization parameters. Simulation results and numerical comparisons for several test functions are provided in Sect. 7. As an illustration we analyze some of the Poisson data.

2 Model formulation and basic notations

2.1 Generalized models

We first briefly describe the class of generalized linear regression models. Consider a pair $(\mathbf{X}, Y)$ of random variables, where Y is real-valued and $\mathbf{X}$ is real vector-valued; here Y is referred to as a response or dependent variable and $\mathbf{X}$ as the vector of covariates or predictor variables. We only consider here the univariate case, but note that the extension of our methods to multiple predictors can be easily made. A basic generalized model (GM) analysis starts with a random sample of size n from the distribution of (X, Y) where the conditional distribution of Y given $X = x$ is assumed to be from a one-parameter exponential family distribution with a density of the form

$$\exp\left(\frac{y\theta(x) - b(\theta(x))}{\phi} + c(y, \phi)\right),$$

where $b(\cdot)$ and $c(\cdot)$ are known functions, while the natural parameter function $\theta(x)$ is unknown and specifies how the response depends on the covariate. The parameter ϕ is a scale parameter which is assumed to be known. The conditional mean and variance of the ith response Y_i are given by

$$\mathbb{E}(Y_i|X = x_i) = \dot{b}(\theta(x_i)) = \mu(x_i) \quad \text{Var}(Y_i|X = x_i) = \phi\ddot{b}(\theta(x_i)).$$

Here a dot denotes differentiation. In the usual GM framework, the mean is related to the GM regression surface via the link function transformation $g(\mu(x_i)) = \eta(x_i)$ where $\eta(x)$ is called the predictor function. A wide variety of distributions can be

modeled using this approach including normal regression with additive normal errors (identity link), logistic regression (logit link) where the response is a binomial variable and Poisson regression models (log link) where the observations are from a Poisson distribution. Many examples are given in McCullagh and Nelder (1989). See also Fahrmeir and Tutz (1994).

A regression analysis seeks to estimate the dependence $\eta(x)$ based on the observed data (x_i, y_i), $i = 1, \ldots, n$. A standard parametric model restricts $\eta(x)$ to a low-dimensional function space through a certain function form $\eta(x, \boldsymbol{\beta})$, where the function is known up to a finite-dimensional parameter $\boldsymbol{\beta}$. A (generalized) linear model results when $\eta(x, \boldsymbol{\beta})$ is linear in $\boldsymbol{\beta}$. When knowledge is insufficient to justify a parametric model, various non-parametric techniques can be used for the estimation of $\eta(x)$. Like O'Sullivan et al. (1986) and Gu and Kim (2002), we consider the penalized log-likelihood estimation of $\eta(x)$ through the maximization of

$$Z_n(\eta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \eta(x_i)) - \lambda J(\eta), \tag{1}$$

where ℓ is the log-likelihood of the model and $J(\cdot)$ is a roughness functional. To estimate the regression function using the penalized maximum likelihood method, one maximizes the functional (1), for a given λ, in some function space in which $J(\eta)$ is defined and finite. For GM models the first term of (1) is concave in η.

The function $\eta(\cdot)$ can be estimated in a flexible manner by representing it as a linear combination of known basis functions $\{h_k, \ k = 1, \ldots, p\}$,

$$\eta(x) = \sum_{k=1}^{p} \beta_k h_k(x), \tag{2}$$

and then to estimate the coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, where $\mathbf{A}^T$ denotes the transposed of a vector or matrix $\mathbf{A}$. Usually the number p of basis functions used in the representation of η should be large in order to give a fairly flexible way for approximating η (this is similar to the high-dimensional setup of Klinger (2001)). Popular examples of such basis functions are wavelets and polynomial splines. A crucial problem with such representations is the choice of the number p of basis functions. A small p may result in a function space which is not flexible enough to capture the variability of the data, while a large number of basis functions may lead to serious overfitting. Traditional ways of "smoothing" are through basis selection see e.g. Friedman and Silverman (1989); Friedman (1991) and Stone et al. (1997) or regularization. The key idea in model selection methods is that overfitting is avoided by careful choice of a model that is both parsimonious and suitable for the data. Regularization takes a different approach: rather than seeking a parsimonious model, one uses a highly parametrized model and imposes a penalty on large fluctuations on the fitted curve.

Given observations $x_i, i = 1, \ldots, n$, let $\mathbf{h}(x_i) = \left(h_1(x_i), h_2(x_i), \ldots, h_p(x_i)\right)^T$ the column-vector containing the evaluations of the basis functions in the point x_i. Since $\eta(x_i) = \mathbf{h}^T(x_i) \boldsymbol{\beta}$, (1) becomes

$$Z_n(\boldsymbol{\beta}) = \frac{1}{n} L_{\mathbf{y}}(\boldsymbol{\beta}) - \lambda J(\boldsymbol{\beta}), \tag{3}$$

where we denoted, for commodity, $Z_n(\boldsymbol{\beta})$ for $Z_n(\eta_{\boldsymbol{\beta}})$ and $J(\boldsymbol{\beta})$ for $J(\eta_{\boldsymbol{\beta}})$, and with

$$L_{\mathbf{y}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell \left(y_i, \mathbf{h}^T(x_i)\boldsymbol{\beta} \right). \tag{4}$$

Minimizing $-Z_n$ with respect to $\boldsymbol{\beta}$, leads to estimation of the parameters $\boldsymbol{\beta}$ in

$$\eta(x_i) = g(\mu(x_i)) = \mathbf{h}^T(x_i)\boldsymbol{\beta}.$$

In this paper we concentrate on polynomial spline methods, which are easy to interpret and useful in many applications. We allow p to be large, but use an adequate penalty J on the coefficients to control the risk of overfitting the data. Hence, our approach is similar in spirit to the penalized regression spline approach for additive white noise models by Eilers and Marx (1996) and Ruppert and Carroll (2000), among others.

2.2 Truncated power basis and B-splines

Polynomial regression splines are continuous piecewise polynomial functions where the definition of the function changes at a collection of knot points, which we write as $t_1 < \cdots < t_K$. Using the notation $z_+ = \max(0, z)$, then, for an integer $d \geq 1$, the truncated power basis for polynomial of degree d regression splines with knots $t_1 < \cdots < t_K$ is

$$\left\{ 1, x, \ldots, x^d, (x - t_1)_+^d, \ldots, (x - t_K)_+^d \right\}.$$

When representing a univariate function f as a linear combination of these basis functions as

$$f(x) = \sum_{j=0}^{d} \beta_j x^j + \sum_{l=1}^{K} \beta_{d+l}(x - t_l)_+^d,$$

it follows that each coefficient β_{d+l} is identified as a jump in the dth derivative of f at the corresponding knot. Therefore, coefficients in the truncated power basis are easy to interpret especially when tracking change-points or more or less abrupt changes in the regression curve.

Use of polynomial regression splines, while easy to understand, is sometimes not desirable because it is computationally less stable (see Dierckx (1993)). Another possible choice, offering analytical and computational advantages, is a B-splines basis. Eilers and Marx (1996), in their paper on penalized splines global non-parametric smoothing, choose the numerically stable B-spline basis, suggesting a moderately large number of knots (usually between 20 and 40) to ensure enough flexibility, and

using a quadratic penalty based on differences of adjacent B-Spline coefficients to guarantee sufficient smoothness of the fitted curves. Details of B-splines and their properties can be found in de Boor (1978). A normalized B-splines basis of order q with interior knots $0 < t_1 < \cdots < t_K < 1$ is a set of degree $q - 1$ spline functions $\{B^q_{Kj}, j = 1, \ldots, q + K\}$, that is a basis of the space of the qth-order polynomial splines on [0, 1] with K ordered interior knots. The functions B^q_{Kj} are positive and are non-zero only on an interval which covers no more than $q + 1$ knots. Equivalently, at any point x there are no more than q B-splines that are non-zero. A recursive relationship can be used to describe B-splines, leading to a very stable numerical computation algorithm.

2.3 Penalized likelihood

By (2), the estimation of η is simplified to the estimation of $\boldsymbol{\beta}$. The use of a criterion function with penalty, as in (1), has a long history which goes back to Whittaker (1923) and Tikhonov (1963). When the penalty is of the form $J(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2$, we talk about quadratic regularization. When the criterion function is based on the log-likelihood of the data, the method of regularization is called penalized maximum likelihood method (PMLE).

The literature on penalized maximum likelihood estimation is abundant, with several recommending strategies for the choice of J and λ. A synthesis has been presented in Titterington (1985), Poggio (1985) and Demoment (1989). In statistics, using a penalty function can also be interpreted as formulating some prior knowledge about the parameters of interest (Good and Gaskins 1971) and leads to the so-called Bayesian MAP estimation. Under such a Bayesian setting, maximizing $Z_n(\boldsymbol{\beta})$ is equivalent to maximize the posterior distribution $p(\boldsymbol{\beta}|\mathbf{y})$ corresponding to the prior $p(\boldsymbol{\beta}) \propto \exp(-\lambda J(\boldsymbol{\beta}))$. Typically, J is of the form

$$J(\boldsymbol{\beta}) = \sum_{k=1}^{r} \gamma_k \psi\left(d_k^T \boldsymbol{\beta}\right), \tag{5}$$

where $\gamma_k > 0$ are weights and d_k are linear operators. Thus the penalty J pushes the solution $\hat{\boldsymbol{\beta}}$ to be such that $|d_k^T \hat{\boldsymbol{\beta}}|$ is small. In particular, if d_k are finite difference operators, neighboring coefficients of $\hat{\boldsymbol{\beta}}$ are encouraged to have similar values in which case $\hat{\boldsymbol{\beta}}$ involves homogeneous zones. If $d_k = e_k$ are the vectors of the canonical basis of $\mathbb{R}^p$, then J encourages the components $\hat{\boldsymbol{\beta}}_k$ to have small magnitude. The setting of Bayesian MAP estimation gave rise to a large variety of functions ψ, especially within the framework of Markov random fields where ψ is interpreted as a potential function (Geman and McClure 1987; Besag 1989).

The choice of $J(\boldsymbol{\beta})$ depends strongly on the basis that is used for representing the predictor function $\eta(x)$. If one uses for example a truncated power basis functions of degree d, the coefficients of the basis functions at the knots involve the jumps of the dth derivative, and therefore J is generally of the form $J(\boldsymbol{\beta}) = \sum_k \gamma_k \psi(\beta_k)$ where $\gamma_k > 0$. Indeed, there is no reason that neighboring coefficients of $\boldsymbol{\beta}$ have close values.
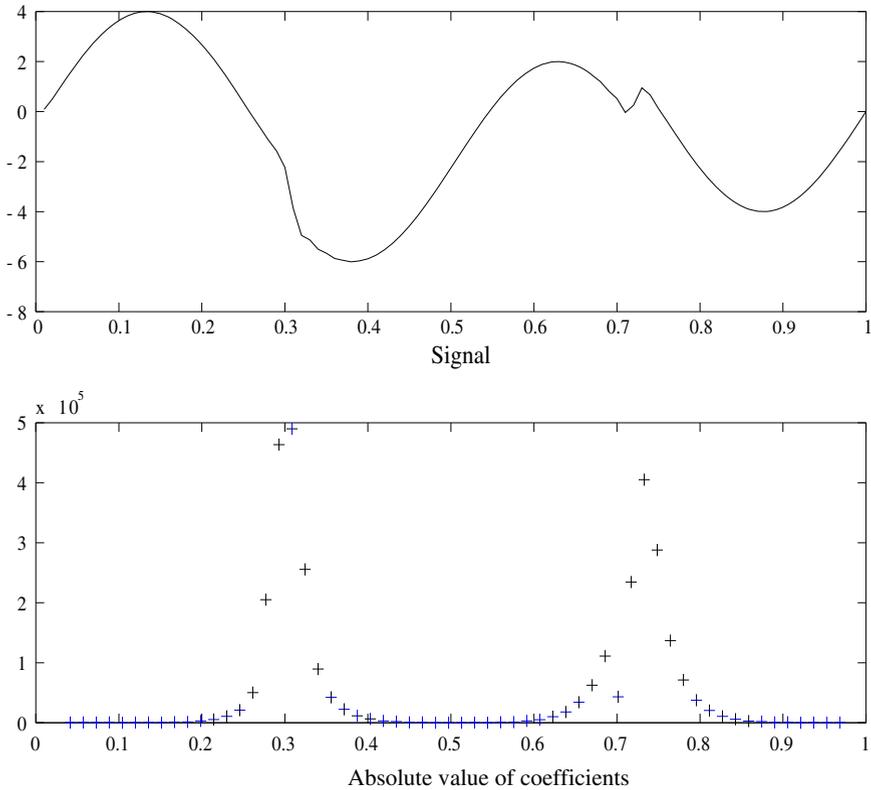
Fig. 1 Behavior of the coefficients of a function in a truncated power basis

This is illustrated in Fig. 1, where large coefficients are associated with singularities in the function that is decomposed in the truncated power basis. Penalties of this kind, with $\psi(\cdot) = |\cdot|$ have been suggested and studied in detail by several authors, including Donoho et al. (1992), Alliney and Ruzinsky (1994), Mammen and Van de Geer (1997), Ruppert and Carroll (1997) and Antoniadis and Fan (2001). Such penalties have the feature that many of the components of $\boldsymbol{\beta}$ are shrunk all the way to 0. In effect, these coefficients are deleted. Therefore, such procedures perform a model selection.

When using B-splines, however, penalties on neighbor B-spline coefficients ensure that neighboring coefficients do not differ too much from each other when η is smooth. As illustrated in Fig. 2, it is the absolute values of first-order or second-order differences that are maximum at singularity points of the decomposed curve and penalties such as the one given in (5) are therefore more adequate.

To end this section, we will discuss a somewhat general type of penalties that we are going to use within the generalized model approach. Several penalty functions have been used in the literature. The L_2 or quadratic penalty $\psi(\beta) = |\beta|^2$ yields a ridge type regression, while the L_1 penalty $\psi(\beta) = |\beta|$ results in LASSO (first proposed by Donoho and Johnstone 1994) in the wavelet setting and extended by Tibshirani 1996 for general least squares settings). The latter is also the penalty used by Klinger (2001)
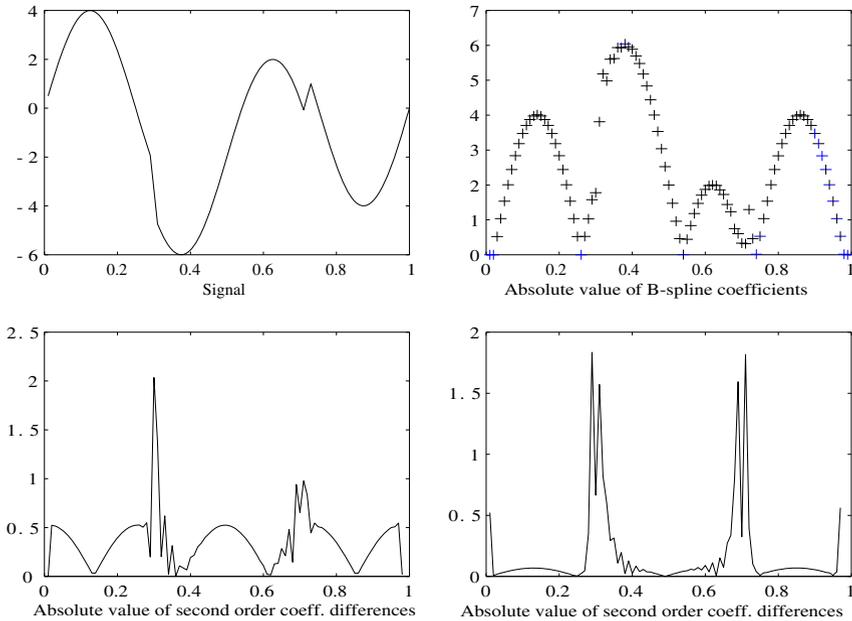
Fig. 2 Behavior of the coefficients of a function in a B-splines basis

for P-spline fitting in generalized linear models. More generally, the L_q $(0 \leq q \leq 1)$ leads to bridge regression (see Frank and Friedman 1993; Ruppert and Carroll 1997; Fu 1998; Knight and Fu 2000).

A well-known justification of regularization by LASSO type penalties is that it usually leads to sparse solutions, i.e. a small number of non-zero coefficients in the basis function expansion, and thus performs model selection. This is generally true for penalties alike the smoothed clipped absolute deviation (SCAD) penalty (item 16 in Table 2) introduced by Fan (1997) and studied in detail by Antoniadis and Fan (2001) and Fan and Li (2001). SCAD penalties present nice oracle properties. For LASSO-like procedures, recent works by Zhao and Yu (2006), Zou (2006) and Yuan and Lin (2007) in the multiple linear regression models have looked precisely at the model consistency of the LASSO, i.e. if we know that the data were generated from a sparse loading vector, does the LASSO actually recover it when the number of observed data points grows? In the case of a fixed number of covariates, the LASSO does recover the sparsity pattern if and only if a certain simple condition on the generating covariance matrices is verified; see Yuan and Lin (2007). In particular, in low correlation settings, the LASSO is indeed consistent. However, in presence of strong correlations, the LASSO cannot be consistent, shedding light on the potential problems of such procedures for variable selection. Adaptive versions where data-dependent weights are added to the L_1-norm then allow to keep the consistency in all situations (see Zou 2006) and our penalization procedures using the weights γ_k (see 5) are in this spirit.

Several conditions on ψ are needed in order for the penalized likelihood approach to be effective. Usually, the penalty function ψ is chosen to be symmetric and increasing

Table 1 Examples of convex penalty functions

Convex	
Smooth at zero	Singular at zero
1. $\psi(\beta) = \|\beta\|^{\alpha}, \alpha > 1$	6. $\psi(\beta) = \|\beta\| \quad \psi'(0^+) = 1$
2. $\psi(\beta) = \sqrt{\alpha + \beta^2}$	7. $\psi(\beta) = \alpha^2 - (\|\beta\| - \alpha)^2 I\{\|\beta\| < \alpha\} \quad \psi'(0^+) = 2\alpha$
3. $\psi(\beta) = \log(\cosh(\alpha\beta))$	
4. $\psi(\beta) = \beta^2 - (\|\beta\| - \alpha)^2 I\{\|\beta\| > \alpha\}$	
5. $\psi(\beta) = 1 + \|\beta\|/\alpha - \log(1 + \|\beta\|/\alpha)$	

Table 2 Examples of non-convex penalty functions

Non-convex	
Smooth at zero	Singular at zero
8. $\psi(\beta) = \alpha\beta^2/(1 + \alpha\beta^2)$	12. $\psi(\beta) = \|\beta\|^{\alpha}, \alpha \in (0, 1) \quad \psi'(0^+) = \infty$
9. $\psi(\beta) = \min\{\alpha\beta^2, 1\}$	13. $\psi(\beta) = \alpha\|\beta\|/(1 + \alpha\|\beta\|) \quad \psi'(0^+) = \alpha$
10. $\psi(\beta) = 1 - \exp(-\alpha\beta^2)$	14. $\psi(0) = 0, \; \psi(\beta) = 1, \forall \beta \neq 0 \quad$ discont.
11. $\psi(\beta) = -\log\left(\exp(-\alpha\beta^2) + 1\right)$	15. $\psi(\beta) = \log(\alpha\|\beta\| + 1) \quad \psi'(0^+) = \alpha$
	16. $\int_0^{\beta} \psi'(u)du \quad \psi'(\|\beta\|) = \alpha\{I\{\|\beta\| \leq \alpha\} + \frac{(a\alpha - \|\beta\|)_+}{(a-1)\alpha}\{\|\beta\| > \alpha\}\},$ $a > 2$

on $[0, +\infty)$. Throughout this paper, we suppose that ψ satisfies these two conditions. Furthermore, ψ can be convex or non-convex, smooth or non-smooth. In the wavelet setting, Antoniadis and Fan (2001) provide some insights into how to choose a penalty function. A good penalty function should result in an estimator that avoids excessive bias (*unbiasedness*), that forces sparse solutions to reduce model complexity (*sparsity*) and, that is continuous in the data to avoid unnecessary variation (*stability*). Moreover, from the computational viewpoint, penalty functions should be chosen in a way that the resulting optimization problem is easily solvable.

As a first contribution of this paper, we will try to summarize and unify the main features of ψ that determine essential properties of the maximizer $\hat{\boldsymbol{\beta}}$ of Z_n. The accent will be put on adequate choices of penalty functions among the many proposed in the literature. Essentially, penalties can be convex or non-convex with or without a singularity at the origin (non-differentiable at 0). Tables 1 and 2 list several examples of such penalties.

In the following section, we further discuss some necessary conditions on penalty functions for obtaining unbiasedness, sparsity and stability that have been derived by Nikolova (2000), Antoniadis and Fan (2001) and Fan and Li (2001). Stability for non-convex and possibly non-smooth regularization has been studied by Durand and Nikolova (2006a), Durand and Nikolova (2006b). Concerning edge-detection and unbiasedness using non-convex regularization (smooth or non-smooth) see for example Nikolova (2005).

3 Penalties and regularization

The truncated power basis for polynomial of degree d regression splines with knots $t_1 < \cdots < t_K$ or the set of order q B-splines with K interior knots may be viewed as a given family of piecewise polynomial functions $\{B_{Kj}^q, j = 1, \ldots, q + K\}$. Assuming the initial location of the knots known, the $K + q$-dimensional parameter vector, $\boldsymbol{\beta}$, describes the $K + q$ necessary polynomial coefficients that parsimoniously represent the function η. A critical component of spline smoothing is the choice of knots, especially for curves with varying shapes and frequencies in their domain. We consider a two-stage knot selection scheme for adaptively fitting regression splines or B-splines to our data. As it is usually done in this context, an initial fixed large number of potential knots will be chosen at fixed quantiles of the independent variable with the intention to have sufficient points at regions where the curve shows rapid changes. Basis selection by non-smooth at zero penalties will then eliminate basis functions when they are non-necessary, retaining mainly basis functions whose support covers regions with sharp features.

The estimation part of our statistical analysis involves first fixing K and estimating $\boldsymbol{\beta}$ by penalized maximum likelihood within the generalized model setup. We first consider the case where the penalty function ψ is convex and smooth at the origin, a case that includes more or less traditional regularization methods and we then proceed to penalized likelihood situations with more complicated and challenging penalty functions that are more efficient in recovering functions that may have singularities of various orders.

3.1 Smooth regularization

Regularization with smooth penalties leads to classical smooth estimates. The optimization problems, however, are quite different when using convex or non-convex penalties, and one needs to distinguish between these two cases.

3.1.1 Convex penalties

The traditional way of smoothing is using maximum likelihood with a roughness penalty placed on the coefficients that involves a penalty proportional to the square of a specified derivative, usually the second. One usually refers to this as quadratic penalization.The original idea traces back to O'Sullivan et al. (1986) and was further developed by Eilers and Marx (1996) using the B-splines basis.

Typically the penalized maximum likelihood estimator of $\boldsymbol{\beta}$ is defined using the penalty

$$J(\boldsymbol{\beta}) = \boldsymbol{\beta}^T D(\boldsymbol{\gamma}) \boldsymbol{\beta}, \tag{6}$$

where D is an appropriate positive definite matrix and λ is a global penalty parameter. For example, in the case of regression P-splines, $D(\boldsymbol{\gamma})$ is a diagonal matrix with its diagonal elements equal to γ_k and the rest equal to 0, which yields spatially-adapted

penalties. In such a case, $J(\boldsymbol{\beta}) = \sum_k \gamma_k \beta_k^2$, i.e. $\psi(\beta) = \beta^2$ and $d_k = e_k$ in (5). Alternatively, $D(\boldsymbol{\gamma})$ can be a banded matrix corresponding to a quadratic form of finite differences in the components of $\boldsymbol{\beta}$, as in Eilers and Marx (1996) (these authors consider a constant vector $\boldsymbol{\gamma}$, that is their penalty weights on the coefficients are constant). The specification of an unknown penalty weight γ_k for every β_k or $d_k^T \boldsymbol{\beta}$ coefficient results in models far too complex to be of any use in an estimation procedure and usually there is a variety of simplifications of this problem in the literature. We postpone the discussion to later sections.

As it is usually done in generalized linear models, the optimal estimator of $\boldsymbol{\beta}$ in (6) is obtained recursively, for fixed λ and $\boldsymbol{\gamma}$ by an iterated re-weighted least squares algorithm that is usually convergent, easy to implement and produces very satisfactory estimates for smooth functions η. The crucial values of the smoothing parameters (λ and $\boldsymbol{\gamma}$) are usually chosen by generalized cross-validation procedures (GCV). See Sect. 6 for a discussion on selection procedures. When dealing with additive Gaussian errors, the negative log-likelihood is quadratic. Then, the estimate $\hat{\boldsymbol{\beta}}$, as long as λ and $\boldsymbol{\gamma}$ are fixed, is explicitly given as an affine function of the data y. When the function η is a spline function with a fixed number of knots, Yu and Ruppert (2002) present several asymptotic results on the strong consistency and asymptotic normality for the corresponding penalized least-squares estimators. Since similar results are going to be derived for our GLM setup under more general situations in further sections, we refer to the paper of Yu and Ruppert (2002) for details on penalized least squares with quadratic penalties.

While the more or less classical maximum penalized likelihood estimation with a quadratic regularization functional makes computations much easier and allows the use of classical asymptotic maximum likelihood theory in the derivation of asymptotic properties of the estimators, it yields smooth solutions, which may not be acceptable in many applications when the function to recover is less regular.

A remedy can be found if the function ψ in (6) imposes a strong smoothing on small coefficients $d_k^T \beta_k$ and only a loose smoothing on large coefficients. This can be partially achieved by using non-quadratic convex penalty functions ψ such as penalties 1–5 in Table 1.

Among the main characteristics of these functions (for function 1 in Table 1 take $\alpha < 2$) are that $\psi'(t)/t$ is decreasing on $(0, \infty)$ with $\lim_{t \to \infty} \psi'(t)/t = 0$, and that $\lim_{t \searrow 0} \psi'(t)/t > 0$ (here the symbol $\searrow$ is to say that t converges to zero by positive values). In other words, ψ has a strict minimum at zero and ψ' is almost constant (but > 0) except on a neighborhood of the origin. Under the additional condition that either $L_{\mathbf{y}}$ is strictly concave and ψ is convex, or that $L_{\mathbf{y}}$ is concave and ψ is strictly convex, the penalized log-likelihood Z_n is guaranteed to have a unique maximizer. Let us mention that the hyperbolic potential $\psi(t) = \sqrt{\alpha + t^2}$ is very frequently used, often as a smooth approximation to $|t|$ since $\psi(t) \to |t|$ as $\alpha \searrow 0$.

3.1.2 Non-convex penalties

Roughly speaking, smoothing of a coefficient $d_k^T \boldsymbol{\beta}$ is determined by the value of $\psi'(d_k^T \boldsymbol{\beta})$. So a good penalty function should result to an estimator that is unbiased

when the true parameter is large in order to avoid unnecessary bias. It is easy to see that when $\psi'(|t|) = 0$ for large $|t|$, the resulting estimator $\hat{\beta}$ is unbiased for large values of $|t|$. We will see when analyzing the asymptotic properties of the resulting estimators that such a condition is necessary and sufficient. This condition implies that the penalty function $\psi(|t|)$ must be (nearly) constant for large $|t|$ which obviously requires that ψ is non-convex. Such a condition may be traced back to the paper by Geman and Geman (1984) where non-convex regularization has been proposed in the context of Markov random modeling for Bayesian estimation. Popular non-convex smooth penalty functions are items 8–11 in Table 2. See also Black and Rangarajan (1996) and Nikolova (2005).

Note, however, that the main difficulty with such penalties is that the penalized log-likelihood Z is non-concave and may exhibit a large number of local maxima. In general, there is no way to guarantee the finding of a global maximizer and the computational cost is generally high.

3.2 Non-smooth regularization

When one wants to estimate less regular functions then it is necessary to use in the regularization process penalties that are singular at zero. As mentioned before, such penalties enforce sparsity of the spline coefficients in the representation of the regression function. A popular penalty is the L_1 LASSO penalty

$$\psi(\beta) = |\beta|, \tag{7}$$

which is non-smooth at zero but convex. In a wavelet denoising context and under least-squares loss it is known that the optimal solution tends to be sparse and produces asymptotically optimal minimax estimators. This explains why the hyperbolic potential, which is a smooth version of the Lasso penalty is often used. Another popular non-smooth but non-convex penalty sharing similar optimality properties but leading to less biased estimators is the SCAD penalty (item 16 in Table 2).

In a image processing context, a commonly used non-smooth at the origin and non-convex penalty function which is $\psi(\beta) = (\alpha|\beta|)/(1 + \alpha|\beta|)$ (see entry 13 in Table 2). Other non-smooth at the origin potential functions are given in Tables 1 and 2. Although being quite different in shape, these potential functions lead to optimal solutions that are characterized by the fact that $d_k^T \hat{\beta} = 0$ for a certain number (or many) of indexes k. Thus, if $d_k^T \beta$ are first-order differences, minimizers involve constant regions which is known as the stair-casing effect. For arbitrary d_k^T see Nikolova (2000, 2004, 2005). Generally, non-smoothness at the origin encourages sparsity in the resulting estimators.

However, finding a maximum penalized likelihood estimator with such non-convex penalties might be a difficult or even impossible task. In the case of convex non-smooth at the origin penalties, the existence and properties of maximum penalized likelihood estimators is a feasible task as discussed in Sect. 4.2.

4 Optimization of the penalized likelihood

It is sometimes challenging to find the maximum penalized likelihood estimator. In this section, we focus on the existence of maximizers of $Z_n(\boldsymbol{\beta}) = Z_n(\boldsymbol{\beta}; \mathbf{y})$, and when possible on their uniqueness in several restricted but important cases.

Before proceeding further, let us recall here some useful definitions from optimization theory (see e.g. Ciarlet 1989 and Rockafellar and Wets 1997).

For $U \subseteq \mathbb{R}^n$, we say that $\mathcal{B} : U \to \mathbb{R}^p$ is a *(local) maximizer function* for the family $Z_n(\beta; U) = \{Z_n(\boldsymbol{\beta}; \mathbf{y}) : \mathbf{y} \in U\}$ if for every $\mathbf{y} \in U$, the function $Z_n(\cdot; \mathbf{y})$ reaches a (local) maximum at $\mathcal{B}(\mathbf{y})$. Given $\mathbf{y} \in \mathbb{R}^n$, we usually write $\hat{\boldsymbol{\beta}} = \mathcal{B}(\mathbf{y})$.

Also the function $\boldsymbol{\beta} \to -Z_n(\boldsymbol{\beta})$ is said to be coercive if

$$\lim_{\|\boldsymbol{\beta}\| \to +\infty} -Z_n(\boldsymbol{\beta}) = +\infty.$$

Since $J(\boldsymbol{\beta})$ is non-negative the function $\boldsymbol{\beta} \to J(\boldsymbol{\beta})$ is bounded by below. If in addition $\boldsymbol{\beta} \to L_{\mathbf{y}}(\boldsymbol{\beta})$ is bounded by above, then $-Z_n$ is coercive if at least one of the two terms J or $-L_{\mathbf{y}}$ is coercive. It is easy to see that for the Gaussian and the Poisson non-parametric GLM models, $-Z_n$ is coercive. This is not the case for the Bernoulli model, the addition of a suitable penalty term (for example a quadratic term) to $J(\boldsymbol{\beta})$ makes $-Z_n$ coercive. Note that such an approach has also been proposed by Park and Hastie (2007) to handle the case of separable data in Bernoulli models.

Under the assumption that $-Z_n$ is coercive, for every $c \in \mathbb{R}$, the set $\{\boldsymbol{\beta} : -Z_n(\boldsymbol{\beta}) \leq c\}$ is bounded. Whenever Z_n is continuous the value $\sup_{\boldsymbol{\beta}} Z_n$ is finite and the set of the optimal solutions, namely

$$\left\{ \hat{\boldsymbol{\beta}} \in \mathbb{R}^p : Z_n(\hat{\boldsymbol{\beta}}) = \sup_{\boldsymbol{\beta} \in \mathbb{R}^p} Z_n \right\} \tag{8}$$

is *non-empty and compact* (see e.g. Rockafellar and Wets 1997, p. 11). In general, beyond its global maxima, Z_n may exhibit other local maxima. However, if in addition Z_n is concave, then Z_n does not have any local minimum which is not global, and the set of the optimal solutions (8) is convex. If moreover Z_n is strictly concave, then for every $\mathbf{y} \in \mathbb{R}^n$, the set in (8) is a singleton, hence there is a unique maximizer function $\mathcal{B}$ and its domain of definition is $\mathbb{R}^n$.

Analyzing the maximizers of Z_n when the latter is not concave is much more difficult. In the Gaussian case and J non-convex, the regularity of local and global maximizers of Z_n has been studied by Durand and Nikolova (2006a,b) and Nikolova (2005).

In our setup of non-parametric GLM models, we consider two situations. First, we study the optimization problem with penalties belonging to the class of symmetric and non-negative functions ψ satisfying the following properties:

- ψ is in $\mathcal{C}^2$ and convex on $[0, +\infty[$
- $t \to \psi(\sqrt{t})$ is concave on $[0, +\infty[$

- $\psi'(t)/t \to M < \infty$ as $t \to \infty$
- $\lim_{t \nearrow 0} \psi'(t)/t$ exists.

For such a class, referred to as Geman's class we shown the existence of a unique solution and discuss a computational algorithm to find it. Penalties belonging to this class among the ones displayed in Table 1 are 2, 3, 4 and 5.

The second class we consider is made by symmetric and non-negative penalty functions ψ such that

- ψ is monotone increasing on $[0, +\infty[$.
- ψ is in $\mathcal{C}^1$ on $\mathbb{R}\backslash\{0\}$ and continuous in 0.
- $\lim_{t \to 0} \psi'(t)t = 0$.

Note that many of the penalties in Tables 1 and 2 belong to this class. Such a class will be named a δ-class since it essentially consists of penalties that are non-smooth at the origin but can be approximated by a quadratic function at a δ-neighbor of the origin. For this class we find an approximate solution to the optimization problem and provide bias and variance expression for the resulting estimators in Sect. 5.

4.1 Optimization with penalties in Geman's class

We now study the optimization problem with penalties belonging to Geman's class. A typical example of a penalty belonging to this class is the penalty $\psi(\beta) = \sqrt{\beta^2 + \alpha}$ in Table 1. For such penalties the penalized log-likelihood Z_n is smooth and concave, and the penalized maximum likelihood solutions can be obtained using standard optimization numerical algorithms such as relaxation, gradient or conjugated gradient. Note, however, that even if the penalties ψ are convex, their second derivative is large near to zero and almost null beyond, so the optimization of the corresponding Z_n may be slow. For this reason, specialized optimization schemes have been conceived.

A very successful approach is *half-quadratic optimization*, proposed in two different forms (called multiplicative and additive) in Geman and Reynolds (1992) and Geman and Yang (1995), for Gaussian distributed data. The idea is to associate with every $d_k^T \boldsymbol{\beta}$ in (5) an auxiliary variable $b_k \in \mathbb{R}$ and to construct an augmented criterion $K_{\mathbf{y}}$ of the form

$$K_{\mathbf{y}}(\boldsymbol{\beta}, \mathbf{b}) = L_{\mathbf{y}}(\boldsymbol{\beta}) - \lambda R(\boldsymbol{\beta}, \mathbf{b}), \tag{9}$$

with

$$R(\boldsymbol{\beta}, \mathbf{b}) = \sum_k \gamma_k \left(Q\left(d_k^T \boldsymbol{\beta}, b_k\right) + \phi(b_k) \right), \tag{10}$$

where for every $\mathbf{b}$ fixed, the function $\boldsymbol{\beta} \to Q(\boldsymbol{\beta}, \mathbf{b})$ is quadratic, and ϕ—the dual function of ψ—is such that for every $\boldsymbol{\beta}$,

$$\inf_{\mathbf{b}} R(\boldsymbol{\beta}, \mathbf{b}) = J(\boldsymbol{\beta}). \tag{11}$$

The last condition ensures that if $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{b}})$ is a maximizer of $K_{\mathbf{y}}$, then $\hat{\boldsymbol{\beta}}$ is a maximizer of $Z_n(\boldsymbol{\beta}) = Z_n(\boldsymbol{\beta}; \mathbf{y})$ as defined by (4) and (5). The interest is that for every $\mathbf{b}$ fixed,

using a weighted local quadratic approximation of the log-likelihood as it is usually done in the iteratively re-weighted least squares (IRLS) approach for GLM's, for each iteration in the IRLS algorithm, the function $\boldsymbol{\beta} \rightarrow K_{\mathbf{y}}(\boldsymbol{\beta}, \mathbf{b})$ is quadratic (hence quadratic programming can be used) whereas for every $\boldsymbol{\beta}$ fixed, each b_k can be computed independently using an explicit formula. At each iteration one realizes an optimization with respect to $\boldsymbol{\beta}$ for $\mathbf{b}$ fixed and a second with respect to $\mathbf{b}$ for $\boldsymbol{\beta}$ fixed. Geman and Reynolds (1992) first considered quadratic terms of the multiplicative form,

$$Q(t, b) = t^2 b.$$

Later, Geman and Yang (1995) proposed quadratic terms of the additive form,

$$Q(t, b) = (t - b)^2.$$

In both cases, the dual function ϕ which gives rise to (11) is obtained using the theory of convex conjugate functions (see for example Rockafellar 1970). These ideas have been pursued and convergence to the sought-after solution under appropriate conditions have been considered by many authors. For example, rate of convergence of the iterations has been examined by Nikolova and Ng (2005). Half-quadratic regularization (9) and (10) may also be used with smooth non-convex penalties as well (see Delanay and Bressler 1998).

4.2 Optimization with penalties in the δ-class

To deal with penalties in the δ-class and especially with the non-differentiabilty at zero of such penalties, we will use an approximation $Z_\delta(\boldsymbol{\beta})$ of the penalized log-likelihood $Z_n(\boldsymbol{\beta})$, replacing the penalty $J(\boldsymbol{\beta}) = \sum_k \gamma_k \psi(\beta_k)$ in (3) by $J_\delta(\boldsymbol{\beta}) = \sum_k \gamma_k \psi_\delta(\beta_k)$, where ψ_δ is a function which is equal to ψ away from 0 (at a distance $\delta > 0$) and is a "smooth quadratic" version of ψ in a δ-neighborhood of zero (see for example Tishler and Zang 1981). More precisely, it is easy to see that, given the conditions on ψ, when ψ belongs to the δ-class, the function ψ_δ may be defined by

$$\psi_\delta(s) = \begin{cases} \psi(s) & \text{if } s > \delta, \\ \frac{\dot{\psi}(\delta)}{2\delta} s^2 + [\psi(\delta) - \dot{\psi}(\delta)\delta/2] & \text{if } 0 \leq s \leq \delta. \end{cases} \tag{12}$$

Note also that

$$\ddot{\psi}_\delta(s) = \begin{cases} \ddot{\psi}(s) & \text{if } s > \delta, \\ \dfrac{\dot{\psi}(\delta)}{\delta} & \text{if } 0 \leq s \leq \delta, \end{cases}$$

and, given the conditions on ψ, we have, for all $s \geq 0$

$$\lim_{\delta \downarrow 0} \psi_\delta(s) = 0.$$

Using the above approximate penalized log-likelihood $Z_\delta(\boldsymbol{\beta})$ the estimating equations for $\boldsymbol{\beta}$ can be derived by looking at the score function

$$u_\delta(\boldsymbol{\beta}) = s(\mathbf{y}, \boldsymbol{\beta}) + \lambda D(\boldsymbol{\gamma}) \mathbf{g}_\delta(\boldsymbol{\beta}), \tag{13}$$

where $s(\mathbf{y}, \boldsymbol{\beta}) = (\partial L_\mathbf{y}(\boldsymbol{\beta}) / \partial \beta_j)_{j=1,\dots,p}$ and $\mathbf{g}_\delta(\boldsymbol{\beta})$ denotes the $(p \times 1)$ vector with corresponding jth component $g_\delta(|\beta_j|)$ defined by

$$g_\delta(|\beta_j|) = \begin{cases} -\dot{\psi}_\delta(|\beta_j|) & \text{if } \beta_j \geq 0 \\ +\dot{\psi}_\delta(|\beta_j|) & \text{if } \beta_j < 0. \end{cases}$$

Note also that, for any $\boldsymbol{\beta}$ fixed, by definition of ψ_δ,

$$\lim_{\delta \downarrow 0} \mathbf{g}_\delta(\boldsymbol{\beta}) = \mathbf{g}(\boldsymbol{\beta}),$$

where $\mathbf{g}(\boldsymbol{\beta}) = (g(|\beta_1|), \dots, g(|\beta_p|))^T$ with $g(|\beta_j|) = -\dot{\psi}(\beta_j) I\{\beta_j \neq 0\}$. It follows that $u_\delta(\boldsymbol{\beta})$ converges to $u(\boldsymbol{\beta})$ as $\delta \downarrow 0$, where

$$u(\boldsymbol{\beta}) = s(\mathbf{y}, \boldsymbol{\beta}) + \lambda D(\boldsymbol{\gamma}) \mathbf{g}(\boldsymbol{\beta}).$$

Let $\hat{\boldsymbol{\beta}}(\delta)$ be a root of the approximate penalized score equations above, i.e. such that $u_\delta(\hat{\boldsymbol{\beta}}(\delta)) = 0$. By penalization, and since the penalty function ψ_δ is strictly convex, such an estimator exists and is unique even in situations where the maximum likelihood principle diverges. Fast computation of the estimator can be done by a standard Fisher scoring procedure.

5 Statistical properties of the estimates

Under the set of assumptions concerning the δ-class of penalties, we have shown in the previous section, that, provided δ vanishes at an appropriate rate, the resulting estimator converges to the penalized likelihood estimator, whenever it exists. The purpose of the next section is to study the quality of this approximation in statistical terms in more detail.

5.1 Bias and variance $p < n$ for δ-class penalties

We will first derive some approximations to the variance and bias of our penalized estimators when $p < n$, which allow to study their small sample properties. Using the same notation as in the previous section, suppose that the diagonal matrix $D(\boldsymbol{\gamma})$ of weights and the penalization parameter λ are fixed and let $\boldsymbol{\beta}^*$ be a maximizer of the expected penalized log-likelihood. In the case of uniqueness, this is equivalent to the root of the expected penalized score equation, i.e. $\mathbb{E}(u(\boldsymbol{\beta}^*)) = 0$. Let us then consider

the estimation error induced by our regularized procedure. A linear Taylor expansion of $u_\delta(\hat{\boldsymbol{\beta}}(\delta))$ gives

$$0 = u_\delta(\hat{\boldsymbol{\beta}}(\delta)) \approx u_\delta(\boldsymbol{\beta}^*) + \left(H(\boldsymbol{\beta}^*) + \lambda D(\boldsymbol{\gamma})G(\boldsymbol{\beta}^*; \delta)\right)(\hat{\boldsymbol{\beta}}(\delta) - \boldsymbol{\beta}^*),$$

where $H(\boldsymbol{\beta}) = \left(\frac{\partial^2 L_{\mathbf{y}}(\boldsymbol{\beta})}{\partial \beta_j \beta_l}\right)_{j,l=1,\dots,p}$ is the Hessian matrix, and where $G(\boldsymbol{\beta}^*; \delta)$ is the diagonal matrix with diagonal entries $\partial g_\delta(|\beta_j|)/\partial\beta_j = -\ddot{\psi}_\delta(|\beta_j|)$. Using the properties of ψ_δ, from the above Taylor approximation we get

$$\hat{\boldsymbol{\beta}}(\delta) - \boldsymbol{\beta}^* \approx -\left(H(\boldsymbol{\beta}^*) + \lambda D(\boldsymbol{\gamma})G(\boldsymbol{\beta}^*; \delta)\right)^{-1} u_\delta(\boldsymbol{\beta}^*) \tag{14}$$

and therefore

$$\mathbb{E}(\hat{\boldsymbol{\beta}}(\delta)) \approx \boldsymbol{\beta}^* - \left(H(\boldsymbol{\beta}^*) + \lambda D(\boldsymbol{\gamma})G(\boldsymbol{\beta}^*; \delta)\right)^{-1} \mathbb{E}(u_\delta(\boldsymbol{\beta}^*)). \tag{15}$$

Since $\boldsymbol{\beta}^*$ is a root of $\mathbb{E}(u(\boldsymbol{\beta}))$ it follows from (13) that $\mathbb{E}(u_\delta(\boldsymbol{\beta}^*)) = \lambda D(\boldsymbol{\gamma})\mathbf{g}_\delta(\boldsymbol{\beta}^*)$ and therefore the estimator $\hat{\boldsymbol{\beta}}(\delta)$ has a bias $-\left(H(\boldsymbol{\beta}^*) + \lambda D(\boldsymbol{\gamma})G(\boldsymbol{\beta}^*; \delta)\right)^{-1} \mathbb{E}(u_\delta(\boldsymbol{\beta}^*))$.

As for the variance, using again the above approximations we obtain

$$\mathrm{var}(\hat{\boldsymbol{\beta}}(\delta)) = \left(H(\boldsymbol{\beta}^*) + \lambda D(\boldsymbol{\gamma})G(\boldsymbol{\beta}^*; \delta)\right)^{-1} \mathrm{var}(s(\mathbf{y}, \boldsymbol{\beta}^*)) \left(H(\boldsymbol{\beta}^*)\right.$$
$$\left. + \lambda D(\boldsymbol{\gamma})G(\boldsymbol{\beta}^*; \delta)\right)^{-1}$$

which has the well-known sandwich form of Hubert (1967).

Therefore, the bias and the variance of our estimator depend on the behavior of the eigenvalues of $(H(\boldsymbol{\beta}^*) + \lambda D(\boldsymbol{\gamma})G(\boldsymbol{\beta}^*; \delta))^{-1}$ and their limits as $\delta \downarrow 0$ with $\lambda > 0$ fixed.

In the case where $\boldsymbol{\beta}_j^* > \delta$ for all j, we have $G(\boldsymbol{\beta}^*; \delta) = \mathrm{diag}(\ddot{\psi}(|\beta_j^*|))$. If we assume that ψ is such that $\max\{\gamma_j|\ddot{\psi}(|\beta_j^*|)|; \beta_j^* \neq 0\} \to 0$, the asymptotic variance of the estimator becomes

$$\mathrm{var}(\hat{\boldsymbol{\beta}}(\delta)) = H(\boldsymbol{\beta}^*)^{-1}\mathrm{var}(s(\mathbf{y}, \boldsymbol{\beta}^*))H(\boldsymbol{\beta}^*)^{-1}.$$

When $\beta_j^* \leq \delta$ for some coefficient then $G(\beta_j^*, \delta) = -\dot{\psi}(\delta)/\delta$ and all depends on the speed at which $\dot{\psi}(\delta)/\delta$ goes to zero. If $\dot{\psi}(\delta)/\delta$ tends to infinity as δ goes to zero then by increasing the diagonal elements of $\left(H(\boldsymbol{\beta}^*) + \lambda D(\boldsymbol{\gamma})G(\boldsymbol{\beta}^*; \delta)\right)$, the diagonal elements of its inverse decrease, resulting in a reduced variance for $\hat{\boldsymbol{\beta}}(\delta)$. The penalty parameter λ tunes this variance reduction by controlling the eigenvalues of $(H(\boldsymbol{\beta}^*) + \lambda D(\boldsymbol{\gamma})G(\boldsymbol{\beta}^*; \delta))^{-1}$. In this case, and in the limit $\delta \to 0$, the diagonal elements of $G(\boldsymbol{\beta}^*; \delta)$ corresponding to $\beta_j^* = 0$, tend to infinity and the limiting covariance becomes singular approximating the components having $\beta_j^* = 0$ by 0. For the remaining components it leads to an approximate variance given by the corresponding diagonal entries of $H(\boldsymbol{\beta}^*)^{-1}\mathrm{var}(s(\mathbf{y}, \boldsymbol{\beta}^*))H(\boldsymbol{\beta}^*)^{-1}$.

While the above approximations are useful, we can now proceed to a general asymptotic analysis of our estimators.

5.2 Asymptotic analysis

In order to get a better insight into the properties of the estimators and to provide a basis for inference we will consider in this section asymptotic results of estimators $\hat{\boldsymbol{\beta}}_n$ minimizing $-Z_n(\boldsymbol{\beta})$ defined in (3). Before stating the results we will first state some regularity conditions on the log-likelihood that are standard regularity conditions for asymptotic analysis of maximum likelihood estimators (see e.g. Fahrmeir and Kaufman (1985)). We will first examine the case of a fixed finite-dimensional approximating basis (p finite) and then proceed to the general case of an increasing dimension sequence of approximating subspaces. Most of the proofs of the presented results are inspired by similar proofs made for model selection in regression models in Fan and Li (2001), Antoniadis and Fan (2001) and Fan and Peng (2004) but are tailored here to our special setup with minor modifications. We include one proof below for sake of completeness.

5.2.1 Limit theorems in the parametric case (fixed number of parameters)

We will first state here some regularity conditions for the log-likelihood.

Regularity conditions

(a) The probability density of the observations has a support that does not depend on $\boldsymbol{\beta}$ and the model is identifiable. Moreover, we assume that $\mathbb{E}_{\boldsymbol{\beta}}(s(\mathbf{Y}, \boldsymbol{\beta})) = 0$ and that the Fisher information matrix exists and is such that

$$I_{j,k}(\boldsymbol{\beta}) = \mathbb{E}(s_j(\mathbf{Y}, \boldsymbol{\beta}) s_k(\mathbf{Y}, \boldsymbol{\beta})) = \mathbb{E}_{\boldsymbol{\beta}}\left(-\frac{\partial^2}{\partial\beta_j \partial\beta_k} L_{\mathbf{Y}}(\boldsymbol{\beta})\right).$$

(b) The Fisher information matrix $I(\boldsymbol{\beta})$ is finite and positive definite at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ where $\boldsymbol{\beta}_0$ is the true vector of coefficients.

(c) There exists an open set Ω of the parameter set containing the true parameter $\boldsymbol{\beta}_0$ such that, for almost all (y, x)'s the density $\exp L_{\mathbf{y}}(\boldsymbol{\beta})$ admits all third derivatives (with respect to $\boldsymbol{\beta}$) for all $\boldsymbol{\beta} \in \Omega$ and

$$\left|\frac{\partial^3}{\partial\beta_i \partial\beta_j \partial\beta_k} L_{\mathbf{Y}}(\boldsymbol{\beta})\right| \leq M_{i,j,k}(\mathbf{Y}) \quad \forall \boldsymbol{\beta} \in \Omega,$$

with $\mathbb{E}_{\boldsymbol{\beta}_0}(M_{i,j,k}(\mathbf{Y})) < +\infty$.

These are standard regularity conditions that usually guarantee asymptotic normality of ordinary maximum likelihood estimates. Let $a_n = \lambda_n \max\{\gamma_j \dot{\psi}(|\beta_{0j}|); \beta_{0j} \neq 0\}$ which is finite. Then, we have

Theorem 1 *Let the probability density of our model satisfy the regularity conditions (a), (b) and (c). Assume also that $\lambda_n \to 0$ as $n \to \infty$. If $b_n := \lambda_n \max\{\gamma_j |\ddot{\psi}(|\beta_{0j}|)|; \beta_{0j} \neq 0\} \to 0$, then there exists a local minimizer $\hat{\boldsymbol{\beta}}_n$ of the penalized likelihood such that $\|\hat{\boldsymbol{\beta}} - \beta_0\| = O_P(n^{-1/2} + a_n)$, where $\|\cdot\|$ denotes the Euclidian norm of $\mathbb{R}^p$.*

It is now clear that by choosing appropriately λ_n and the γ_j's, there exists a root-n consistent estimator of $\boldsymbol{\beta}_0$.

Proof First note that the result will follow if for all $\epsilon > 0$, there exists a large enough constant C_ϵ such that

$$\mathbb{P}\left\{ \sup_{\|\mathbf{u}\|=C_\epsilon} Z_n(\boldsymbol{\beta}_0 + \alpha_n\mathbf{u}) < Z_n(\boldsymbol{\beta}_0) \right\} \geq 1 - \epsilon.$$

In order to prove the above, let

$$W_n(\mathbf{u}) := Z_n(\boldsymbol{\beta}_0 + \alpha_n\mathbf{u}) - Z_n(\boldsymbol{\beta}_0).$$

Using the expression of $Z_n(\boldsymbol{\beta})$, we have

$$W_n(\mathbf{u}) : = Z_n(\boldsymbol{\beta}_0 + \alpha_n\mathbf{u}) - Z_n(\boldsymbol{\beta}_0) = \frac{1}{n}L\mathbf{y}(\boldsymbol{\beta}_0 + \alpha_n\mathbf{u}) - \frac{1}{n}L\mathbf{y}(\boldsymbol{\beta}_0)$$

$$- \lambda_n \sum_{j=1}^{p} \gamma_j \{\psi(|\beta_{0j} + \alpha_n u_j|) - \psi(|\beta_{0j}|)\}.$$

Given our regularity assumptions, a Taylor's expansion of the likelihood function $L\mathbf{y}$ and a Taylor's expansion for ψ give

$$W_n(\boldsymbol{\beta}_0) = \frac{1}{n}\alpha_n\nabla L\mathbf{y}(\boldsymbol{\beta}_0)^T\mathbf{u} - \frac{1}{2}\mathbf{u}^T\mathcal{J}_n(\boldsymbol{\beta}_0)\mathbf{u}\frac{1}{n}\alpha_n^2 + o\left(\mathbf{u}^T\mathcal{J}_n(\boldsymbol{\beta}_0)\mathbf{u}\frac{1}{n}\alpha_n^2\right)$$

$$- \sum_{j=1}^{p} \gamma_j \left\{ \lambda_n\alpha_n\dot{\psi}(|\beta_{0j}|)\mathrm{sgn}(\beta_{0j})u_j + \lambda_n\alpha_n^2\ddot{\psi}(|\beta_{0j}|)u_j^2(1 + o(1)) \right\},$$

where $\mathcal{J}_n(\boldsymbol{\beta}_0) = [\partial^2 L\mathbf{y}(\boldsymbol{\beta}_0)/\partial\beta_i\partial\beta_j]$ denotes the observed information matrix. By assumption (b) and the Law of large numbers we have that $\nabla L\mathbf{y}(\boldsymbol{\beta}_0) = O_P(\sqrt{n})$ and also that $\mathcal{J}_n(\boldsymbol{\beta}_0) = nI(\boldsymbol{\beta}_0) + o_P(n)$. It follows that

$$W_n(\boldsymbol{\beta}_0) = \frac{1}{n}\alpha_n\nabla L\mathbf{y}(\boldsymbol{\beta}_0)^T\mathbf{u} - \frac{1}{2}\mathbf{u}^T I(\boldsymbol{\beta}_0)\mathbf{u}\alpha_n^2\{1 + o_P(1)\}$$

$$- \sum_{j=1}^{p} \gamma_j \left\{ \lambda_n\alpha_n\dot{\psi}(|\beta_{0j}|)\mathrm{sgn}(\beta_{0j})u_j + \lambda_n\alpha_n^2\ddot{\psi}(|\beta_{0j}|)u_j^2(1 + o(1)) \right\}.$$

The first term on the right-hand side of the above equality is of the order $O_P(n^{-1/2}\alpha_n)$ and the second term of the order $O_P(\alpha_n^2)$. By choosing a sufficiently large C_ϵ the second term dominates the first one, uniformly in $\mathbf{u}$ such that $\|\mathbf{u}\| = C_\epsilon$. Now the third term is bounded above by

$$\sqrt{p}\left\|\mathbf{u}\right\|\left\|a_n\alpha_n + \alpha_n^2 b_n\right\|\left\|\mathbf{u}\right\|^2,$$

which is also dominated by the second term of order $O_P(\alpha_n^2)$. By choosing therefore a large enough C_ϵ the theorem follows. □

5.2.2 Limit theorems when $p \to \infty$

In the previous subsection we have considered the case where the dimension of the spline bases was fixed and finite. We now consider the case where the dimension may grow as the sample size increases, as this allows a better control of the approximation bias when the regression function is irregular. This case has been extensively studied by Fan and Peng (2004) for some non-concave penalized likelihood function. The proofs can be adapted to the various penalties that we have considered in this paper with little effort and some extra regularity conditions and follow closely (with appropriate modifications) the proof of the previous subsection. We will not present the proofs and the interested reader is referred to the above-mentioned paper. We only would like here to state the extra conditions that allows us to extend the results to the case of a growing dimension.

We will need the following extra regularity conditions on the penalty and on the rate of growth of the dimension p_n.

Regularity conditions

(a) $\liminf_{\beta \to 0^+} \dot{\psi}(\beta) > 0$
(b) $a_n = O(n^{-1/2})$
(c) $a_n = o((np_n)^{-1/2})$
(d) $b_n = \max_{1 \leq j \leq p_n} \{\gamma_j |\ddot{\psi}(|\beta_j|)|; \beta_j \neq 0\} \to 0$
(e) $b_n = o_P(p_n^{-1/2})$
(f) There exist C and D such that when x_1 and $x_2 > C\lambda_n$,

$$\lambda_n |\ddot{\psi}(x_1) - \ddot{\psi}(x_2)| \leq D|x_1 - x_2|.$$

Under such conditions the theorem of the previous section extends to the case with $p_n \to \infty$. Conditions (b) and (c) allow us to control the bias when $p_n \to \infty$ and ensure the existence of the root-n consistent penalized likelihood estimators while conditions (d) and (e) dump somehow the influence of the penalties on the asymptotic behavior of the estimators. Condition (f) is a regularity assumption on the penalty away from 0 that allows an efficient asymptotic analysis.

6 Choosing the penalization parameters

Recall from (3) that the estimation procedure consists of minimizing

$$-\sum_{i=1}^{n} \ell\left(y_i, \mathbf{h}^T(x_i)\boldsymbol{\beta}\right) + n\lambda \sum_{k=1}^{p} \gamma_k \psi(\beta_k), \tag{16}$$

with respect to $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, where $\gamma_k, k = 1, \ldots, p$, are positive weights, and $\lambda > 0$ is a general smoothing parameter. Denote by

$$\rho_k = n\lambda\,\gamma_k, \quad k = 1, \ldots, p, \tag{17}$$

the *regularization parameters*. Then minimization problem (16) is equivalent to

$$-\sum_{i=1}^{n} \ell\left(y_i, \mathbf{h}^T(x_i)\boldsymbol{\beta}\right) + \sum_{k=1}^{p} \rho_k\psi\left(\beta_k\right). \tag{18}$$

For a given log-likelihood function $\ell(\cdot)$, and a given penalty function $\psi(\cdot)$ one needs to choose the regularization parameters $\rho_1, \ldots, \rho_p$. The choice of these parameters will have an influence on the quality of the estimators. To select this vector of parameters we propose to use an extension of the concept of L-curve. This idea of selecting the whole *vector* of regularization weights ρ_k is similar to the non-linear L-curve regularization method used for determining a proper regularization parameter in penalized non-linear least-squares problems (see Gullikson and Wedin 1998). In our context, the L-curve is a very efficient procedure that allows to choose a multidimensional hyperparameter. Such a choice seems to be computationally impossible to achieve using standard cross-validation procedures on a multidimensional grid.

We first briefly explain the basics of this L-curve approach for the particular case of a Gaussian likelihood (a quadratic loss function) and then discuss the extension to the current context of generalized linear models.

In our discussion of this section we assume that $\psi(\cdot)$ satisfies the assumptions:

- ψ is a continuously differentiable and convex function
- ψ is a non-negative and symmetric function
- ψ is such that $\psi'(t) \geq 0$, $\forall t \geq 0$
- $\lim_{t \nearrow 0} \psi'(t)/t = C$, with $0 < C < \infty$.

Functions ψ that satisfies these conditions are functions 1, 3 and 4 in Table 1.

6.1 Multiple regularization, L-curves and Gaussian likelihood

In case of a Gaussian likelihood, minimization problem (18) reduces to

$$\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - H(\mathbf{x})\boldsymbol{\beta}\|_2^2 + \sum_{k=1}^{p} \rho_k\psi\left(\beta_k\right) \right\}, \tag{19}$$

where $\mathbf{x} = (x_1, \ldots, x_n)^T$ and $H(\mathbf{x})$ is the matrix of dimension $n \times p$ built up of the rows $\mathbf{h}^T(x_i)$, $i = 1, \ldots, n$. Now denote by $\boldsymbol{\beta}^s$ the solution to the above minimization problem, i.e.

$$\boldsymbol{\beta}^s(\boldsymbol{\rho}) = \operatorname{argmin}_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - H(\mathbf{x})\boldsymbol{\beta}\|_2^2 + \sum_{k=1}^{p} \rho_k\psi\left(\beta_k\right) \right\}, \tag{20}$$

where $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_p)^T$. Putting

$$z(\boldsymbol{\rho}) = \|\mathbf{y} - H(\mathbf{x})\beta^s(\boldsymbol{\rho})\|_2^2 \quad \text{and} \quad z_k(\boldsymbol{\rho}) = \psi\left(\beta_k^s(\boldsymbol{\rho})\right) \quad k = 1, \ldots, p,$$

an optimal choice of the regularization parameters $(\rho_1, \ldots, \rho_p)$ would consists of choosing them such that the estimation error

$$\|\mathbf{y} - H(\mathbf{x})\boldsymbol{\beta}^s(\boldsymbol{\rho})\|_2^2 + \sum_{k=1}^{p} \rho_k \psi \left(\beta_k^s(\boldsymbol{\rho}) \right) = z(\boldsymbol{\rho}) + \sum_{k=1}^{p} \rho_k z_k(\boldsymbol{\rho}),$$

is minimal.

The L-hypersurface is defined as a subset of $\mathbb{R}^{p+1}$, associated with the map

$$L(\boldsymbol{\rho}) : \mathbb{R}^p \rightarrow \mathbb{R}^{p+1}$$
$$\boldsymbol{\rho}^T \mapsto \left(t[z_1(\boldsymbol{\rho})], \ldots, t[z_p(\boldsymbol{\rho})], t[z(\boldsymbol{\rho})] \right),$$

where $t(\cdot)$ is some appropriate scaling function such as

$$t(u) = \log(u) \quad \text{or} \quad t(u) = \sqrt{u}.$$

The L-hypersurface is a plot of the residual norm term $z(\boldsymbol{\rho})$ plotted against the constraint (penalty) terms $z_1(\boldsymbol{\rho}), \ldots, z_p(\boldsymbol{\rho})$ drawn in an appropriate scale. In the one-dimensional case, when $p = 1$, the L-hypersurface reduces to the L-curve, the curve that plots the perturbation error $t[z(\boldsymbol{\rho})]$ against the regularization error $t[z_1(\boldsymbol{\rho})]$ for all possible values of the parameter $\boldsymbol{\rho}$. The 'corner' of the L-curve corresponds to the point where the perturbation and regularization errors are approximately balanced. Typically to the left of the corner, the L-curve becomes approximately horizontal and to the right of the corner the L-curve becomes approximately vertical. Under regularity conditions on the L-curve, the corner is found as the point of maximal Gaussian curvature.

Similarly, in the multidimensional case ($p > 1$) the idea is to search for the point on the L-hypersurface for which the Gaussian curvature is maximal. Such a point represents a generalized corner of the surface, i.e. a point around which the surface is maximally warped. Examples and illustrations of L-curves and L-hypersurfaces, their Gaussian curvatures and (generalized) corners, can be found in Belge et al. (2002), in the framework of a regularized least-squares problem.

The Gaussian curvature of $L(\boldsymbol{\rho})$ can be computed via the first- and second-order partial derivatives of $t[z(\boldsymbol{\rho})]$ with respect to $t[z_k(\boldsymbol{\rho})]$, $1 \le k \le p$, and is given by

$$\kappa(\boldsymbol{\rho}) = \frac{(-1)^p}{w^{p+2}} \det(P), \tag{21}$$

where

$$w^2 = 1 + \sum_{k=1}^{p} \left(\frac{\partial t(z)}{\partial t(z_k)} \right)^2 \quad \text{and} \quad P_{k,l} = \frac{\partial^2 t(z)}{\partial t(z_k) \, \partial t(z_l)},$$

with the derivatives calculated at the point $(z_1(\boldsymbol{\rho}), \ldots, z_p(\boldsymbol{\rho}), z(\boldsymbol{\rho}))$.

Evaluating the Gaussian curvature in (21) for a large number of regularization parameters in searching for the point where we get maximal Gaussian curvature, is computationally expensive. In addition, the use of conventional optimization techniques to locate the maximum Gaussian curvature point might run into difficulties by the fact that the Gaussian curvature often possesses multiple local extrema.

A way out to these difficulties is to look at an approximate optimization problem that is easier to solve, and that is such that the solutions from both optimization problems (the exact and the approximate) are close. As discussed in Belge et al. (2002), an appropriate surrogate function is to look at the minimum distance function (MDF), defined as the distance from the origin $(a, b_1, \ldots, b_p)$ of the coordinate system to the point $L(\boldsymbol{\rho})$ on the L-hypersurface:

$$v(\boldsymbol{\rho}) = \|t[z(\boldsymbol{\rho})] - a\|^2 + \sum_{k=1}^{p} \|t[z_k(\boldsymbol{\rho})] - b_k\|^2. \tag{22}$$

Intuititively, this distance is minimal for a point close to the 'corner' of the L-hypersurface. The minimum distance point is then defined as

$$\boldsymbol{\rho}^s = \mathrm{argmin}_{\boldsymbol{\rho}} v(\boldsymbol{\rho}). \tag{23}$$

The relationship between this minimum MDF point and the point of maximum Gaussian curvature has been studied in Belge et al. (2002), showing in particular the proximity of the two points for the case of the L-curve.

Finding the minimum distance point, defined in (23), can be done via a fixed-point approach. For the scale function $t(u) = \log(u)$ this leads to the following iterative algorithm to approximate $\boldsymbol{\rho}^s$:

$$\rho_k^{(j+1)} = \frac{z(\boldsymbol{\rho}^{(j)})}{z_k(\boldsymbol{\rho}^{(j)})} \left(\frac{\log\{z_k(\boldsymbol{\rho}^{(j)})\} - b_k}{\log\{z(\boldsymbol{\rho}^{(j)})\} - a} \right), \quad k = 1, \ldots, p, \tag{24}$$

where $\boldsymbol{\rho}^{(j)}$ is the vector of the regularization parameters at step j in the iterative algorithm. The algorithm is started with an initial regularization parameter $\boldsymbol{\rho}^{(0)} = (\rho_1^{(0)}, \ldots, \rho_p^{(0)})^T$ and is then iterated until convergence (i.e. until the relative change in the iterates becomes sufficiently small).

6.2 Multiple regularization, L-curves and generalized linear models

Let us now look at the general case of a general log-likelihood function $\ell(\cdot)$ with contributions $\ell(y_i, \mathbf{h}^T(x_i)\boldsymbol{\beta})$ in the log-likelihood part. Now suppose that minus the log-likelihood function is a strict convex function and that we have a convex penalty function. Then the minimization problem in (16) will have a unique minimum. For given parameters $\lambda, \gamma_k, 1 \le k \le p$, this minimum is found by a kind of Newton–Raphson algorithm, resulting in the so-called Fisher scoring method. This procedure

is equivalent with an iteratively re-weighted least-squares procedure. See e.g. Hastie and Tibshirani (1990).

The procedure of Sect. 6.1 can now be extended to this more general case as follows. For given parameters λ, γ_k, $1 \leq k \leq p$ denote by $\boldsymbol{\beta}^s$ the solution retained at the last (convergent) step of the iteratively re-weighted least-squares procedure. For this (approximate) solution we then compute the error term (minus the log-likelihood term) as well as the regularization errors. With these we produce the corresponding L-hypersurface, and proceed as before.

6.3 Other selection methods

The selection method described in Sects. 6.1 and 6.2 provides a data-driven way for choosing the regularization parameters $\rho_1, \ldots, \rho_p$ as defined in (17).

An alternative approach to choose the parameters λ, $\gamma_1, \ldots, \gamma_p$ in optimization problem (16) is as follows. As mentioned by Klinger (2001), the penalized likelihood estimation procedure is not invariant under linear transformations of the basisfunctions $h_k(x)$: an estimate of $\widetilde{\beta}_j$, the coefficient which is associated to the basisfunction $\kappa[H(\mathbf{x})]_j$, does not equal $\widehat{\beta}_j/\kappa$, where $\widehat{\beta}_j$ is the estimated coefficient associated with the basisfunction $[H(\mathbf{x})]_j$. In other words, the estimated predictor depends on the scaling of the basisfunctions.

To overcome this possible drawback, one can standardize the basisfunctions in advance by considering

$$\overline{h}_j = \frac{1}{n} \sum_{i=1}^{n} h_j(x_i),$$

and by calculating

$$\widetilde{s}_j^2 = \frac{1}{n} \sum_{i=1}^{n} \left[h_j(x_i) - \overline{h}_j \right]^2 .$$

A possibility is then to adjust the threshold parameters γ_k appropriately by taking then equal to

$$\gamma_k = \sqrt{\widetilde{s}_k^2}.$$

With this choice, any scaled version $\kappa[H(\mathbf{x})]_j$ would yield the threshold $\widetilde{\gamma}_k = |\kappa|\gamma_k$. A data-driven choice of the parameters λ, $\gamma_1, \ldots, \gamma_p$ is obtained by choosing $\gamma_k = \sqrt{\widetilde{s}_k^2}$ and by selecting λ by generalized cross validation.

The two data-driven procedures can also be combined, by choosing λ and $\gamma_1, \ldots, \gamma_p$ as explained above, and then calculating from this the initial regularization parameter $\boldsymbol{\rho}^{(0)} = (\rho_1^{(0)}, \ldots, \rho_p^{(0)})^T$ as needed in the selection procedure of Sects. 6.1 and 6.2. These alternative selection procedures are not further explored in this paper.

7 Simulations and example

7.1 Simulation study

Here, we report some results on two sets of numerical experiments, which are part of an extensive simulation study that has been conducted to investigate into the properties of the penalization methods proposed in this paper and to compare them with some other popular approaches found in the literature. In all the experiments we have chosen some test functions that present either some jumps or some discontinuities in their derivatives. In each set of experiments, we have two inhomogeneous test functions, advocated by Donoho and Johnstone (1995) for testing several wavelet-based denoising procedures. The noise for the first set of experiments is assumed to be Gaussian and therefore the loss function considered for these experiments is quadratic. For the second set of experiments, we have used a Poisson noise, in order to illustrate the performances of the various procedure under GLM settings.

7.1.1 Quadratic loss

Data were generated from a regression model with two test functions. Both test functions named `heavisine` and `corner` are displayed in the figures below. Each basic experiment was repeated 100 times for a signal-to-noise (SNR) ratio of level 4. SNRs are defined as $\text{SNR} = \{\sqrt{\text{var(f)}/\sigma^2}\}^{1/2}$, with f the target function to be estimated, as in Donoho and Johnstone (1995). For each combination of test functions and each Monte Carlo simulation we have used the same design points x_i, $i = 1, \ldots, n$, obtained by simulating a uniform distribution on [0, 1]. To save some space we report here results for only $n = 200$ and SNR $= 4$, since results from other SNRs and n combinations are similar. Figures 3 and 4 respectively, depict a typical simulated data set with a SNR $= 4$, together with the corresponding mean function as the solid curve. Altogether four regularization procedures were compared. The four procedures are all based on regression splines and are RIDGE (quadratic loss and L_2-penalty on the coefficients), LASSO (quadratic loss and L_1 penalty on the coefficients), the spatially adaptive regression splines (SARS) developed by Zhou and Shen (2001) which is particularly suited for functions that have jumps by themselves or in their derivatives and HQ, the half-quadratic regularization procedure (quadratic loss and a penalty within the δ-class (penalty 2 in Table 1). We recall here that SARS is locally adaptive to variable smoothness and automatically places more knots in the regions where the function is not smooth, and has been proved as effective in estimating such functions.

For each simulated data set, the above cited smoothing procedures were applied to estimate the test functions. The numerical measure used to evaluate the quality of an estimated curve was the MSE, defined as $\text{MSE}(\hat{f}) = \frac{1}{n} \sum_{i=1}^{n} (\hat{f}(x_i) - f(x_i))^2$. Typical curve estimates for each test function obtained applying the four procedures on the same noisy data set are plotted in Figs. 3 and 4, where also boxplots of the $\text{MSE}(\hat{f})$ values of each $\hat{f}$ are also plotted. To compute the SARS estimates we have used the default values supplied by the code of Zhou and Shen for the hyperparameters that are required to be pre-selected. For other procedures we have used a maximum of 40 equispaced knots for the truncated power basis and the smoothing parameters
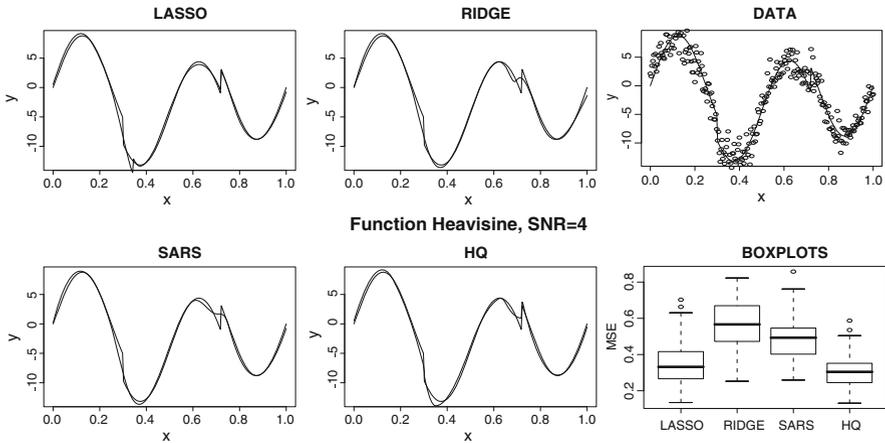
Fig. 3 A typical simulated data set with a SNR = 4, together with the corresponding `heavisine` function as a *solid curve*, together with typical fits obtained with the various regularization procedures and the *boxplots* of their MSE over 100 simulations
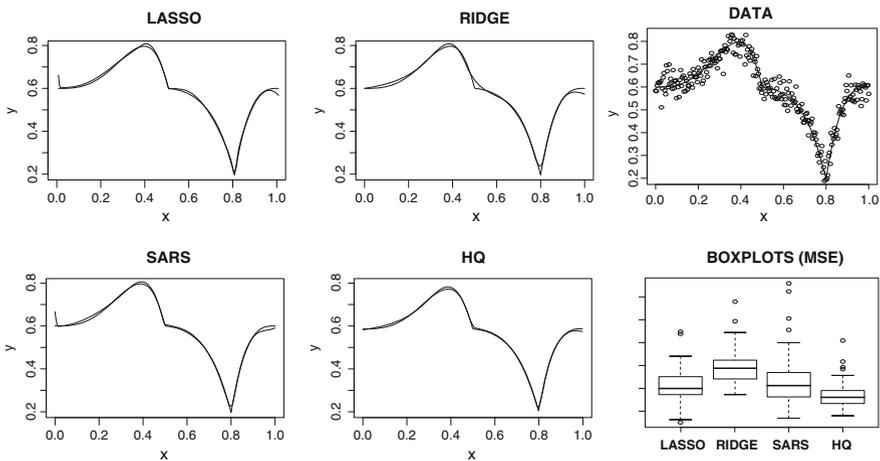

Fig. 4 A typical simulated data set with a SNR = 4, together with the corresponding `corner` function as a *solid curve*, together with typical fits obtained with the various regularization procedures and the *boxplots* of their MSE over 100 simulations

where selected by 10-fold generalized cross-validation. The threshold parameters γ_k were adjusted according to the average standard deviation of each basis function as advocated in Sect. 6. The L-curve criterion gave, in these simulation models, very similar results (and hence are not reported on here).

In view of our simulations, some empirical conclusions can be drawn. For the `heavisine` which presents two jump points, the boxplots in Fig. 3 suggest that LASSO and HQ have the smallest MSE and that the performance of LASSO depends heavily on the position of the retained knots. This is not surprising since a large observation can be easily mistaken as jump points by LASSO. The RIDGE and SARS
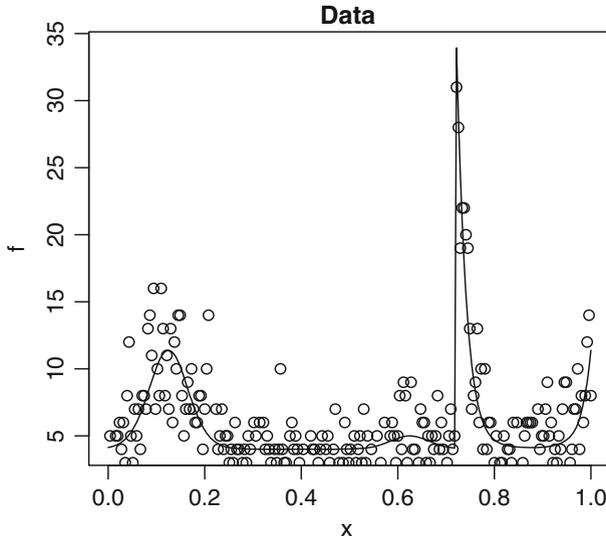
Fig. 5 A typical simulated Poisson regression data set together with the corresponding `exp (heavisine)` function as a *solid curve*

procedures perform roughly the same with an advantage to SARS which chooses the knots in a somewhat adaptive way. For the `corner` function all procedure give similar results with a preference in HQ with respect to MISE. The fact that all other three procedures are better than RIDGE is not surprising since RIDGE is not designed for recovering non-smooth functions.

7.1.2 Poisson regression

In the Monte Carlo simulation described here repeated random samples $\{(x_i, y_i), i = 1, \ldots, 200\}$ were generated form the Poisson regression model $Y_i \sim \text{Poisson}(f(x_i))$, with a function f given by the exponential of the `heavisine` function of the previous subsection. While the inhomogeneous character of this function is retained by exponentiation, the above intensity is ensured to be strictly positive on the design interval. For this case too, the basic experiment was repeated 100 times and for each Monte Carlo simulation we have used the same design points x_i, $i = 1, \ldots, n$, obtained by simulating a uniform distribution on [0, 1]. Figure 5 depicts a typical simulated data set, together with the true corresponding intensity function as a solid curve. Since SARS is not designed for treating Poisson distributed data, we have used instead a B-splines procedure based on an information criterion designed recently by Imoto and Konishi (2003) for our comparisons.

Three regularization procedures were compared. The RIDGE procedure based on a P-splines approach by Eilers and Marx (1996), our HQ procedure for the GLM case and finally the SPIC procedure by Imoto and Konishi (2003). The first two procedures are again based on regression splines with a maximum of 40 equispaced knots for the truncated power basis and the smoothing parameters were all selected by 10-fold
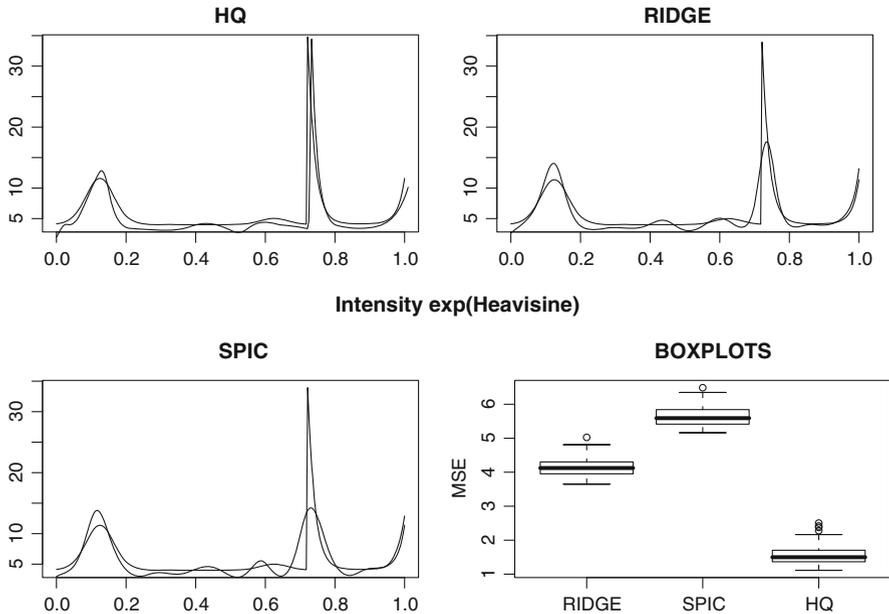
Fig. 6 Typical fits obtained with the various regularization procedures and the *boxplots* of their MSE over 100 simulations

cross-validation. The SPIC procedure is based on B-splines with 30 knots and the smoothing parameter for the Imoto and Konishi (2003) method is selected by their SPIC-criterion.

In view of these simulations and the boxplots in Fig. 6 the HQ procedure has the smallest MSE (due to a better tracking of the discontinuity and a less biased estimation of the small bump near to the point 0.64) followed by the RIDGE regression and the SPIC procedure. The RIDGE and SPIC procedures perform roughly the same, and seem to over-smooth the peak.

7.2 Analysis of real data

We illustrate our proposed procedure (HQ) through the analysis of the AIDS data (Stasinopoulos and Rigby 1992). This data set concerns the quarter yearly frequency count of reported AIDS cases in the UK from January 1983 to September 1990 and is reproduced in Stasinopoulos and Rigby (1992). As suggested by these authors, after deseasonalising this time series, one suspects a break in the relationship between the number of AIDS cases and the time measured in quarter years.

We model the dependent variable Y (deseasonalised frequency of AIDS cases) by a Poisson distribution with mean a polynomial spline function of x, the time measured in quarter years and have used the appropriate half quadratic procedure (HQ) with a spline basis based on 12 knots to fit the data. The deseasonalized data and the resulting fit, and the derivative of the fit are plotted in Fig. 7. One clearly sees a change in the
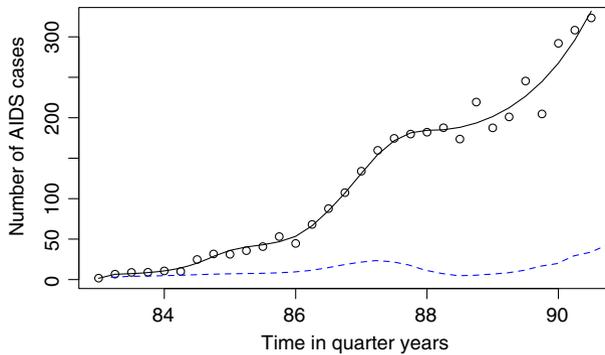
Fig. 7 Half Quadratic penalized fit (*solid line*) to the deseasonalized AIDS data. The *dashed line* is the derivative of the fitted regression curve

derivative at about July 1987. Stasinopoulos and Rigby (1992) suggested that this change was caused by behavioral changes in high risk groups. See also Jandhyala and MacNeill (1997) for an analysis of these data.

References

Alliney, S., Ruzinsky, S. A. (1994). An algorithm for the minimization of mixed l_1 and l_2 norms with application to Bayesian estimation. *IEEE Transactions on Signal Processing, 42*, 618–627.

Antoniadis, A., Fan, J. (2001). Regularization of wavelets approximations (with discussion). *Journal of the American Statistical Association, 96*, 939–967.

Belge, M., Kilmer, M. E., Miller, E. L. (2002). Efficient determination of multiple regularization parameters in a generalized L-curve approach. *Inverse Problems, 18*, 1161–1183.

Besag, J. E. (1989). Digital image processing: Towards Bayesian image analysis. *Journal of Applied Statistics, 16*, 395–407.

Biller, C. (2000). Adaptive Bayesian regression splines in semiparametric generalized linear models. *Journal of Computational and Graphical Statistics, 9*, 122–140.

Black, M., Rangarajan, A. (1996). On the unification of line processes, outlier rejection, and robust statistics with applications to early vision. *International Journal of Computer Vision, 19*, 57–91.

Ciarlet, P. G. (1989). *Introduction to numerical linear algebra and optimisation*. Cambridge Texts in Applied Mathematics. New York: Cambridge University Press.

Cox, D. D., O'Sullivan, F. (1990). Asymptotic analysis of penalized likelihood and related estimators. *The Annals of Statistics, 18*, 1676–1695.

de Boor, C. (1978). *A practical guide to splines*. New York: Springer.

Delanay, A. H., Bressler, Y. (1998). Globally convergent edge-presering regularized reconstruction: An application to limited-angle tomography. *IEEE Transactions on Image Processing, 7*, 204–221.

Demoment, G. (1989). Image-reconstruction and restoration—overview of common estimation structures and problems. *IEEE Transactions on Acoustics Speech and Signal Processing, 37*, 2024–2036.

Denison, D. G. T., Mallick, B. K., Smith, A. F. M. (1998). Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society, Series B, 60*, 333–350.

Dierckx, P. (1993). *Curve and surface fitting with splines*. Oxford Monographs on Numerical Analysis. New York: Oxford University Press.

DiMatteo, I., Genovese, C. R., Kass, R. E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika, 88*, 1055–1071.

Donoho, D., Johnstone, I., Hoch, J., Stern, A. (1992). Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society, Series B, 52*, 41–81.

Donoho, D. L., Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika, 81*, 425–455.

Donoho, D. L., Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association, 90*, 1200–1224.

Durand, S., Nikolova, M. (2006a). Stability of the minimizers of least squares with a non-convex regularization. Part I: Local behavior. *Applied Mathematics and Optimization, 53*, 185–208.

Durand, S., Nikolova, M. (2006b). Stability of the minimizers of least squares with a non-convex regularization. Part II: Global behavior. *Applied Mathematics and Optimization, 53*, 259–277.

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

Fahrmeir, L., Kaufman, H. (1985). Consistency of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics, 13*, 342–368.

Fahrmeir, L., Tutz, G. (1994). *Multivariate statistical modelling based on generalized linear models*. Springer Series in Statistics. New York: Springer.

Fan, J. (1997). Comments on "Wavelets in statistics: A Review", by A. Antoniadis. *Journal of the Italian Statistical Society, 6*, 131–138.

Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96*, 1348–1360.

Fan, J., Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics, 32*, 928–961.

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

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics, 19*, 1–67.

Friedman, J. H., Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling. *Technometrics, 31*, 3–21.

Fu, W. J. (1998). Penalized regressions: The bridge versus the Lasso. *Journal of Computational and Graphical Statistics, 7*, 397–416.

Geman, S., Geman, D. (1984). Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligenced, 6*, 721–741.

Geman, S., McClure, D. E. (1987). Statistical methods for tomographic image reconstruction. In *Proceedings of the 46-th session of the ISI*, Bulletin of the ISI, *52*, 22–26.

Geman, D., Reynolds, G. (1992). Constrained restoration and the recovery of discontinuities. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 14*, 367–383.

Geman, D., Yang, C. (1995). Nonlinear image recovery with half-quadratic regularization. *IEEE Transactions on Image Processing, 4*, 932–946.

Good, I. J., Gaskins, R. A. (1971). Nonparametric roughness penalities for probability densities. *Biometrika, 58*, 255–277.

Green, P. J., Yandell, B. (1985). Semi-parametric generalized linear models. In R. Gilchrist, B. J. Francis, J. Whittaker (Eds.), *Generalized linear models*, *Vol. 32*. Lecture notes in statistics (pp. 44–55). Berlin: Springer.

Gu, C., Kim, Y. J. (2002). Penalized likelihood regression: General formulation and efficient approximation. *The Canadian Journal of Statistics, 30*, 619–628.

Gu, C., Qiu, C. F. (1994). Penalized likelihood regression—a simple asymptotic analysis. *Statistica Sinica, 4*, 297–304.

Gullikson, M., Wedin, P.-A. (1998). Analyzing the nonlinear *L*-curve. *Technical Report*, Sweden: Department of Computer Science, University of Umeøa.

Hastie, T. J., Tibshirani, R. J. (1990). *Generalized additive models*. Monographs on Statistics and Applied Probability. New York: Chapman and Hall.

Hubert, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1*, 221–233.

Imoto, S., Konishi, S. (2003). Selection of smoothing parameters in B-spline nonparametric regression models using information criteria. *The Annals of the Institute of Statistical Mathematics, 55*, 671–687.

Jandhyala, V. K., MacNeill, I. B. (1997). Iterated partial sum sequences of regression residuals and tests for changepoints with continuity constraints. *Journal of the Royal Statistical Society, Series B, 59*, 147–156.

Klinger, A. (2001). Inference in high-dimensional generalized linear models based on soft-thresholding. *Journal of the Royal Statistical Society, Series B, 63*, 377–392.

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

Lindstrom, M. J. (1999). Penalized estimation of free-knot splines *Journal of Computational and Graphical Statistics, 8*, 333–352.

Mammen, E., Van de Geer, S. (1997). Locally adaptive regression splines. *The Annals of Statistics, 25*, 387–413.

McCullagh, P., Nelder, J. A. (1989). *Generalized linear models*, 2nd edn. London: Chapman & Hall.

Nikolova, M. (2000). Local strong homogeneity of a regularized estimator. *SIAM Journal of Applied Mathematics, 61*, 633–658.

Nikolova, M. (2004). Weakly constrained minimization. Application to the estimation of images and signals involving constant regions. *Journal of Mathematical Imaging and Vision, 21*, 155–175.

Nikolova, M. (2005). Analysis of the recovery of edges in images and signals by minimizing nonconvex regularized least-squares. *SIAM Journal on Multiscale Modeling and Simulation, 4*, 960–991.

Nikolova, M., Ng, M. K. (2005). Analysis of half-quadratic minimization methods for signal and image recovery. *SIAM Journal on Scientific Computing, 27*, 937–966.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion). *Statistical Science, 1*, 505–527.

O'Sullivan, F., Yandell, B. S., Raynor, W. J. (1986). Automatic smoothing of regression functions in generalized linear models. *Journal of the American Statistical Association, 81*, 96–103.

Park, M. Y., Hastie, T. (2007). L_1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society, Series B, 69*, 659–677.

Poggio, T. (1985). Early vision: From computational structure to algorithms and parallel hardware. *Computer Vision, Graphics, and Image Processing, 31*, 139–155.

Rockafellar, R. T. (1970). *Convex analysis*. New Jersy: Princeton University Press.

Rockafellar, T., Wets, R. J.-B. (1997). *Variational analysis*. Mathematics. Berlin: Springer.

Ruppert, D., Carroll, R. (1997). Penalized regression splines. *Working paper*, School of Operations Research and Industrial Engineering, Cornell University. Available at http://www.orie.cornell.edu/~davidr/papers.

Ruppert, D., Carroll, R. J. (2000). Spatially-adaptive penalties for spline fitting. *Australian and New Zealand Journal of Statistics, 42*, 205–223.

Stasinopoulos, D. M., Rigby, R. A. (1992). Detecting break points in generalized linear models. *Computational Statistics & Data Analysis, 13*, 461–471.

Stone, C. J., Hansen, M. H., Kooperberg, C., Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling. *The Annals of Statistics, 25*, 1371–1470.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B, 58*, 267–288.

Tikhonov, A. N. (1963). Solution of incorrectly formulated problems and the regularization method. *Soviet Math Dokl, 4*, 1035–1038. (English translation).

Tishler, A., Zang, I. (1981). A new maximum likelihood algorithm for piecewise regression. *Journal of the American Statistical Association, 76*, 980–987.

Titterington, D. M. (1985). Common structure of smoothing techniques in statistics. *International Statistical Review, 53*, 141–170.

Whittaker, E. (1923). On a new method of graduation. *Proceedings of Edinburgh Mathematical Society, 41*, 63–75.

Yu, Y., Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association, 97*, 1042–1054.

Yuan, M., Lin, Y. (2007). On the non-negative garrotte estimator. *Journal of the Royal Statistical Society, Series B, 69*, 143–161.

Zhao, P., Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research, 7*, 2541–2563.

Zhou, S., Shen, X. (2001). Spatially adaptive regression splines and accurate knot selection schemes. *Journal of the American Statistical Association, 96*, 247–259.

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