TEST (2023) 32:1055–1078
https://doi.org/10.1007/s11749-023-00866-x

ORIGINAL PAPER

Robust and efficient estimation of nonparametric generalized linear models

Ioannis Kalogridis[1] · Gerda Claeskens[2] · Stefan Van Aelst[1]

Received: 15 June 2022 / Accepted: 19 April 2023 / Published online: 16 May 2023

Abstract

Generalized linear models are flexible tools for the analysis of diverse datasets, but the classical formulation requires that the parametric component is correctly specified and the data contain no atypical observations. To address these shortcomings, we introduce and study a family of nonparametric full-rank and lower-rank spline estimators that result from the minimization of a penalized density power divergence. The proposed class of estimators is easily implementable, offers high protection against outlying observations and can be tuned for arbitrarily high efficiency in the case of clean data. We show that under weak assumptions, these estimators converge at a fast rate and illustrate their highly competitive performance on a simulation study and two real-data examples.

Keywords Generalized linear model · Robustness · Penalized splines · Reproducing kernel Hilbert space · Asymptotics

Mathematics Subject Classification 62G08 · 62R20 · 62G35

✉ Ioannis Kalogridis
ioannis.kalogridis@kuleuven.be

Gerda Claeskens
gerda.claeskens@kuleuven.be

Stefan Van Aelst
stefan.vanaelst@kuleuven.be

[1] Department of Mathematics, KU Leuven, Leuven, Belgium

[2] ORStat and Leuven Statistics Research Centre, KU Leuven, Leuven, Belgium

1 Introduction

Based on data $(t_1, Y_1), \ldots, (t_n, Y_n)$ with fixed $t_i \in [0, 1]$, consider the generalized linear model (GLM) stipulating that

$$Y_i \sim F_{\theta_{0,i},\phi_0}, \quad (i = 1, \ldots, n), \tag{1}$$

where $F_{\theta_{0,i},\phi_0}$ is a discrete or continuous exponential family of distributions over $\mathbb{R}$. Here, $\theta_{0,i}$ is the “canonical" parameter depending on t_i and ϕ_0 is a common nuisance or scale parameter. Each $F_{\theta_{0,i},\phi_0}$ has density with respect to either the Lebesgue or counting measure of the form

$$f_{\theta_{0,i},\phi_0}(y) = \exp\left[\frac{y\theta_{0,i} - b(\theta_{0,i})}{\phi_0} + c(y, \phi_0)\right], \quad (i = 1, \ldots, n), \tag{2}$$

for some functions $b : \mathbb{R} \to \mathbb{R}$ and $c : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ that determine the shape of the density. In the classical formulation (see, e.g. McCullagh and Nelder 1983), the systematic component of the model is defined indirectly through the relation $G(b'(\theta_{0,i})) = \beta_0 + \beta_1 t_i$ for a known link function $G : \mathbb{R} \to \mathbb{R}$ and some unknown $(\beta_0, \beta_1) \in \mathbb{R}^2$. In the present paper, following Green and Silverman (1994), we avoid this heavy parametric assumption and simply require that $\theta_{0,i} = g_0(t_i)$ for a smooth function $g_0 : [0, 1] \to \mathbb{R}$ to be estimated from the data.

The limitations of parametric inference in GLMs have been previously noticed by several authors and a large number of broader alternatives have been proposed through the years. The interested reader is referred to the dedicated monographs of Green and Silverman (1994), Ruppert et al. (2003) and Gu (2013) for extensive relevant discussions and illustrative examples of nonparametric GLM estimators. From a practical standpoint, an important drawback of most of these methods is their reliance on a correctly specified probability model for all the data, whereas, as Huber and Ronchetti (2009, p. 3) note, mixture distributions and aberrant observations not following the model frequently occur in practice. The need for resistant estimators in the nonparametric framework has been acknowledged as early as Hastie and Tibshirani (1990), but the literature in the meantime has remained relatively sparse and overwhelmingly focused on continuous responses with constant variance. For this type of response, resistant penalized spline estimators have been considered by Kalogridis (2021) and Kalogridis and Van Aelst (2021), but unfortunately these ideas do not naturally extend to the setting of GLMs.

A resistant estimation procedure encompassing many generalized linear models was proposed by Azadeh and Salibian-Barrera (2011), who devised a robust backfitting algorithm based on the robust quasi-likelihood estimator of Cantoni and Ronchetti (2001b) along with local polynomial estimation and showed the asymptotic unbiasedness of their estimates under a suitable set of conditions. Bianco et al. (2011) investigated general weighted M-estimators defined and proved their pointwise consistency and asymptotic normality. Croux et al. (2012) combined the approach of Cantoni and Ronchetti (2001b) with the P-spline approach of Eilers and Marx (1996) in order to construct robust estimates of both mean and dispersion, but without theoretical support. For smoothing spline type robust quasi-likelihood estimators, Wong et al. (2014) showed that for smooth score functions and under certain limit conditions the robust estimator inherits the rate of convergence of the non-robust quasi-likelihood estimator. More recently, Aeberhard et al. (2021) proposed obtaining robust estimates by wrapping log-likelihood contributions with a smooth convex function that downweights small log-likelihood contributions, but again with limited theoretical support.

A drawback shared by all aforementioned approaches is the lack of automatic Fisher consistency, that is, these methods do not estimate the target quantities without ad hoc corrections. These corrections may seem unnatural to researchers and practitioners that are well acquainted with classical maximum likelihood estimators. Moreover, the estimators of Azadeh and Salibian-Barrera (2011), Croux et al. (2012) and Wong et al. (2014) depend on monotone score functions, which implies that outliers in both response and predictor spaces can still exert undue influence on the estimates. Croux et al. (2012) proposed using a weight function to limit the influence of high leverage observations in the predictor space, but this greatly complicates theoretical investigations and does not help with gross outliers in the response space. On the other hand, the pointwise nature of the estimators of Bianco et al. (2011) poses a computational challenge and makes uniform asymptotic results much more difficult to obtain.

To overcome these issues, we propose a new family of nonparametric spline estimators for GLMs that is based on the concept of density power divergence between distributions, as developed by Basu et al. (1998). We propose both a full-rank smoothing spline type estimator and a lower-rank estimator based on the principle of penalized splines, which is particularly advantageous for large datasets. The proposed estimators possess a number of notable advantages. First, they are inherently Fisher consistent for all GLMs and thus do not require ad hoc corrections. The proposed class of estimators is adaptive and can combine high efficiency in clean data with robustness against gross outliers. Moreover, these estimators can be efficiently computed through combination of the locally supported B-splines and fast iterative algorithms. For this family of estimators, we establish asymptotic rates of convergence in a Sobolev norm, a rate of uniform convergence and rates of convergence of the derivatives, the latter two of which are novel outside the classical nonparametric likelihood framework.

2 The proposed family of estimators

2.1 Density power divergence

We begin by reviewing the definition of density power divergence, as introduced by Basu et al. (1998) for i.i.d. data and modified by Ghosh and Basu (2013) for independent but not identically distributed data. Consider two densities f and h, which for clarity we take to be Lebesgue densities. Densities with respect to the counting measure are also covered by the subsequent arguments provided essentially that integrals are replaced by sums. The density power divergence $d_\alpha(h, f)$ is defined as

$$d_\alpha(h, f) = \begin{cases} \int_{\mathbb{R}} \left\{ f^{1+\alpha}(x) - \left(1 + \frac{1}{\alpha}\right) h(x) f^\alpha(x) + \frac{1}{\alpha} h^{1+\alpha}(x) \right\} \mathrm{d}x & \alpha > 0 \\ \int_{\mathbb{R}} h(x) \log\{h(x)/f(x)\} \mathrm{d}x & \alpha = 0. \end{cases}$$

It is easy to see that $d_\alpha(h, f) \geq 0$ and that $d_\alpha(h, f) = 0$ if and only if $h = f$ almost everywhere. Moreover, it can be readily verified that $d_\alpha(h, f)$ is continuous for $\alpha \to 0$. In fact, $d_0(h, f)$ is the Kullback–Leibler divergence, which is closely associated with classical maximum likelihood estimators (see, e.g. Claeskens and Hjort 2008), whereas $d_1(h, f)$ is the L_2-error between densities, which has been used by Scott (2001) for robust parametric modelling. Hence, as Basu et al. (1998) note, density power divergence provides a smooth bridge between maximum likelihood estimation and L_2-distance minimization.

In the GLM setting, we assume that the densities of the Y_i share a possibly infinite-dimensional parameter θ_0 determining the form of their conditional means through the relationship $G(\mu_i) = \ell_i(\theta_0)$ where ℓ_i is the evaluation functional for each t_i, i.e. $\ell_i(\theta_0) = \theta_0(t_i)$. Moreover, θ_0 belongs to some metric space Θ, which is known a priori. Henceforth, we shall place emphasis on estimating θ_0, and assume that ϕ_0, the dispersion parameter, is either known or suitably substituted. For popular GLMs such as logistic, Poisson and exponential models, ϕ_0 is indeed known whereas for other GLMs, e.g. with Gaussian responses, ϕ_0 can be estimated without any model fitting with the resistant Rice-type estimators proposed by Ghement et al. (2008) and Boente et al. (2010). Therefore, we set $\phi_0 = 1$ without loss of generality and drop it from the notation from now on. Furthermore, for convenience we shall from now on use θ_i to denote $\ell_i(\theta)$.

To derive estimators from the density power divergence, we fix an $\alpha > 0$ and minimize $n^{-1} \sum_{i=1}^n d_\alpha(f_{\theta_{0,i}}, f_{\theta_i})$ over all $\theta \in \Theta$. For this minimization problem, we may drop terms only involving $f_{\theta_{0,i}}$ as they do not depend on θ. An unbiased estimator for the unknown cross term $\int f_{\theta_i}^\alpha dF_{\theta_{0,i}}$ is given by $f_{\theta_i}^\alpha(Y_i)$. Thus, we may replace each summand by

$$l_\alpha(Y_i, \theta_i) = \int_{\mathbb{R}} f_{\theta_i}^{1+\alpha}(x) \mathrm{d}x - \left(1 + \frac{1}{\alpha}\right) f_{\theta_i}^\alpha(Y_i), \quad (i = 1 \ldots, n). \tag{3}$$

Below we explain how we use this principle to construct a new class of nonparametric GLM estimators with good asymptotic properties.

In the remainder of the paper, we assume that Θ is a reproducing kernel Hilbert space (RKHS) of functions that is generated by a kernel $\mathcal{R} : [0, 1]^2 \to \mathbb{R}$. Let us denote this space by $\mathcal{H}(\mathcal{R})$. A crucial property of $\mathcal{H}(\mathcal{R})$ is that for every $f \in \mathcal{H}(\mathcal{R})$, it holds that $f(t_i) = \langle f, \mathcal{R}(t_i, \cdot)\rangle_{\mathcal{H}(\mathcal{R})}$. A popular measure of robustness is the influence function (IF), which roughly measures the effect of a small proportion of contamination on the estimator (see, e.g. Hampel et al. 2011). Estimators with bounded IF are considered robust, as in that case a small amount of contamination can only have a limited effect on the estimator.

It can be shown that for every $\alpha > 0$, the associated density power divergence estimator possesses a bounded influence function. Indeed, using standard M-estimation theory, it may be verified that the influence function for the density power divergence functional is proportional to

$$\sum_{i=1}^{n} \left[u_{\theta_{0,i}}(y_0) f_{\theta_{0,i}}^{\alpha}(y_0) - \int_{\mathbb{R}} u_{\theta_{0,i}}(y) f_{\theta_{0,i}}^{1+\alpha}(y) \mathrm{d}y \right] \langle \mathcal{R}(t_i, \cdot), \cdot \rangle_{\mathcal{H}(\mathcal{R})} \tag{4}$$

where $u_\theta(y) = (y - b'(\theta))$ is the derivative of the log-density with respect to its canonical parameter and $u_{\theta_{0,i}}(y)$ is the derivative evaluated at $\theta_{0,i}$. An analogous result holds for the discrete case. For $\alpha = 0$, the IF of the maximum likelihood estimator is obtained, which is clearly an unbounded function of y_0. However, for $\alpha > 0$ it may be seen that the IF is bounded, ensuring that density power divergence estimators are robust against small amounts of contamination. Their degree of resistance depends on the magnitude of α, as larger values of α lead to a faster decay of $u_\theta(y_0) f_\theta^\alpha(y_0)$ to zero, ensuring greater robustness.

2.2 Smoothing spline type estimators

Consider now the specific GLM (1) with densities (2) where $\theta_{0,i} = \theta_0(t_i) = g_0(t_i)$ for $i = 1, \ldots, n$, and g_0 needs to be estimated from the data. In this section, we only require that g_0 belongs to the Hilbert–Sobolev space $\mathcal{W}^{m,2}([0,1])$ for some $m \geq 1$, which is defined as

$$\mathcal{W}^{m,2}([0,1]) = \{f : [0,1] \to \mathbb{R}, f \text{ has } m-1 \text{ absolutely continuous derivatives} \\ f^{(1)}, \ldots, f^{(m-1)} \text{ and } \int_0^1 |f^{(m)}(t)|^2 \mathrm{d}t < \infty\}.$$

The space $\mathcal{W}^{m,2}([0,1])$ is well suited for nonparametric regression problems, as it forms a RKHS so that for each $x \in [0,1]$ the evaluation functionals $\mathcal{W}^{m,2}([0,1]) \to \mathbb{R} : f \mapsto f(x)$ are continuous, see, e.g. Wahba (1990).

As a compromise between goodness of fit and complexity, we propose to estimate g_0 by the function $\widehat{g}_n$ solving

$$\inf_{g \in \mathcal{W}^{m,2}([0,1])} \left[\frac{1}{n} \sum_{i=1}^{n} l_{\widehat{\alpha}_n}(Y_i, g(t_i)) + \lambda \int_{[0,1]} |g^{(m)}(t)|^2 \mathrm{d}t \right], \tag{5}$$

for some $\lambda > 0$, that acts as the tuning parameter. Here, we also allow for a random tuning parameter $\widehat{\alpha}_n$. This random tuning parameter may depend on the data itself leading to an adaptive estimator whose robustness and efficiency automatically adjust to the data. In particular, for $\widehat{\alpha}_n$ close to zero the objective function approaches the penalized likelihood considered, for example, by Cox and O'Sullivan (1990), Mammen and van de Geer (1997) and Kauermann et al. (2009). As discussed previously, these estimators are efficient but not robust. On the other hand, for large $\widehat{\alpha}_n$, estimators minimizing (5) are robust but not efficient. In practice, we aim to balance robustness and efficiency and select an $\widehat{\alpha}_n$ in (0, 1], although higher values can also be considered. Section 4 outlines a possible strategy in this respect.

For bounded densities that are continuous with respect to their parameter, the objective function is bounded from below and continuous in $\mathcal{W}^{m,2}([0, 1])$. Reasoning along the same lines as in the proof of Theorem 1 of Kalogridis (2021) reveals that for $n \geq m$, this minimization problem is well defined and there exists at least one minimizer in $\mathcal{W}^{m,2}([0, 1])$. Arguing now in a standard way (see, e.g. Eubank 1999) shows that this minimizer must be an easily computable n-dimensional natural spline with knots at the unique $t_1, \ldots, t_n$. As we discuss in Sect. 4 though, unrestricted B-splines may also be used in the computation of the estimator.

Even for Gaussian responses, the smoothing spline type estimator in (5) has not been previously considered. In this case, the form of the loss function is rather simple. Indeed, for Gaussian Y_i the first term in (5) is constant as a function of $g \in \mathcal{W}^{m,2}([0, 1])$. Hence, apart from constants the objective function becomes

$$\frac{1}{n}\sum_{i=1}^{n} \rho_{\widehat{\alpha}_n}(Y_i - g(t_i)) + \lambda \int_{[0,1]} |g^{(m)}(t)|^2 \mathrm{d}t,$$

with $\rho_\alpha(x) = -e^{-\alpha x^2/2}$. This exponential loss function has attractive properties for robust estimation because it is a bounded loss function which is infinitely differentiable with bounded derivatives of all orders. In the parametric setting, the exponential squared loss function has been used, e.g. by Wang et al. (2013). In the nonparametric setting considered herein, the penalized exponential squared loss gives rise to a novel estimator that may be viewed as a more robust alternative to the Huber and least absolute deviations smoothing spline estimators studied in van de Geer (2000); Kalogridis (2021). See Sect. 5 for interesting comparisons.

A noteworthy property of the penalty functional in (5) is the shrinkage of the estimator towards a polynomial of order m. To see this, assume that $\widehat{g}_n$ lies in the null space of the penalty so that $\|\widehat{g}_n^{(m)}\| = 0$. A Taylor expansion with integral remainder term shows that

$$\widehat{g}_n(t) = P_m(t) + \int_0^1 \frac{\widehat{g}_n^{(m)}(x)}{(m-1)!}(t-x)_+^{m-1}\mathrm{d}x,$$

where $P_m(t)$ is the Taylor polynomial of order m. The Schwarz inequality shows that the integral remainder vanishes for all $t \in [0, 1]$, whence $\sup_{t\in[0,1]} |\widehat{g}_n(t) - P_m(t)| = 0$. This implies that letting $\lambda \to \infty$ will cause the estimator to become a polynomial of order m, as the dominance of the penalty term in (5) forces the estimator to lie in its null space.

A crucial property underlying the construction of all our GLM estimators is their inherent Fisher consistency (Hampel et al. 2011, p. 83). In particular, our previous discussion shows that, for each $\alpha > 0, \theta_{0,i}$ minimizes $\mathbb{E}\{l_\alpha(Y_i, \theta_i)\}$ for each $i = 1, \ldots, n$. Hence, our estimation method is Fisher consistent for every model distribution. To the best of our knowledge, this is the first robust nonparametric estimator that enjoys this property without corrections, although other divergence based-estimators may also enjoy this property. Since Fisher consistency corrections are model dependent, the inherent Fisher consistency yields an important practical bonus.

2.3 Penalized spline type alternatives

A drawback of smoothing spline type estimators is their dimension, which grows linearly with the sample size. This implies that for large n smoothing spline estimators can be computationally cumbersome. Moreover, as noted by Wood (2017, p.199), in practice the value of λ is almost always high enough such that the effective degrees of freedom of the resulting spline are much smaller than n. Penalized spline estimators offer a compromise between the complexity of smoothing splines and the simplicity of (unpenalized) regression splines (O'Sullivan 1986; Eilers and Marx 1996). We now discuss penalized spline estimators for our setting as a simpler alternative to the smoothing spline estimator discussed above.

Fix a value $K \in \mathbb{N}_+$ and define the interior knots $0 = x_0 < x_1, \dots, x_K < x_{K+1} = 1$, which do not have to be design points. For a fixed $p \in \mathbb{N}_+$, let $S_K^p([0,1])$ denote the set of spline functions on $[0, 1]$ of order p with knots at the x_i. For $p = 1$, $S_K^1([0,1])$ is the set of step functions with jumps at the knots while for $p \geq 2$,

$$S_K^p([0,1]) = \{s \in \mathcal{C}^{p-2}([0,1]) : s(x) \text{ is a polynomial of degree } (p-1) \text{ on each } [x_i, x_{i+1}]\}.$$

Thus, p controls the smoothness of the functions in S_K^p while the number of interior knots K represents the degree of flexibility of spline functions in $S_K^p([0,1])$, see, e.g. Ruppert et al. (2003); Wood (2017). It is easy to see that $S_K^p([0,1])$ is a $(K+p)$-dimensional subspace of $\mathcal{C}^{p-2}([0,1])$ and B-spline functions yield a stable basis for $S_K^p([0,1])$ with good numerical properties (de Boor 2001).

For any $m \in \mathbb{N}_+$ satisfying $m < p$, we define the penalized spline type estimator $\widehat{g}_n$ as the solution of the optimization problem

$$\min_{g \in S_K^p([0,1])} \left[\frac{1}{n} \sum_{i=1}^{n} l_{\widehat{\alpha}_n}(Y_i, g(t_i)) + \lambda \int_{[0,1]} |g^{(m)}(t)|^2 \mathrm{d}t \right], \tag{6}$$

with $\lambda \geq 0$. Hence, we have replaced $\mathcal{W}^{m,2}([0,1])$ in (5) by a $(K+p)$- dimensional spline subspace. For $K \ll n$, this yields sizeable computational gains in relation to the smoothing spline estimator. Moreover, it turns out that penalized spline estimators do not sacrifice much in terms of accuracy if K is large enough, but still smaller than n. See Claeskens et al. (2009); Xiao (2019) for interesting comparisons in classical nonparametric regression models and Sect. 3 for a comparison in the present context.

Penalized spline estimators retain a number of important mathematical properties of their full-rank smoothing spline counterparts. In particular, for $\lambda > 0$ and $p = 2m$ it can be shown that the penalized spline estimator is a natural spline of order $2m$. Moreover, the null space of the penalty consists exactly of polynomials of order $\leq m$.

In the frequently used setting of equidistant interior knots, the latter property is also retained if one replaces the derivative penalty with the simpler difference (P-spline) penalty $\sum_{j=m+1}^{K+p} |\Delta^m \beta_j|^2$ proposed by Eilers and Marx (1996). Here, Δ^m is the mth backward difference operator and β_j, $j = 1, \ldots, K + p$ are the coefficients of the B-spline functions. In this case, these two penalties are scaled versions of one another with the scaling factor depending on K, p and m (see, e.g. Proposition 1 of Kalogridis and Van Aelst 2021). Thus, P-spline estimators are also covered by the asymptotic results of the following section.

3 Asymptotic behaviour of the estimators

3.1 Smoothing spline type estimators

As noted before, an essential characteristic of $\mathcal{W}^{m,2}([0, 1])$ is that it is a RKHS. The reproducing kernel depends on the inner product that $\mathcal{W}^{m,2}([0, 1])$ is endowed with. We shall make use of the inner product

$$\langle f, g\rangle_{m,\lambda} = \langle f, g\rangle + \lambda\langle f^{(m)}, g^{(m)}\rangle, \quad f, g \in \mathcal{W}^{m,2}([0, 1])$$

where $\langle\cdot, \cdot\rangle$ denotes the standard inner product on $\mathcal{L}^2([0, 1])$. It is interesting to observe that $\langle\cdot, \cdot\rangle_{m,\lambda}$ is well-defined and depends on the smoothing parameter λ, which typically varies with n. Eggermont and LaRiccia (2009, Chapter 13) show that there exists a finite positive constant c_0 such that for all $f \in \mathcal{W}^{m,2}([0, 1])$ and $\lambda \in (0, 1]$ we have

$$\sup_{t\in[0,1]} |f(t)| \leq c_0\lambda^{-1/(4m)} \|f\|_{m,\lambda}, \tag{7}$$

with $\|f\|_{m,\lambda} = \langle f, f\rangle_{m,\lambda}^{1/2}$. Hence, for any $\lambda \in (0, 1]$, $\mathcal{W}^{m,2}([0, 1])$ is indeed a RKHS under $\langle\cdot, \cdot\rangle_{m,\lambda}$. The condition $\lambda \leq 1$ is not restrictive, as for our main result below we assume that $\lambda \to 0$ as $n \to \infty$ and our results are asymptotic in nature.

The assumptions needed for our theoretical development are as follows.

(A1) The support $\mathcal{Y} := \overline{\{y : f_\theta(y) > 0\}}$ does not depend on $\theta \in \mathbb{R}$.

(A2) There exists an $\alpha_0 > 0$ such that $\widehat{\alpha}_n \xrightarrow{P} \alpha_0$.

(A3) The densities $\{f_\theta(y), \theta \in \mathbb{R}\}$ are uniformly bounded, twice differentiable as a function of their canonical parameter θ in a neighbourhood of the true parameter θ_0 and there exist $\delta > 0$ and $M < \infty$ such that

$$\sup_{t\in[0,1]} \sup_{\substack{|\alpha-\alpha_0|<\delta \\ |u|<\delta}} \sup_{y\in\mathcal{Y}} \left| \frac{\partial f_\theta^\alpha(y)}{\partial\theta}\Big|_{\theta=\theta_0(t)+u} \log\left(f_{\theta_0(t)+u}(y)\right)\right| \leq M.$$

(A4) In the case of densities w.r.t. Lebesgue measure, the families of functions $\{m_t(u,\alpha), t \in [0,1]\}$ and $\{n_t(u,\alpha,y), t \in [0,1]\}$ defined by

$$m_t(u,\alpha) = \int_{\mathcal{Y}} \frac{\partial^2 f_\theta^{1+\alpha}(y)}{\partial\theta^2}\bigg|_{\theta=\theta_0(t)+u} \mathrm{d}y, \quad n_t(u,\alpha,y) = \frac{\partial^2 f_\theta^{\alpha}(y)}{\partial\theta^2}\bigg|_{\theta=\theta_0(t)+u},$$

are equicontinuous at $u = 0$, for every $\alpha \in (\alpha_0 - \delta, \alpha_0 + \delta)$ with δ as in (A3) and $y \in \mathcal{Y}$. Moreover, there exists an $M' > 0$ such that

$$\sup_{t\in[0,1]} \sup_{\substack{|\alpha-\alpha_0|<\delta \\ |u|<\delta}} \left[|m_t(u,\alpha)| + \sup_{y\in\mathcal{Y}} |n_t(u,\alpha,y)| \right] \leq M'.$$

For densities w.r.t. counting measure, the sum over $y \in \mathcal{Y}$ replaces the integral in the definition of $m_t(u,\alpha)$.

(A5) For $\delta > 0$, as given in (A3), there exist $0 < c_0 \leq C_0 < \infty$ such that

$$c_0 \leq \inf_n \min_{i\leq n} \mathbb{E}\{f_{\theta_{0,i}}^{\alpha}(Y_i)|u_{\theta_{0,i}}(Y_i)|^2\} \leq \sup_n \max_{i\leq n} \mathbb{E}\{f_{\theta_{0,i}}^{\alpha}(Y_i)|u_{\theta_{0,i}}(Y_i)|^2\} \leq C_0,$$

for all $\alpha \in (\alpha_0 - \delta, \alpha_0 + \delta)$, with $u_{\theta_{0,i}}(y) = \partial \log(f_\theta(y))/\partial\theta$ the derivative of the log-density evaluated at $\theta_{0,i}$.

(A6) The family of design points t_i is asymptotically quasi-uniform in the sense of Eggermont and LaRiccia (2009), that is, there exists an $n_0 \in \mathbb{N}$ such that,

$$\sup_{f\in\mathcal{W}^{1,1}([0,1])} \frac{\left| n^{-1}\sum_{i=1}^n f(t_i) - \int_0^1 f(t)\mathrm{d}t \right|}{\int_0^1 |f'(t)|\mathrm{d}t} = O(n^{-1}),$$

for all $n \geq n_0$.

Condition (A1) is standard in the theory of exponential families and may be viewed as an identifiability condition. It is worth noting that imposing the convergence in probability of nuisance (estimated) parameters as in (A2) is a common way of dealing with them theoretically, see, e.g. van der Vaart (1998, theorems 5.31 and 5.55). Furthermore, since assumption (A2) requires no rate of convergence whatsoever, the methodology employed herein can be used to weaken such assumptions in other works, e.g. in Kalogridis (2021, Theorem 3) whose assumptions require a specific rate of convergence of the preliminary scale estimator in the context of robust nonparametric regression with continuous responses.

Assumptions (A3)–(A4) are regularity conditions, which are largely reminiscent of classical maximum likelihood conditions (cf. Theorem 5.41 in van der Vaart 1998). These conditions essentially impose some regularity of the first and second derivatives in a neighbourhood of the true parameter. Since for any $\alpha > 0$,

$$\frac{\partial^2 f_\theta^{\alpha}(y)}{\partial\theta^2} = \alpha^2|y - b'(\theta)|^2 f_\theta^{\alpha}(y) - \alpha b''(\theta) f_\theta^{\alpha}(y),$$

conditions (A3) and (A4) are satisfied for a wide variety of GLMs due to the rapid decay of $f_{\theta}^{\alpha}(y)$ for large values of $|y|$. In particular, they are satisfied for the Gaussian, logistic and Poisson models. Similarly condition (A5) is an extension of the assumptions underpinning classical maximum likelihood estimation. For $\alpha = 0$, these moment conditions entail that the Fisher information is strictly positive and finite. The following examples demonstrate that condition (A5) holds for popular GLMs.

Example 1 (Gaussian responses) Clearly, $u_{\theta_{0,i}} = y - b'(\theta_{0,i}) = y - \theta_{0,i}$ in this case and

$$\int_{\mathbb{R}} f_{\theta_{0,i}}^{1+\alpha}(y)|u_{\theta_{0,i}}(y)|^2 \mathrm{d}y = \frac{1}{\{(2\pi)^{\alpha}(1+\alpha)\}^{1/2}},$$

so that (A5) is satisfied without any additional conditions.

Example 2 (Binary responses) The canonical parameter is $\theta_i = \log(p_i/(1-p_i))$ with p_i denoting the probability of success for the ith trial. Thus,

$$\sum_{y \in \{0,1\}} f_{\theta_{0,i}}^{1+\alpha}(y)|u_{\theta_{0,i}}(y)|^2 = (1-p_{0,i})^{1+\alpha} p_{0,i}^2 + p_{0,i}^{1+\alpha}(1-p_{0,i})^2,$$

where $p_{0,i} = 1/(1+e^{-\theta_{0,i}})$. Thus, (A5) is satisfied whenever $p_{0,i}$ is bounded away from zero and one, precisely as required by Cox and O'Sullivan (1990) and Mammen and van de Geer (1997) for classical nonparametric logistic regression.

Example 3 (Exponential responses) Let $f_{\theta_{0,i}}(y) = \exp[y\theta_{0,i} - \log(-\theta_{0,i})]\mathcal{I}_{(0,\infty)}(y)$ with $\theta_{0,i} = -\lambda_{0,i}$. A lengthy calculation shows that

$$\int_{\mathbb{R}} f_{\theta_{0,i}}^{1+\alpha}(y)|u_{\theta_{0,i}}(y)|^2 \mathrm{d}y = \frac{1+\alpha^2}{(1+\alpha)^3} \frac{1}{\lambda_{0,i}^{2-\alpha}}.$$

Hence, (A5) is satisfied provided that $\lambda_{0,i}$ stays away from zero, since by compactness and continuity, $\lambda_0(t)$ is always bounded.

Example 4 (Poisson responses) Here, for positive rate parameters $\lambda_{0,i}$, we have $\theta_{0,i} = \log(\lambda_{0,i})$, $i = 1, \ldots, n$. It can be shown that

$$\lambda_{0,i}^2 \exp[-\lambda_{0,i}(1+\alpha)] \leq \sum_{y=0}^{\infty} f_{\theta_{0,i}}^{1+\alpha}(y)|u_{\theta_{0,i}}(y)|^2 \leq \lambda_{0,i},$$

so that (A5) is satisfied provided that $\lambda_{0,i}$ stays away from zero.

It is easy to see that assumptions (A1)–(A5) also cover density power divergence estimators based on a fixed $\alpha_0 > 0$, for example, L_2-distance estimators with $\alpha_0 = 1$, as in this case condition (A2) is trivially satisfied and conditions (A3)–(A5) only need to hold for that particular α_0. Finally, condition (A6) ensures that the design points are well spread throughout the interval of interest. Call Q_n the distribution function of $t_1, \ldots, t_n$. Then, for each $f \in \mathcal{W}^{1,1}([0,1])$, an integration by parts argument reveals that

$$\left| n^{-1} \sum_{i=1}^{n} f(t_i) - \int_0^1 f(t) \mathrm{d}t \right| \leq \sup_{t \in [0,1]} |Q_n(t) - t| \int_0^1 |f'(t)| \mathrm{d}t.$$

Consequently, (A6) is satisfied provided that Q_n approximates well the uniform distribution function, for example if $t_i = i/(n+1)$ or $t_i = 2i/(2n+1)$.

Theorem 1 contains the main result for smoothing type spline estimators defined in (5).

Theorem 1 *If assumptions (A1)–(A6) hold and* $\lambda \to 0$ *in such a way that* $n\lambda^{1/m} \to \infty$ *and* $n\lambda^{1/(2m)} \exp[-\lambda^{-1/m}] \to 0$ *as* $n \to \infty$*. Then,*

$$\lim_{D \to \infty} \liminf_{n \to \infty} \Pr[\text{there exists a sequence of local minimizers } \widehat{g}_n \text{ of (5) satisfying}$$
$$\|\widehat{g}_n - g_0\|_{m,\lambda}^2 \leq D(n^{-1}\lambda^{-1/(2m)} + \lambda)] = 1.$$

The result in Theorem 1 is local in nature as the objective function (6) is not convex, unless $\widehat{\alpha}_n = 0$. Similar considerations exist in Basu et al. (1998), Fan and Li (2001) and Wang et al. (2013). The limit requirements of the theorem are met, e.g. whenever $\lambda \asymp n^{-2m/(2m+1)}$, in which case we are led to the optimal rate $\|\widehat{g}_n - g_0\|_{m,\lambda}^2 = O_P(n^{-2m/(2m+1)})$. This rate is similar to the rate obtained in nonparametric regression for continuous responses (Kalogridis 2021). A faster $n^{-4m/(4m+1)}$-rate can be obtained whenever g_0 and its derivatives fulfil certain boundary conditions, see Eggermont and LaRiccia (2009); Kalogridis (2021). Since we assume that, for all large n, $\widehat{\alpha}_n > 0$ with high probability, the theorem does not cover penalized likelihood estimators, but these estimators may be covered with similar arguments as in Mammen and van de Geer (1997); van de Geer (2000) and the same rates of convergence would be obtained.

As for all $\lambda > 0$, $\|\cdot\| < \|\cdot\|_{m,\lambda}$ the same rate applies for the more commonly used $\mathcal{L}^2([0,1])$-norm. However, the fact that our results are stated in terms of the stronger $\|\cdot\|_{m,\lambda}$ leads to two notable consequences. First, for $\lambda \asymp n^{-2m/(2m+1)}$ the bound in (7) immediately yields

$$\sup_{t \in [0,1]} |\widehat{g}_n(t) - g_0(t)| = O_P(n^{(1/2-m)/(2m+1)}),$$

which implies that convergence can be made uniform. Although this uniform rate is slower than the $\log^{1/2}(n)n^{-m/(2m+1)}$-rate in Eggermont and LaRiccia (2009, Chapter 21) for the classical smoothing spline in case of continuous data with constant variance, this rate is guaranteed for a much broader setting encompassing many response distributions. Secondly, by using Sobolev embeddings we can also describe the rate of convergence of the derivatives of $\widehat{g}_n$ in terms of the $\mathcal{L}^2([0,1])$-metric. These are given in Corollary 1. To the best of our knowledge, these are the first results on uniform convergence and convergence of derivatives in among all robust estimation methods for nonparametric generalized linear models.

Corollary 1 *Assume the conditions of Theorem 1 hold. Then, for any* $\lambda \asymp n^{-2m/(2m+1)}$, *the sequence of minimizers in Theorem 1 satisfies*

$$\left\| \widehat{g}_n^{(j)} - g_0^{(j)} \right\|^2 = O_P(n^{-2(m-j)/(2m+1)}), \quad (j = 1, \ldots, m).$$

3.2 Penalized spline type estimators

The assumptions for the penalized spline type estimators are for the most part identical to those for smoothing spline type estimators. In addition to (A1)–(A5), we require the following assumptions.

- (B6) The number of knots $K = K_n \to \infty$ and there exists a $\delta' > 0$ such that $n^{\delta'-1}K^2 \to 0$ as $n \to \infty$.
- (B7) Let $h_i = x_i - x_{i-1}$ and $h = \max_i h_i$. Then, $\max_i |h_{i+1} - h_i| = o(K^{-1})$ and there exists a finite $M > 0$ such that $(h/\min_i h_i) \leq M$.
- (B8) Let Q_n denote the empirical distribution of the design points $t_i, i = 1, \ldots, n$. Then, there exists a distribution function Q with corresponding Lebesgue density bounded away from zero and infinity such that $\sup_{t\in[0,1]} |Q_n(t) - Q(t)| = o(K^{-1})$.

Assumptions (B6)–(B8) have been extensively used in the treatment of penalized spline estimators, (see, e.g. Claeskens et al. 2009; Kauermann et al. 2009; Xiao 2019). Assumption (B6) imposes a weak condition on the rate of growth of the interior knots which is not restrictive in practice, as it is in line with the primary motivation behind penalized spline type estimators. Assumption (B7) concerns the placement of the knots and is met if the knots are equidistant, for example. Finally, assumption (B8) is the lower-rank equivalent of assumption (A6) and holds in similar settings.

Recall that $S_K^p([0, 1])$ is a finite-dimensional space and, for any $f : [0, 1] \to \mathbb{R}$ which is continuous or a step function, $\|f\| = 0$ implies that $f = 0$. Consequently, endowing $S_K^p([0, 1])$ with $\langle \cdot, \cdot \rangle_{m,\lambda}$ makes it a Hilbert space. Proposition 1 shows that $S_K^p([0, 1])$ is an RKHS allowing us to draw direct parallels between $\mathcal{W}^{m,2}([0, 1])$ and its dense subspace $S_K^p([0, 1])$.

Proposition 1 *If assumptions (B6)–(B7) hold and* $p > m \geq 1$, *then there exists a positive constant* c_0 *such that for every* $f \in S_K^p([0, 1])$, $K \geq 1$ *and* $\lambda \in [0, 1]$ *it holds that*

$$\sup_{t\in[0,1]} |f(t)| \leq c_0 \min\{K^{1/2}, \lambda^{-1/(4m)}\} \|f\|_{m,\lambda},$$

where for $\lambda = 0$ *we define* $\lambda^{-1/(4m)} = \infty$.

Proposition 1 implies the existence of a symmetric function $\mathcal{R}_{m,K,\lambda} : [0, 1] \times [0, 1] \to \mathbb{R}$ depending on K, λ, m and p such that, for every $y \in [0, 1]$ the map $x \mapsto \mathcal{R}_{m,K,\lambda}(x, y) \in S_K^p([0, 1])$ and for every $f \in S_K^p([0, 1])$, $f(t) = \langle \mathcal{R}_{m,K,\lambda}(t, \cdot), f \rangle_{m,\lambda}$. Hence, $\mathcal{R}_{m,K,\lambda}$ is the reproducing kernel which can be derived explicitly in this setting. Let $\mathbf{H}_p$ denote the $(K + p) \times (K + p)$ matrix consisting of the inner products $\langle B_i, B_j\rangle$ with $B_1, \ldots, B_{K+p}$, the B-spline functions of order p, and let $\mathbf{P}_m$ denote the penalty matrix with elements $\langle B_i^{(m)}, B_j^{(m)}\rangle$. Set $\mathbf{G}_\lambda = \mathbf{H}_p + \lambda\mathbf{P}_m$, then

$$\mathcal{R}_{m,K,\lambda}(x, y) = \mathbf{B}^\top(x)\mathbf{G}_\lambda^{-1}\mathbf{B}(y).$$

Since, for $f, g \in S_K^p([0, 1])$ we have $\langle f, g\rangle_{m,\lambda} = \mathbf{f}^\top\mathbf{G}_\lambda\mathbf{g}$, with $\mathbf{f}, \mathbf{g} \in \mathbb{R}^{K+p}$ the vectors of scalar coefficients, it is easy to see that $\mathcal{R}_{m,K,\lambda}$ satisfies the required properties.

Since $S_K^p([0, 1]) \subset \mathcal{W}^{m,2}([0, 1])$ for $p > m$ it follows from (7) that for all $f \in S_K^p([0, 1])$ it holds that $\sup_{t\in[0,1]} |f(t)| \leq c_0\lambda^{-1/(4m)}\|f\|_{m,\lambda}$. However, the bound in Proposition 1 is tighter if $K = o(\lambda^{-1/(2m)})$, that is, if the rate of growth of K is outpaced by the rate of decay of $\lambda^{1/(2m)}$. This suggests that the rate of growth of K and the rate of decay of λ jointly determine the asymptotic properties of penalized spline estimators. This relation is formalized in Theorem 2.

Theorem 2 *Suppose that assumptions (A1)–(A5) and (B6)–(B8) hold and assume that* $\lambda \to 0$ *in a such way that* $K \min\{\lambda^2 K^{2m}, \lambda\} \to 0$ *and* $n \max\{K^{-1}, \lambda^{1/(2m)}\} \exp[-\min\{K^2, \lambda^{-1/m}\}] \to 0$ *as* $n \to \infty$. *Then, if* $g_0 \in \mathcal{C}^j([0, 1])$ *with* $m \leq j \leq p$,

$$\lim_{D\to\infty} \liminf_{n\to\infty} \Pr[\textit{there exists a sequence of local minimizers } \widehat{g}_n \textit{ of (6) satisfying}$$
$$\|\widehat{g}_n - g_0\|^2 \leq D(n^{-1}\min\{K, \lambda^{-1/(2m)}\} + \min\{\lambda^2 K^{2m}, \lambda\} + K^{-2j})] = 1.$$

Theorem 2 presents the $\mathcal{L}^2([0, 1])$-error as a decomposition of three terms accounting for the variance, the modelling bias and the approximation bias of the estimator, respectively. It is interesting to observe that, apart from the term K^{-2j} stemming from the approximation of a generic $\mathcal{C}^j([0, 1])$-function with a spline, the error simultaneously depends on K and λ, highlighting the intricate interplay between knots and penalties in the asymptotics of penalized spline estimators.

For $K < \lambda^{-1/(2m)}$, Theorem 2 leads to the regression spline asymptotics established by Claeskens et al. (2009); Xiao (2019) for Gaussian responses (see also Kim and Gu 2004), whereas for $g_0 \in \mathcal{C}^m([0, 1])$ and $K \geq \lambda^{-1/(2m)}$ we obtain

$$\|\widehat{g}_n - g_0\|^2 = O_P(n^{-1}\lambda^{-1/(2m)}) + O_P(\lambda) + O_P(K^{-2m}),$$

which, apart from the approximation error K^{-2m}, corresponds to the $\mathcal{L}^2([0, 1])$ error decomposition of smoothing spline estimators in Theorem 1. This fact has important practical implications, as it allows for an effective dimension reduction without any theoretical side effects. Indeed, taking $\lambda \asymp n^{-2m/(2m+1)}$ and $K \asymp n^\gamma$ for any $\gamma \geq 1/(2m + 1)$ leads to $\|\widehat{g}_n - g_0\|^2 = O_P(n^{-2m/(2m+1)})$, which is the same rate of convergence as for smoothing spline estimators. Moreover, with this choice of tuning parameters, the convergence rates of the derivatives given in Corollary 1 carry over to the present lower-rank estimators, as summarized in Corollary 2.

Corollary 2 *Assume the conditions of Theorem 2 hold and* $g_0 \in \mathcal{C}^m([0, 1])$. *Then, for any* $\lambda \asymp n^{-2m/(2m+1)}$ *and* $K \asymp n^\gamma$ *with* $\gamma \geq 1/(2m + 1)$,

$$\left\|\widehat{g}_n^{(j)} - g_0^{(j)}\right\|^2 = O_P(n^{-2(m-j)/(2m+1)}), \quad (j = 1, \ldots, m).$$

While there have been attempts in the past to cast penalized spline estimators in an RKHS framework (e.g. in Pearce and Wand 2006), this was from a computational perspective. To the best of our knowledge, our theoretical treatment of penalized spline estimators based on the theory of RKHS is the first of its kind in both the classical and the robustness literature and may be of independent mathematical interest. The interested reader is referred to the accompanying supplementary material for the technical details.

4 Practical implementation

4.1 Computational algorithm

The practical implementation of the smoothing and penalized spline estimators based on density power divergence requires a computational algorithm as well as specification of their parameters, namely the tuning parameter α, the penalty parameter λ and K, the dimension of the spline basis, in case of penalized splines. Practical experience with penalized spline estimators has shown that the dimension of the spline basis is less important than the penalty parameter, provided that K is taken large enough (Ruppert et al. 2003; Wood 2017). Hence, little is lost by selecting K in a semi-automatic manner. We now discuss a unified computational algorithm for the smoothing and penalized type estimators described in Sect. 3 and briefly discuss the selection of K.

By using the B-spline basis, the computation of the proposed estimators can be unified. Recall from our discussion in Sect. 3 that a solution to (5) is a natural spline of order $2m$. Assume for simplicity that all the t_i are distinct and that $\min_i t_i > 0$ and $\max_i t_i < 1$. Then, the natural spline has n interior knots and we may represent the candidate minimizer $g \in S_K^p([0,1])$ as $g = \sum_{k=1}^{n+2m} g_k B_k$ where the B_k are the B-spline basis functions of order $2m$ supported by the knots at the interior points t_i and the g_k are scalar coefficients (de Boor 2001,p. 102). The penalty term now imposes the boundary conditions. The reasoning is as follows: if that were not the case, it would always be possible to find a $2m$th order natural spline $\widehat{h}_n(t)$, written in terms of $B_1, \ldots, B_{K+p}$, that interpolates $\widehat{g}_n(t_i)$, $i = 1, \ldots, n$ leaving the first term in (5) unchanged, but due to it being a polynomial of order m outside $[\min_i t_i, \max_i t_i] \subset [0,1]$ we would always have $\|\widehat{h}_n^{(m)}\| < \|\widehat{g}_n^{(m)}\|$.

The above discussion shows that to compute either the smoothing spline type or the penalized spline type estimators, it suffices to minimize

$$L_n(g) = \sum_{i=1}^{n} l_{\widehat{\alpha}_n}(Y_i, g(t_i)) + \lambda \int_{[0,1]} |g^{(m)}(t)|^2 \mathrm{d}t, \tag{8}$$

over $g \in S_K^p([0,1])$, where in the smoothing spline case the knots satisfy $x_i = t_i$, $i = 1, \ldots, n$ and the order $p = 2m$ while in the penalized spline case $0 < x_1 < \ldots < x_K < 1$ and $p > m$. By default we set $p = 2m$ in our implementation of the algorithm and use all interior knots, i.e. $x_i = t_i$ for $i \leq n$ if $n \leq 50$. For $n > 50$, we adopt a thinning strategy motivated by the theoretical results of the previous section as well as the strategy employed by the popular smooth.spline function in the R language (see Hastie et al. 2009, p. 189). In particular, for $n > 50$ we only employ $K \simeq n^{1/(2m+1)}$ interior knots, which we spread in the [0, 1]-interval in an equidistant manner. While this leads to large computational gains, Theorem 2 assures that no accuracy is sacrificed. For example, with $m = 2$ and $n = 5000$ our strategy amounts to using only 83 knots; a dramatic dimension reduction.

We solve (8) with a Newton–Raphson procedure, which we initiate from the robust estimates of Kalogridis and Van Aelst (2021) and Croux et al. (2012). The updating steps of the algorithm can be reduced to penalized iteratively reweighted least-squares updates, in the manner outlined by Green and Silverman (1994, Chapter 5.2.3). The weights in our case are given by

$$\begin{aligned} w_i &= (1+\widehat{\alpha}_n)^2 \mathbb{E}_{g_i}\{|Y_i - b'(g_i)|^2 f_{g_i}^{\widehat{\alpha}_n}(Y_i)\} - (1+\alpha) b''(g_i) \mathbb{E}_{g_i}\{f_{g_i}^{\widehat{\alpha}_n}(Y_i)\} \\ &\quad + (1+\widehat{\alpha}_n) b''(g_i) f_{g_i}^{\widehat{\alpha}_n}(Y_i) - \widehat{\alpha}_n (1+\widehat{\alpha}_n)|Y_i - b'(g_i)|^2 f_{g_i}^{\widehat{\alpha}_n}(Y_i), \quad (i = 1, \ldots, n), \end{aligned}$$

where $g_i = \mathbf{B}^\top(t_i)\boldsymbol{g}$ and the vector of "working" data $\mathbf{z} \in \mathbb{R}^n$ consists of

$$z_i = g_i - (1+\widehat{\alpha}_n)\frac{\mathbb{E}_{g_i}\{(Y_i - b'(g_i)) f_{g_i}^{\widehat{\alpha}_n}(Y_i)\} - \{Y_i - b'(g_i)\} f_{g_i}^{\widehat{\alpha}_n}(Y_i)}{w_i}.$$

Alternatively, taking expected values we may replace the weights with $(1 + \widehat{\alpha}_n)\mathbb{E}_{g_i}\{|Y_i - b'(g_i)|^2 f_{g_i}^{\widehat{\alpha}_n}(Y_i)\}$ thereby obtaining a variant of the Fisher scoring algorithm. For $\widehat{\alpha}_n = 0$, these formulae reduce to those for penalized likelihood estimators and canonical links, cf. Green and Silverman (1994).

4.2 Selection of the tuning and penalty parameters

The tuning parameter α determines the trade-off between robustness and efficiency of the estimators. Selecting α independent of the data could lead to lack of resistance towards atypical observations or an undesirable loss of efficiency. To determine α in a data-driven way we modify the strategy of Ghosh and Basu (2015) and rely on a suitable approximation of the integrated mean integrated squared error (MISE) of $\widehat{g}_n$, i.e. $\mathbb{E}\{\|\widehat{g}_n - g_0\|^2\}$. To show the dependence of the estimator on α, we now denote it by $\widehat{g}_{n,\alpha}$. For each $\alpha \geq 0$, its MISE can be decomposed as

$$\mathrm{MISE}(\alpha) = \left\|\mathbb{E}\left\{\widehat{g}_{n,\alpha}\right\} - g_0\right\|^2 + \mathbb{E}\left\{\left\|\widehat{g}_{n,\alpha} - \mathbb{E}\left\{\widehat{g}_{n,\alpha}\right\}\right\|^2\right\}, \tag{9}$$

where the first term on the RHS represents the integrated squared bias, while the second term is the integrated variance of the estimator. Neither of these terms can be computed explicitly since $\mathbb{E}\{\widehat{g}_{n,\alpha}\}$ and g_0 are both unknown. Therefore, we seek approximations. Following Warwick and Jones (2005) and Ghosh and Basu (2015), we replace $\mathbb{E}\{\widehat{g}_{n,\alpha}\}$ in the bias term by $\widehat{g}_{n,\alpha}$ and use a "pilot" estimator instead of the unknown g_0. To limit the influence of aberrant observations on the selection of α, we propose to replace g_0 by the robust estimate $\widehat{g}_{n,1}$ which minimizes the L_2-distance between the densities, as described in Sect. 2.

To approximate the variance term, observe that $\mathbb{E}\{\|\widehat{g}_{n,\alpha}-\mathbb{E}\{\widehat{g}_{n,\alpha}\}\|^2\} = \operatorname{Tr}\{\mathbf{H}_p \operatorname{Cov}\{\widehat{\mathbf{g}}_{n,\alpha}\}\}$. Using the notation

$$l'_\alpha(Y_i, g(t_i)) = \frac{\partial l_\alpha(Y_i, x)}{\partial x}\Big|_{x=g(t_i)}, \quad (i = 1, \ldots, n).$$

for the first derivatives and analogously for the second derivatives $l''_\alpha(Y_i, g(t_i))$, define $\mathbf{C}_\alpha = \operatorname{diag}\{|l'_\alpha(Y_1, \widehat{g}_{n,\alpha}(t_1))|^2, \ldots, |l'_\alpha(Y_n, \widehat{g}_{n,\alpha}(t_n))|^2\}$ as well as $\mathbf{D}_\alpha = \operatorname{diag}\{l''_\alpha(Y_1, \widehat{g}_{n,\alpha}(t_1)), \ldots, l''_\alpha(Y_n, \widehat{g}_{n,\alpha}(t_n))\}$. A first-order Taylor expansion of the score function of (8) readily yields

$$\operatorname{Cov}\{\widehat{\mathbf{g}}_{n,\alpha}\} \approx \left[\mathbf{B}^\top \mathbf{D}_\alpha \mathbf{B} + 2\lambda \mathbf{P}_m\right]^{-1} \mathbf{B}^\top \mathbf{C}_\alpha \mathbf{B} \left[\mathbf{B}^\top \mathbf{D}_\alpha \mathbf{B} + 2\lambda \mathbf{P}_m\right]^{-1},$$

where $\mathbf{B}$ is the $n \times (K + p)$ spline design matrix with ijth element given by $B_j(t_i)$.

Inserting the approximations of the bias and variance in (9), we obtain the approximate mean-squared error (AMISE) given by

$$\begin{aligned}\operatorname{AMISE}(\alpha) =& (\widehat{\mathbf{g}}_{n,\alpha} - \widehat{\mathbf{g}}_{n,1})^\top \mathbf{H}_p (\widehat{\mathbf{g}}_{n,\alpha} - \widehat{\mathbf{g}}_{n,1}) \\ &+ \operatorname{Tr}\left\{\mathbf{H}_p \left[\mathbf{B}^\top \mathbf{D}_\alpha \mathbf{B} + 2\lambda \mathbf{P}_m\right]^{-1} \mathbf{B}^\top \mathbf{C}_\alpha \mathbf{B} \left[\mathbf{B}^\top \mathbf{D}_\alpha \mathbf{B} + 2\lambda \mathbf{P}_m\right]^{-1}\right\}.\end{aligned}$$

We propose to select $\widehat{\alpha}_n$ by minimizing AMISE(α) over a grid of 20 equidistant candidate values in [0, 1]. This grid includes both the maximum likelihood ($\alpha = 0$) and L_2-distance estimators ($\alpha = 1$) as special cases.

The bias approximation and thus the selection of α depends on the pilot estimator $\widehat{g}_{n,1}$. To reduce this dependence, Basak et al. (2021) propose to iterate the selection procedure. That is, using $\widehat{g}_{n,1}$ as pilot estimator determine the value $\widehat{\alpha}$ minimizing AMISE(α) and use the corresponding density power divergence estimate as the new “pilot" estimate instead of $\widehat{g}_{n,1}$. This procedure is repeated until $\widehat{\alpha}$ converges, which in our experience takes between 1 and 3 iterations for the vast majority of cases in our setting.

The computation of the estimator for a given value of α requires an appropriate value of the penalty parameter λ. To determine λ, we utilize the Akaike information criterion in the form proposed by Hastie and Tibshirani (1990) given by

$$\operatorname{AIC}(\lambda) = 2\sum_{i=1}^n l_\alpha(Y_i, \widehat{g}_{n,\alpha}(t_i)) + 2\operatorname{Tr}\left\{\left[\mathbf{B}^\top \mathbf{D}_\alpha \mathbf{B} + 2\lambda \mathbf{P}_m\right]^{-1} \mathbf{B}^\top \mathbf{D}_\alpha \mathbf{B}\right\}.$$

Implementations and illustrative examples of the density power divergence smoothing/penalized type spline estimators are available at https://github.com/ioanniskalogridis/Robust-and-efficient-estimation-of-nonparametric-GLMs.

5 Finite-sample performance

We now illustrate the practical performance of the proposed density power divergence spline estimators for GLM with Gaussian, Binomial and Poisson responses. We compare the following estimators.

- The adaptive density power divergence estimator discussed in the previous Section, denoted by DPD($\widehat{\alpha}$).
- The L_2-distance estimator corresponding to $\alpha = 1$, denoted by DPD(1).
- The standard penalized maximum likelihood estimator corresponding to $\alpha = 0$, abbreviated as GAM.
- The robust Huber-type P-spline estimator of Croux et al. (2012) with 40 B-spline functions, denoted by RQL, as implemented in the R-package DoubleRobGam.
- The robust local polynomial estimator of Azadeh and Salibian-Barrera (2011), denoted by RGAM, as implemented in the R-package rgam.

For the first three estimators, we use the default settings described in Sect. 4 with B-splines of order $p = 4$ combined with penalties of order $m = 2$. For Gaussian data, we use the resistant Rice-type estimator discussed in Kalogridis (2021) to estimate ϕ. This estimate is used for the DPD($\widehat{\alpha}$), DPD(1) and RQL estimators. Since RGAM is not implemented for Gaussian responses, we have replaced it in this case by the robust local linear estimator obtained with the loess function and reweighting based on Tukey's bisquare, see Cleveland (1979). Both RQL and RGAM are by default tuned for nominal 95% efficiency for Gaussian data. The penalty parameter is selected via robust AIC for RQL and robust cross-validation for RGAM.

Our numerical experiments assess the performance of the competing estimators on both ideal data and data that contain a small number of atypical observations. For our experiments, we set $t_i = i/(n+1)$, $i = 1, \ldots, n$, and consider the following two test functions:

- $g_1(t) = -\sin(25t/6)/0.8 - 1$,
- $g_2(t) = 1.8\sin(3.4x^2)$.

which were also considered by Azadeh and Salibian-Barrera (2011) and Croux et al. (2012). For Gaussian responses, we generate each Y_i from a mixture of two Gaussian distributions with mixing parameter ϵ, equal mean $g_j(t_i)$, $j \in \{1, 2\}$ and standard deviations equal to 1 and 9, respectively. This yields a proportion $(1-\epsilon)$ of ideal data distorted with a proportion ϵ of outliers originating from a Gaussian distribution with heavier tails.

For Binomial and Poisson responses, each Y_i has mean $\mu_j(t_i) = G^{-1}(g_j(t_i))$ for $j = 1$ or 2 with G the respective canonical links. We then contaminate a fraction ϵ of the responses as follows. For Binomial data, Y_i is set to 0 if the original value was equal to 1 and vice versa. For Poisson data, Y_i is replaced by a draw from a Poisson distribution with mean (and variance) equal to $3\mu_j(t_i)$. In the Binomial case, this contamination represents the frequently occurring situation of missclassified observations while we have a small number of overdispersed observations in the Poisson case. The contamination fraction ϵ takes values in $\{0, 0.05, 0.1\}$ reflecting increasingly severe contamination scenarios. For each setting, we generated 1000 samples

Table 1 Mean and median MSEs ($\times 100$) for Gaussian responses from 1000 simulated datasets with $n = 200$

g_0	ϵ	DPD($\widehat{\alpha}$)		DPD(1)		GAM		RQL		RGAM	
		Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median
g_1	0	2.98	2.28	5.29	4.29	2.81	2.22	3.23	2.72	2.93	2.31
	0.05	3.56	2.84	5.95	4.56	12.35	8.87	4.68	4.03	12.65	9.24
	0.1	4.48	3.49	5.93	4.98	20.99	15.72	7.85	6.54	22.89	16.90
g_2	0	4.34	3.37	7.57	6.72	3.40	2.66	4.64	3.91	3.38	2.75
	0.05	5.21	4.23	8.27	7.37	15.86	11.16	6.72	5.93	14.80	11.25
	0.1	6.24	5.26	8.67	7.88	28.52	19.95	11.19	10.21	25.85	19.11

Table 2 Mean and median MSEs ($\times 100$) for Binomial responses from 1000 simulated datasets with $n = 200$

g_0	ϵ	DPD($\widehat{\alpha}$)		DPD(1)		GAM		RQL		RGAM	
		Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median
g_1	0	0.34	0.32	0.40	0.34	0.54	0.33	0.50	0.47	0.68	0.43
	0.05	0.53	0.46	0.54	0.48	0.64	0.52	0.62	0.55	0.87	0.54
	0.1	0.88	0.80	0.87	0.80	1.05	0.86	0.95	0.86	1.36	1.01
g_2	0	0.55	0.47	0.58	0.49	0.78	0.55	0.76	0.72	0.98	0.66
	0.05	0.65	0.55	0.66	0.58	0.84	0.63	0.87	0.83	1.18	0.75
	0.1	0.90	0.82	0.91	0.82	1.10	0.86	1.05	1.02	1.38	0.98

of size $n = 200$. We evaluate the performance of the competing estimators via their mean-squared error (MSE), given by $\sum_{i=1}^{n} |\widehat{\mu}_i - \mu_i|^2/n$. Since MSE distributions are right-skewed in general, we report both the average and median of the MSEs. Tables 1, 2 and 3 report the results for data with Gaussian, Binomial and Poisson responses, respectively.

The simulation experiments lead to several interesting findings. For uncontaminated Gaussian data, Table 1 shows that GAM and RGAM perform slightly better than DPD($\widehat{\alpha}$) and RQL while the L_2-distance estimator DPD(1) is less efficient. However, even with a small amount of contamination, the advantage of GAM and RGAM evaporates. RQL offers more protection against outlying observations, but it is outperformed by DPD($\widehat{\alpha}$) and also by DPD(1) in case of more severe contamination. This difference may be explained by the use of the monotone Huber function in RQL versus the redescending squared exponential loss function in DPD (Maronna et al. 2019).

As seen in Table 2, for binomial data DPD($\widehat{\alpha}$) remains highly efficient and robust. In ideal data, DPD($\widehat{\alpha}$) even attains a lower mean and median MSE than GAM, although this is likely an artefact of sampling variability. A surprising finding here is the exceptionally good performance of DPD(1) in clean data. It outperforms both RQL and RGAM, performing almost as well as DPD($\widehat{\alpha}$). This fact suggests that efficiency loss for Binomial responses is much less steep as a function of α than for the Gaussian case. See Ghosh and Basu (2016) for efficiency considerations in the parametric case.

Table 3 Mean and median MSEs ($\times 100$) for Poisson responses from 1000 simulated datasets with $n = 200$

g_0	ϵ	DPD($\widehat{\alpha}$)		DPD(1)		GAM		RQL		RGAM	
		Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median
g_1	0	1.02	0.80	1.03	0.82	1.15	0.76	1.17	0.94	2.57	1.61
	0.05	1.13	0.94	1.21	1.04	1.67	0.98	1.27	1.00	3.15	1.93
	0.1	1.39	1.06	1.40	1.10	2.76	1.75	1.72	1.34	4.27	2.69
g_2	0	10.61	7.62	17.70	14.57	10.12	7.62	14.93	12.05	11.89	9.24
	0.05	13.53	10.22	18.61	15.82	59.32	44.28	15.81	13.97	15.34	12.01
	0.1	17.72	14.17	20.67	17.26	152.7	140.1	22.57	21.34	26.06	21.20

As evidenced from Table 3, the situation for count responses is reminiscent of the situation with Gaussian responses. In clean data, GAM exhibits the best median performance, closely followed by DPD($\widehat{\alpha}$). RQL performs slightly worse than DPD($\widehat{\alpha}$) in clean data, but the difference between these two estimators becomes more pronounced in favour of DPD($\widehat{\alpha}$) the heavier the contamination. In ideal data, DPD(1) performs adequately for $\mu_1(t)$, but is rather inefficient for $\mu_2(t)$. RGAM performs better for $\mu_2(t)$ than for $\mu_1(t)$, but is outperformed by DPD($\widehat{\alpha}$) in both cases. Some further insights on the good performance of DPD($\widehat{\alpha}$) are given in the supplementary material.

6 Applications

6.1 Type 2 diabetes in the USA

The National Health and Nutrition Examination Survey is a series of home interviews and physical examinations of randomly selected individuals conducted by the US Centers for Disease Control and Prevention (CDC), normally on a yearly basis. Due to the Covid-19 pandemic, the cycle of data collection between 2019 and 2020 was not completed and, in order to obtain a nationally representative sample, the CDC combined these measurements with data collected in the 2017–2018 period leading to a dataset consisting of 9737 individuals. We refer to https://wwwn.cdc.gov/nchs/nhanes and the accompanying documentation for a detailed description of the survey and the corresponding data.

For this example, we study the relationship between type 2 diabetes and high-density-lipoprotein (HDL) cholesterol as well as the relationship between type 2 diabetes and the amount of glycohemoglobin in one's bloodstream. Measuring the concentration of the latter often constitutes an expedient way of detecting diabetes while low HDL cholesterol values are considered a risk factor. The response variable is binary with value 1 if the interviewee has been diagnosed with type 2 diabetes and 0 otherwise, while the covariates are continuous. The response variable is plotted versus either covariate in the left and right panels of Fig. 1.

Since the way that the covariates influence the probability of diabetes cannot be specified in advance, we apply our methodology to estimate the two regression func-

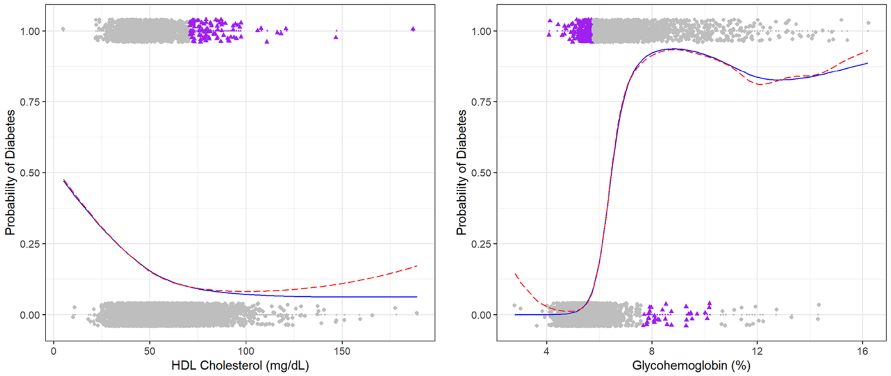

Fig. 1 Left: Incidence of diabetes versus HDL cholesterol (mg/dL). Right: Incidence of diabetes versus glycohemoglobin (%). The lines (——, – – –) correspond to DPD($\widehat{\alpha}$) and GAM estimates, respectively. Observations indicated with (▲) exhibit large Anscombe residuals according to DPD ($\widehat{\alpha}$) (color figure online)

tions in a nonparametric manner. The algorithm described in Sect. 4 selects $\alpha = 1$ in both cases, indicating the presence of several atypical observations among our data. The estimates are depicted with the solid blue lines in the left and right panels of Fig. 1. For comparison, the standard GAM estimates obtained with the `gam` function of the `mgcv` package (Wood 2017) are depicted by dashed red lines in the figure.

Comparing the robust and GAM estimates reveals that, despite some areas of agreement, there is notable disagreement between the estimates. In particular, the estimated probabilities of Type 2 diabetes differ significantly for large values of HDL cholesterol and for both small and medium-large concentrations of glycohemoglobin. Inspection of the panels suggests that the GAM estimates are strongly drawn towards a number of atypical observations, corresponding to individuals with high HDL cholesterol but no diabetes and diabetes patients with low and medium concentrations of glycohemoglobin, respectively. The results in both cases are counter-intuitive, as, for healthy individuals with good levels of HDL cholesterol or low levels of glucose, GAM predicts a non-negligible probability of diabetes. By contrast, the robust DPD($\widehat{\alpha}$)-estimates remain unaffected by these atypical observations leading to more intuitive estimates.

Since robust estimates are less attracted to outlying observations, such observations can be detected from their residuals. For GLMs, we may make use of Anscombe residuals (McCullagh and Nelder 1983, p. 29), which more closely follow a Gaussian distribution than their Pearson counterparts. For Bernoulli distributions, these are given by

$$r_{A,i} = \frac{\mathrm{IB}(Y_i, 2/3, 2/3) - \mathrm{IB}(\widehat{\mu}_i, 2/3, 2/3)}{\widehat{\mu}_i^{1/6}(1-\widehat{\mu}_i)^{1/6}}, \quad (i = 1, \ldots, n),$$

where $\mathrm{IB}(x, a, b) = \int_0^x t^{a-1}(1-t)^{b-1}\mathrm{d}t$. We classify an observation as an outlier if $|r_{A,i}| \geq 2.6$, which is a conventional cut-off value for the standard Gaussian distribution. The outliers for our examples are shown in Fig. 1 with a different shape and

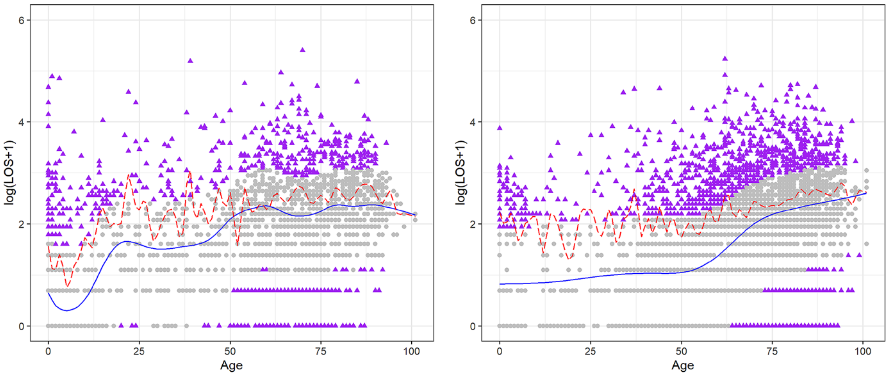

Fig. 2 Left: Logarithm of LOS+1 versus age for patients with disorders of the respiratory system. Right: Logarithm of LOS+1 versus age for patients with disorders of the circulatory system. The lines (——, — —) correspond to DPD($\widehat{\alpha}$) and GAM estimates, respectively. Observations indicated with (▲) exhibit large Anscombe residuals according to DPD($\widehat{\alpha}$) (color figure online)

colour coding. These plots show that these outliers are largely located in the areas in which the DPD($\widehat{\alpha}$) and GAM estimates differ, thus confirming the sensitivity of GAM estimates.

6.2 Length of hospital stay in Switzerland

Patients admitted into Swiss hospitals are classified into homogeneous diagnosis related groups (DRG). The length of stay (LOS) is an important variable in the determination of the cost of treatment so that the ability to predict LOS from the characteristics of the patient is helpful. Herein, we use the age of the patient (in years) to predict the average LOS (in days) for two DRG comprising *Diseases and disorders of the respiratory system* and *Diseases and disorders of the circulatory system*, which we henceforth abbreviate as DRG 1 and DRG 2. We assume that LOS can be modelled as a Poisson random variable with

$$\log(\mathbb{E}\{\text{LOS}_j\}) = g_j(\text{Age}), \quad j \in \{1, 2\},$$

for unknown functions g_1 and g_2 corresponding to DRG 1 and DRG 2, respectively. The data consisting of 2807 and 3922 observations are plotted in the left and right panels of Fig. 2, together with the DPD($\widehat{\alpha}$) and GAM estimates of their regression functions.

The plots show that the GAM estimates lack smoothness and are shifted upwards in relation to the DPD($\widehat{\alpha}$)-estimates. Both facts are attributable to a lack of resistance of GAM estimates towards the considerable number of patients with atypically lengthy hospital stays given their age. It should be noted that, while it is always possible to manually increase the smoothness of the GAM estimates, non-robust automatic methods are very often affected by outlying observations resulting in under or oversmoothed estimates, as observed by Cantoni and Ronchetti (2001a). Thus, in practice, robust methods of estimation need to be accompanied by robust model selection criteria.

On the other hand, our algorithm selects $\alpha = 1$ in both cases which combined with the robust AIC proposed in Sect. 4 leads to reliable estimates for the regression functions, even in the presence of numerous outlying observations. These estimates largely conform to our intuition, as they suggest that older patients are, on average, more likely to experience longer hospital stays. To detect the outlying observations, we may again use the Anscombe residuals of the DPD($\widehat{\alpha}$)-estimates, which in the Poisson case are given by

$$r_{A,i} = \frac{\frac{3}{2}(Y_i^{2/3} - \widehat{\mu}_i^{2/3})}{\widehat{\mu}_i^{1/6}}, \quad (i = 1, \ldots, n).$$

Observations with $|r_{A,i}| \geq 2.6$ are indicated with a different shape and colour coding in Fig. 2. These panels suggest that while there exist patients with atypically brief stays, the vast majority of outliers is in the opposite direction, thereby explaining the upper vertical shift of the sensitive GAM estimates.

7 Discussion

This paper greatly extends penalized likelihood methods for nonparametric estimation in GLMs and derives new and important theoretical properties for this broad class of estimators. In practice, the proposed class of estimators behaves similarly to non-robust GAM estimators in the absence of atypical observations, but exhibits a high degree of robustness in their presence. These properties make the proposed methodology particularly well suited for the analysis of many complex datasets commonly encountered nowadays, such as the diabetes and length of hospital stay data analysed in Sect. 6.

There are a number of interesting and practically useful extensions we aim to consider in future work. These include the case of higher-dimensional nonparametric components, modelled, for example, with thin-plate or tensor product penalties (Wood 2017, Chapter 5), as well as more general semi-parametric models based on density power divergence that would allow for both parametric and nonparametric components. Currently, our density power divergence estimator depends on the tuning parameter α and for the selection of α we have developed a data-dependent scheme. An intriguing alternative would be a composite penalized estimator involving several values of α, as proposed by Zou and Yuan (2008) in the context of quantile regression. Such an approach has the potential of producing another resistant yet highly efficient competitor to standard maximum likelihood estimators.

Supplementary information The supplementary material of this paper contains the proofs of our theoretical results, namely Theorem 1, Corollary 1, Proposition 1, Theorem 2 and Corollary 2, as well as a number of auxiliary results. We also provide three additional figures and remarks on the simulation study of Sect. 5.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11749-023-00866-x.

Acknowledgements The authors are grateful to two anonymous referees and the associate editor for helpful remarks that lead to a much improved manuscript. They are also grateful to Professor Alfio Marazzi (Lausanne University Hospital) for providing the length of hospital stay data. The research of I. Kalogridis was supported by the Research Foundation-Flanders (project 1221122N). G. Claeskens acknowledges support from the KU Leuven Research Fund C1-project C16/20/002.

References

Aeberhard WH, Cantoni E, Marra G, Radice R (2021) Robust fitting for generalized additive models for location, scale and shape. Statistics Comput 31:1–16

Azadeh A, Salibian-Barrera M (2011) An outlier-robust fit for generalized additive models with applications to disease outbreak detection. J Am Statistical Assoc 106:719–731

Basak S, Basu A, Jones MC (2021) On the optimal density power divergence tuning parameter. J Appl Statistics 48:536–556

Basu A, Harris IR, Hjort NL, Jones MC (1998) Robust and efficient estimation by minimising a density power divergence. Biometrika 85:549–559

Bianco AM, Boente G, Sombielle S (2011) Robust estimation for nonparametric generalized regression. Statistics Probab Lett 81:1986–1994

Boente G, Ruiz M, Zamar R (2010) On a robust local estimator for the scale function in heteroscedastic nonparametric regression. Statistics Probab Lett 80:1185–1195

Cantoni E, Ronchetti E (2001) Resistant selection of the smoothing parameter for smoothing splines. Statistics Comput 11:141–146

Cantoni E, Ronchetti E (2001) Robust inference for generalized linear models. J Am Statistical Assoc 96:1022–1030

Claeskens G, Hjort NL (2008) Model selection and model averaging. Cambridge University Press, Cambridge

Claeskens G, Krivobokova T, Opsomer JD (2009) Asymptotic properties of penalised spline estimators. Biometrika 96:529–544

Cleveland WS (1979) Robust locally weighted regression and smoothing scatterplots. J Am Statistical Assoc 74:829–836

Croux C, Gijbels I, Prosdocimi I (2012) Robust estimation of mean and dispersion functions in extended generalized additive models. Biometrics 68:31–44

Cox DD, O'Sullivan F (1990) Asymptotic analysis of penalized likelihood and related estimators. Ann Statistics 18:1676–1695

de Boor C (2001) A practical guide to splines, Revised. Springer, New York

Eggermont PPB, LaRiccia VN (2009) Maximum penalized likelihood estimation, vol II. Regression. Springer, New York

Eilers PHC, Marx BD (1996) Flexible smoothing with B-splines and penalties. Statistical Sci 11:89–102

Eubank RL (1999) Nonparametric regression and spline smoothing, 2nd edn. CRC Press, New York

Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Statistical Assoc 96:1348–1360

Ghement IR, Ruiz M, Zamar R (2008) Robust estimation of error scale in nonparametric regression models. J Statistical Plann Inference 138:3200–3216

Ghosh A, Basu A (2013) Robust estimation for independent non-homogeneous observation using density power divergence with applications to linear regression. Electron J Statistics 7:2420–2456

Ghosh A, Basu A (2015) Robust estimation for non-homogeneous data and the selection of the optimal tuning parameter: the density power divergence approach. J Appl Statistics 42:2056–2072

Ghosh A, Basu A (2016) Robust estimation in generalized linear models: the density power divergence approach. TEST 25:269–290

Green PJ, Silverman BW (1994) Nonparametric regression and generalized linear models. Chapman & Hall, London

Gu C (2013) Smoothing spline ANOVA models, 2nd edn. Springer, New York

Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (2011) Robust statistics: the approach based on influence functions. Wiley, New York

Hastie TJ, Tibshirani RJ (1990) Generalized additive models. Chapman & Hall, Suffolk

Hastie TJ, Tibshirani RJ, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, New York

Huber PJ, Ronchetti EM (2009) Robust statistics, 2nd edn. Wiley, Hoboken, NJ

Kalogridis I (2021) Asymptotics for M-type smoothing splines with non-smooth objective functions. TEST 31:373–389

Kalogridis I, Van Aelst S (2021) Robust penalized spline estimation with difference penalties. Econometrics and Statistics. https://doi.org/10.1016/j.ecosta.2021.07.005

Kauermann G, Krivobokova T, Fahrmeir L (2009) Some asymptotic results on generalized penalized spline smoothing. J R Stat Soc Ser B Statistical Methodol 71:487–503

Kim Y-J, Gu C (2004) Smoothing spline Gaussian regression: more scalable computation via efficient approximation. J R Stat Soc Ser B Statistical Methodol 66:337–356

Mammen E, van de Geer S (1997) Penalized quasi-likelihood estimation in partial linear models. Ann Statistics 25:1014–1035

Maronna RA, Martin D, Salibián-Barrera M, Yohai VJ (2019) Robust statistics: theory and methods, 2nd edn. Wiley, Chichester

McCullagh P, Nelder JA (1983) Generalized linear models. Chapman & Hall, London

O'Sullivan F (1986) A statistical perspective of ill-posed problems. Statistical Sci 1:502–518

Pearce ND, Wand MP (2006) Penalized splines and reproducing kernel methods. Am Statistician 60:233–240

Ruppert D, Wand MP, Carroll RJ (2003) Semiparametric regression. Cambridge University Press, Cambridge

Scott DW (2001) Parametric statistical modeling by minimum integrated square error. Technometrics 43:274–285

van de Geer S (2000) Empirical processes in M-estimation. Cambridge University Press, New York, NY

van der Vaart AW (1998) Asymptotic statistics. Cambridge University Press, New York, NY

Wahba G (1990) Spline models for observational data. SIAM, Philadelphia, Pen

Wang X, Jiang Y, Huang M, Zhang H (2013) Robust variable selection with exponential squared loss. J Am Statistical Assoc 108:632–643

Wong RKW, Yao F, Lee TCM (2014) Robust estimation for generalized additive models. J Comput Graph Statistics 23:270–289

Warwick J, Jones MC (2005) Choosing a robustness tuning parameter. J Statistical Comput Simul 75:581–588

Wood SN (2017) Generalized additive models, 2nd edn. CRC Press, Boca Raton, FL

Xiao L (2019) Asymptotic theory of penalized splines. Electron J Statistics 13:747–794

Zou H, Yuan M (2008) Composite quantile regression and the oracle model selection theory. Ann Statistics 36:1108–1126

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.