Journal Pre-proof

Robust penalized estimators for functional linear regression

Ioannis Kalogridis, Stefan Van Aelst

PII: S0047-259X(22)00095-1
DOI: https://doi.org/10.1016/j.jmva.2022.105104
Reference: YJMVA 105104

To appear in: *Journal of Multivariate Analysis*

Received date : 26 August 2021
Revised date : 12 September 2022
Accepted date : 12 September 2022

Please cite this article as: I. Kalogridis and S. Van Aelst, Robust penalized estimators for functional linear regression, *Journal of Multivariate Analysis* (2022), doi: https://doi.org/10.1016/j.jmva.2022.105104.

Robust penalized estimators for functional linear regression

Ioannis Kalogridis[a,*], Stefan Van Aelst[a]

[a]Department of Mathematics, KU Leuven (University of Leuven), Celestijnenlaan 200B, 3001 Leuven, Belgium

Abstract

Functional data analysis is a fast evolving branch of statistics, but estimation procedures for the popular functional linear model either suffer from lack of robustness or are computationally burdensome. To address these shortcomings, a flexible family of penalized lower-rank estimators based on a bounded loss function is proposed. The proposed class of estimators is shown to be consistent and can attain high rates of convergence with respect to prediction error under weak regularity conditions. These results can be generalized to higher dimensions under similar assumptions. The finite-sample performance of the proposed family of estimators is investigated by a Monte-Carlo study which shows that these estimators reach high efficiency while offering protection against outliers. The proposed estimators compare favourably to existing robust as well as non-robust approaches. The good performance of our method is also illustrated on a complex real dataset.

Keywords: Asymptotics, functional data, regularization, robustness
2000 MSC: 62G35, 62R10, 62G20

1. Introduction

In recent years, technological innovations and improved storage capabilities have led practitioners to observe and record increasingly complex high-dimensional data. Among others, data that are characterized by an underlying functional structure have attracted considerable research interest, following works such as [32–34]. Particular interest has been devoted to the functional linear model, relating a scalar response Y to a random function X, which is viewed as an element of $(\Omega, \mathcal{A}, \mathcal{P})$ with sample paths in $\mathcal{L}^2(\mathcal{I})$, through the model

$$Y = \alpha_0 + \int_{\mathcal{I}} X(t) f_0(t) dt + \sigma_0 \epsilon. \tag{1}$$

Here, $\alpha_0 \in \mathbb{R}$ is the intercept, f_0 is a square integrable coefficient (weight) function defined on a compact interval $\mathcal{I}$ of a Euclidean space, σ_0 is an unknown scale parameter and ϵ is a random error, that is assumed to be independent of X. Typically, ϵ is also assumed to possess finite second moments, but this assumption is not needed for the theoretical results in this paper.

The vast domain of applications of the model, ranging from meteorology [34] and chemometrics [11] to diffusion tensor imaging tractography [14], has spurred the development of numerous novel estimation methods. Since estimating the coefficient function β is an infinite dimensional problem, regularization through dimension reduction or penalization is crucial for the success of these methods. Regressing on the scores of the leading functional principal components [3] is the oldest and perhaps to this day the most popular method of estimation. However, although consistent [15], functional principal component regression may fail to yield smooth estimates of the coefficient function, even in moderately large samples. This fact has motivated proposals that explicitly impose smoothness of the estimated coefficient function. Cardot et al. [4] proposed estimation through a penalized spline expansion while functional extensions of smoothing splines have been proposed and studied by Crambes et al. [6] and Yuan and Cai [48].

*Corresponding author.
Email addresses: ioannis.kalogridis@kuleuven.be (Ioannis Kalogridis), stefan.vanaelst@kuleuven.be (Stefan Van Aelst)

A hybrid approach between principal component and penalized spline regression has been developed by Reiss and Ogden [36] and Goldsmith et al. [13], who combine these methods in order to attain greater flexibility. Variable selection ideas have also been adapted to the functional regression setting. James et al. [17] proposed imposing sparsity on higher order derivatives of a high dimensional basis expansion of β in order to produce more interpretable estimates. Expressing the coefficient function in the wavelet domain, Zhao et al. [52] proposed an ℓ_1 regularization scheme in order to select the most relevant resolutions and ensure stable and accurate estimates of a wide variety of coefficient functions. For more details on existing estimation methods as well as informative comparisons, one may consult the comprehensive review papers of Morris [28] and Reiss et al. [37].

Since all of the above methods rely on generalized least-squares type estimators, a drawback in their use is that the presence of outliers can have a serious effect on the resulting estimates. To address this lack of robustness, more robust estimation procedures have been introduced. Maronna and Yohai [26] proposed a robust version of the smoothing spline estimator of Crambes et al. [6] but did not study theoretical properties of their method. Shin and Lee [42] have extended the work of Yuan and Cai [48] by considering more outlier-resistant loss functions and showed that under regularity conditions their M-type smoothing spline estimator attains the same rates of convergence as its least-squares counterpart. Similarly, Qingguo [31] generalized the work of Hall and Horowitz [15] to functional principal component regression with a general convex loss function. More recently, Boente et al. [2] proposed a family of sieves estimators based on bounded loss functions and B-spline expansions and investigated rates of convergence with respect to the prediction error.

In general, sieves estimators based on either functional principal components or B-splines and smoothing spline estimators can be considered to be situated on the two ends of a spectrum. Unpenalized sieves estimators are easy to implement, yet frequently result in either undersmoothed or oversmoothed estimates of the regression function. This undesirable feature results from the discrete nature of their smoothing parameter, which in this case is the dimension of the basis. On the other hand, smoothing spline estimators, while capable of yielding estimates with the right amount of smoothness, can be unwieldy due to their high dimension. In particular, the requirement to have as many basis functions as the sample size leads to computationally challenging estimators that are prone to instabilities due to the often complex nature of functional data. In the nonparametric regression framework, the case for lower-rank representations on the grounds of simplicity has already been made by Wahba [46]. For functional regression, an even stronger case can be made due to the lack of banded matrices that enable fast computational algorithms for smoothing splines in this setting.

As a compromise between these two types of estimators, this paper introduces and studies a family of lower-rank penalized estimators based on the principle of MM-estimation, as described by Yohai [50]. The proposed class of estimators exhibit a high degree of robustness against both vertical outliers and leverage points, while also maintaining high efficiency under Gaussian errors. In our opinion, this class of estimators fills an important void in the literature by providing a family of flexible and resistant estimators that is also computationally feasible. Our framework does not only include the popular B-spline basis combined with a quadratic roughness penalty, but also many other basis systems combined with a wealth of possible penalties. Examples include the Fourier basis with the harmonic acceleration penalty introduced by Ramsay and Silverman [34] and the wavelet basis with bounded variation or Besov penalties [44, Chapter 10]. It should be noted that the theory for functional linear regression developed herein cannot be deduced from earlier results in the field of nonparametric regression with robust penalized spline estimators, such as Kalogridis and Van Aelst [20, 21], due to the more complex nature of the functional linear regression model in (1), which involves an infinite-dimensional predictor rather than a one-dimensional scalar.

The remainder of the paper is organized as follows. Section 2 introduces the proposed family of penalized estimators and discusses some popular choices of basis systems and penalties in more detail. In Section 3 we study asymptotic properties of these estimators. We show that under mild regularity conditions the estimators achieve a high rate of convergence with respect to the commonly considered prediction error. Our regularity conditions do not require the existence of any moments of the error term, allowing in effect for very heavy-tailed error distributions. Our analysis also uncovers a useful error decomposition pointing to the roles of the variance as well as the twin biases stemming from modelling and regularization. Sections 4 and 5 illustrate the competitive finite-sample performance of the proposed estimator in a Monte Carlo study and in real data. Section 6 contains a final discussion while all proofs are collected in the appendix.

2. Robust penalized estimators for functional linear regression

2.1. Penalized MM-estimators with general bases and penalties

Let us consider independent and identically distributed tuples $(X_1, Y_1), \ldots, (X_n, Y_n)$ which satisfy model (1). For simplicity we shall identify $\mathcal{I}$ with $[0, 1]$, without loss of generality. A popular estimation approach for the functional linear model [34, Chapter 15] expands the functional slope f_0 in terms of a dense set of $\mathcal{L}^2([0, 1])$ functions $\{f_i\}_{i\in\mathbb{N}}$, then truncates this expansion and finally estimates the coefficients using a roughness penalty. Let $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ denote the usual $\mathcal{L}^2([0, 1])$ norm and inner product, respectively. Moreover, let Θ_K denote the K-dimensional linear subspace of $\mathcal{L}^2([0, 1])$ spanned by $f_1, \ldots, f_K$, then this strategy amounts to solving

$$(\widehat{\alpha}_{LS}, \widehat{f}_{LS}) = \underset{\alpha\in\mathbb{R}, f\in\Theta_K}{\operatorname{argmin}} \left[\frac{1}{n}\sum_{i=1}^{n} |Y_i - \alpha - \langle X_i, f\rangle|^2 + \lambda\|f^{(q)}\|^2\right]. \tag{2}$$

Hence, the roughness penalty is placed on the integrated squared qth derivative of f and it is weighted by a penalty parameter $\lambda \geq 0$, which is usually chosen in a data-driven way. The penalty parameter places a premium on the roughness of the estimated function as measured by its integrated squared qth derivative. Large values of λ force the estimated coefficient function to behave essentially like a polynomial of degree at most $q - 1$ while small values of λ produce more wiggly estimates. It is important to note that for such estimators regularization is accomplished by both restricting the basis functions ($f \in \Theta_K$) and penalizing roughness. This strategy leads to more complex estimators than unpenalized sieve estimators but considerably less complex estimators than the smoothing spline estimators of Crambes et al. [6], Yuan and Cai [48] and Shin and Lee [42].

It is well-known that the least-squares criterion employed in (2) yields estimators that are susceptible to outlying observations. To protect against such anomalies we propose to replace the square loss function by a bounded loss function ρ and estimate the unknown quantities according to

$$(\widehat{\alpha}_n, \widehat{f}_n) = \underset{\alpha\in\mathbb{R}, f\in\Theta_K}{\operatorname{argmin}} \left[\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{Y_i - a - \langle X_i, f\rangle}{\widehat{\sigma}_n}\right) + \lambda\mathcal{J}(f)\right], \tag{3}$$

where $\widehat{\sigma}_n$ is a robust estimator of the scale of the error and $\mathcal{J}(f) : \Theta_K \to \mathbb{R}_+$ is a general penalty functional on Θ_K, usually a seminorm. For $\lambda = 0$ the penalty term vanishes and we obtain an unpenalized sieve estimator, such as the B-spline estimator proposed by Boente et al. [2]. On the other hand, for $\lambda \to \infty$ the penalty will dominate the objective function and forces the estimator to lie in the null-space of $\mathcal{J}(\cdot)$. The present set-up is very general and allows for a wide variety of approximating subspaces and penalties. To illustrate this flexibility, we now discuss three important examples of basis systems and penalties that are permitted within our framework.

Example 1 (B-splines with derivative or difference penalties)**.** Fix an integer $p \geq 1$, select $0 < t_1 < \ldots < t_K < 1$ and define the spline subspace

$$\Theta_{K+p} = \Big\{f : f(x) = \sum_{j=1}^{K+p} f_j B_{j,p}(x)\Big\},$$

where $B_{j,p}, j \in \{1, \ldots, K + p\}$, are the B-splines of order p supported by $t_1 \ldots, t_K$, with $2p$ arbitrary boundary knots. For $p = 1$, Θ_{K+p} consists of all step functions with jumps at the knots t_i while for $p \geq 2$, Θ_{K+p} is a subspace of $C^{p-2}([0, 1])$ with the property that each $f \in \Theta_{K+p}$ is a polynomial of order p on each subinterval $[t_i, t_{i+1}]$. The common choice $\mathcal{J}(f) = \|f^{(q)}\|^2$ for some integer $q < p$ was introduced by O'Sullivan [30]. Another popular choice is the P-spline penalty [10], given by $\mathcal{J}(f) = \sum_{j=q+1}^{K+p} |\Delta^q f_j|^2$, where Δ^q refers to the qth-order backward difference operator. This difference penalty largely retains the mathematical properties of the derivative penalty, but results in much simpler expressions. In the frequently used setting of equidistant knots, the derivative and difference penalties on spline subspaces are scaled versions of one another, see, e.g., Proposition 1 of Kalogridis and Van Aelst [21].

Example 2 (Fourier expansion with derivative or harmonic acceleration penalties)**.** Consider the trigonometric sieve given by

$$\Theta_K = \Big\{f : f(x) = \alpha_0 + \sum_{j=1}^{K} \{\alpha_j \cos(2\pi jx) + \beta_j \sin(2\pi jx)\}\Big\}.$$

This sieve consists of infinitely differentiable functions with increasing frequency. Unlike the B-spline basis, the Fourier basis is not local, but it is orthonormal and its derivatives are orthogonal, resulting in simple expressions. For instance, taking $\mathcal{J}(f) = \|f^{(q)}\|^2$, as in Li and Hsing [23], leads to $\mathcal{J}(f) = \mathbf{f}^\top \mathbf{D}\mathbf{f}$ with $\mathbf{f} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $\mathbf{D} = \text{diag}\{(2\pi)^{2q}, (2\pi)^{2q}, (4\pi)^{2q}, (4\pi)^{2q}, \ldots, (2\pi K)^{2q}\}$. Another possibility is the harmonic acceleration penalty proposed in Ramsay and Silverman [34, Chapter 15] which is given by $\mathcal{J}(f) = \int_0^1 \left|(4\pi^2)f'(x) + f^{(3)}(x)\right|^2 dx$. Interestingly, this penalty shrinks the solution towards a function of the form $f(x) = \alpha_0 + \alpha_1 \sin(2\pi x) + \beta_1 \cos(2\pi x)$.

Example 3 (Wavelets with total variation or ℓ_1 penalties)**.** Choose a scaling function ϕ and a mother wavelet ψ that are orthonormal in $\mathcal{L}^2([0,1])$. Now, for $(j,k) \in \mathbb{N} \times \mathbb{N}$ put $\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k)$ and $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$. Fix $j_0 \in \mathbb{N}$ and set $J = \log_2 K - 1$ for $K \in \mathbb{N}$. The wavelet subspace with primary decomposition level j_0 is given by

$$\Theta_K = \left\{ f : f(x) = \sum_{k=0}^{2^{j_0}-1} f'_{j_0,k}\phi_{j_0,k}(x) + \sum_{j=j_0}^{J}\sum_{k=0}^{2^j-1} f_{j,k}\psi_{j,k}(x) \right\}.$$

This wavelet subspace involves $2^{J+1} = K$ coefficients. Possible penalties are the total variation penalty with $q = 1$, i.e., $\mathcal{J}(f) = \int_0^1 |f'(x)|dx$ or the ℓ_1 penalty on all the coefficients given by

$$\mathcal{J}(f) = \sum_{k=0}^{2^{j_0}-1} |f'_{j_0,k}| + \sum_{j=j_0}^{J}\sum_{k=0}^{2^j-1} |f_{j,k}|,$$

as used by Zhao et al. [52] in the context of least-squares estimation.

A widely used family of bounded smooth ρ-functions that is useful for our purposes is the family of Tukey bisquare loss functions, defined as

$$\rho_c(x) = \begin{cases} 1 - \left\{1 - (x/c)^2\right\}^3 & |x| \le c \\ 1 & |x| > c, \end{cases},$$

where $c > 0$ is a tuning parameter that determines the trade-off between robustness and efficiency [27]. In the central part, that is, for $|x| \le c$, ρ the loss function is strictly increasing and it smoothly transitions to a constant function as $|x| \to c$. Thus, the loss incurred by large residuals is constant leading to regression estimators that are impervious to large outliers.

The scale estimate $\widehat{\sigma}_n$ is an important part of the estimator, as it essentially acts as an additional tuning parameter for the loss function. A robust scale estimate may be obtained from an M-scale of the residuals of an S-estimator [38]. In particular, for $\alpha \in \mathbb{R}$ and $f \in \Theta_K$ let $\widehat{\sigma}_n = \widehat{\sigma}_n(\mathbf{r})$ be an M-scale estimate based on a vector of residuals $\mathbf{r}(\alpha, f) = (r_1(\alpha, f), \ldots, r_n(\alpha, f))$, with $r_i(a, f) = Y_i - \alpha - \langle X_i, f\rangle$, $i = 1 \ldots, n$. Then an S-estimator $(\widehat{\alpha}^S, \widehat{f}^S)$ is defined as

$$(\widehat{\alpha}^S, \widehat{f}^S) = \underset{\alpha\in\mathbb{R}, f\in\Theta_K}{\text{argmin}}\ \widehat{\sigma}(\mathbf{r}(\alpha, f)). \tag{4}$$

We set $\widehat{\sigma}_n$ equal to the S-scale estimate, which is given by the minimum of the objective function in (4), i.e., $\widehat{\sigma}_n = \widehat{\sigma}(\mathbf{r}(\alpha^S, f^S))$. Note that S-estimators are well-defined in our setting as $K < n$, i.e., there are fewer parameters than observations. This will also be a requirement for our asymptotic results, see Section 3 below. Hence, computationally efficient algorithms, such as the fast-S algorithm proposed by Salibian-Barrera and Yohai [40], can be applied to obtain the solution of (4).

2.2. Computational aspects

The penalized MM-estimator in (3) depends on the choice of the approximating subspace, its dimension and the penalty parameter. In this section, we outline a number of possible strategies for their selection, but first we briefly discuss the computation of penalized MM-estimates. To this end, we need to differentiate between quadratic and non-quadratic penalties. For quadratic penalties, that is, for penalties which can be written as $\mathcal{J}(f) = \mathbf{f}^\top \mathbf{D}\mathbf{f}$ for

some positive semi-definite $\mathbf{D}$, a fast computational procedure may be developed along the lines of the penalized variant of iteratively reweighted least-squares given in Maronna [25]. To better guarantee that the algorithm returns a global minimum, we recommend initiating the iterations from the robust unpenalized S-estimate given by (4). For non-quadratic penalties, such as the ℓ_1-penalty for instance, we recommend the use of the iterative LARS algorithm proposed by Smucler and Yohai [43], again starting from the unpenalized S-estimate.

Let us now consider the choices that need to be made for the penalized MM-estimator. The dimension K of the subspace seems to be the least critical for the success of the estimator. Indeed, extensive experience with lower-rank penalized estimators [39, 47] has shown that the dimension does not make much difference for the resulting solution as long as the approximating subspace is rich enough, but K is still smaller than the sample size n. In our experience, a choice such as $K = [\min\{40, n/4\}]$, which ensures at least 4 observations per basis function and puts a cap at 40 basis functions, is appropriate for many situations. The number of basis functions can be increased beyond 40 in highly complex situations, but these tend to be rather rare in practice.

For the choice of the basis system some guidelines already exist in the literature. For instance, Ramsay and Silverman [34] recommend using the Fourier system for periodic data and the B-spline system otherwise. As we shall see in Section 3, both systems require smoothness of the coefficient function f_0, in order to attain high rates of convergence. For cases in which the regression function is suspected to be less smooth, possibly with local characteristics, such as spikes, one may opt for the wavelet system instead. However, in this case special attention must be devoted not only to the tuning parameter λ, but also to the level of the decomposition j_0. As this is a discrete parameter, however, the additional computational burden is not excessive.

Let $\mathbf{r}_- = (r_{-1}, \ldots, r_{-n})^\top$ denote an approximation to the leave-one out residuals as given by Maronna [25], for example. To determine the penalty parameter λ in a data-driven way, we propose to select the value of λ that minimizes the robust cross-validation (RCV) criterion

$$\mathrm{RCV}(\lambda) = \tau(\mathbf{r}_-)^2,$$

where τ denotes the robust and efficient τ-scale introduced by Yohai and Zamar [51] with constants equal to $c_1 = 3$ and $c_2 = 5$. This criterion has also been used by Maronna [25] and Maronna and Yohai [26] and may be viewed as a robustification of the classical leave-one-out criterion [see, e.g., 46] in which all the τ scale is replaced by the sum of squares of the r_{-i}. For the simulation experiments and real-data examples in this paper we have adopted a two-step approach to identify the minimizer of $\mathrm{RCV}(\lambda)$. First, we have determined the approximate location of the minimizer by evaluating $\mathrm{RCV}(\lambda)$ on a grid and then employed a numerical optimizer based on golden section search and parabolic interpolation [29] in the neighborhood of this approximate optimum. Such a hybrid approach is often advisable due to the possible local minima and near-flat regions of the CV criterion. Implementations and illustrative examples of the penalized MM-estimator may be found in `https://github.com/ioanniskalogridis/Robust-functional-linear-regression`.

3. Asymptotic properties

3.1. Consistency

We now study asymptotic properties of the penalized MM-estimators defined in Section 2. For notational convenience we assume that the variables are centred so that $\alpha_0 = 0$ and the object of interest is the coefficient function f_0. As is common for spline estimators in the functional linear model, see, e.g., Cardot et al. [4], Crambes et al. [6] and Boente et al. [2], we focus on the distance criterion given by

$$\pi(f, g) = [\mathbb{E}\{|\langle X, f - g\rangle|^2\}]^{1/2}, \quad (f, g) \in \mathcal{L}^2([0, 1]) \times \mathcal{L}^2([0, 1]),$$

which may be rewritten as $\pi(f, g) = \{\langle \Gamma(f - g), f - g\rangle\}^{1/2}$ with Γ denoting the self-adjoint Hilbert-Schmidt covariance operator of X. This criterion is directly linked to the average squared prediction error that arises when using $\langle X_{n+1}, \widehat{f_n}\rangle$ to predict $\langle X_{n+1}, f_0\rangle$, where X_{n+1} is a new random function possessing the same distribution as X.

For our theoretical development we require assumptions on the loss function, ρ, the error, ϵ, the functional predictor, X, and the coefficient function, f_0. For each K-dimensional approximating subspace Θ_K we define the element closest to f_0 as $\widetilde{f_K} = \mathrm{argmin}_{f \in \Theta_K} \|f_0 - f\|$. Since Θ_K is closed and convex, the Hilbert projection theorem ensures that $\widetilde{f_K}$ is a well-defined and unique element of Θ_K. Note that $\widetilde{f_K}$ is an abstract quantity to which we have no access in practice, but its existence and properties are essential for the results to follow. We require the following assumptions.

(A1) The loss function ρ satisfies $\rho(0) = 0$ and is even, non-decreasing on $[0, \infty)$, bounded and twice continuously differentiable with bounded derivatives ψ and ψ'. Furthermore, $\sup_{x\in\mathbb{R}} |x\psi(x)| < \infty$. Without loss of generality we assume that $\|\rho\|_\infty = 1$.

(A2) The scale estimate $\widehat{\sigma}_n$ satisfies $\widehat{\sigma}_n \xrightarrow{\mathbb{P}} \sigma_0$, where σ_0 is defined in (1).

(A3) The error ϵ is independent of X and possesses a Lebesgue-density $g_0(t)$ that is even, decreasing in $|t|$ and strictly decreasing in $|t|$ in a neighbourhood of zero. Furthermore, $\mathbb{E}\{\psi'(\epsilon)\} > 0$.

(A4) There exists a $C > 0$ such that $\mathbb{P}(\|X\| \leq C) = 1$ and for every $f \in \mathcal{L}^2([0, 1])$ and $\alpha \in \mathbb{R}$ such that $(f, \alpha) \neq (0, 0)$, $\mathbb{P}(\langle X, f\rangle) = \alpha) < 1$.

(A5) The coefficient function f_0 belongs to a Banach space of functions, $\mathcal{B}([0, 1])$, that is embeddable in $C([0, 1])$. Furthermore, the unit ball $\{f \in \mathcal{B}([0, 1]) : \|f\|_{\mathcal{B}} \leq 1\}$ is compact in the topology of the norm $\|\cdot\|_\infty$.

(A6) There exists a $c \in (0, 1)$ such that $\mathbb{P}(\langle X, f\rangle = 0) < c$ for any $f \in \mathcal{B}([0, 1])$ such that $f \neq 0$.

(A7) $\Theta_K \subset \mathcal{B}([0, 1])$ and the dimension K satisfies $K \asymp n^\beta$ for some $\beta \in (0, 1)$. Furthermore, $\|\widetilde{f}_K - f_0\| \to 0$ as $K \to \infty$ and $\lambda\mathcal{J}(\widetilde{f}_K) \xrightarrow{\mathbb{P}} 0$, as $n \to \infty$.

Assumptions (A1)–(A3) are standard for MM-estimators, see Yohai [50]. In combination with (A6) they imply that the estimators are Fisher-consistent so that at the population level we are indeed estimating the target function f_0, see Lemma 1 in the appendix. Assumption (A1) is satisfied, for example, by the Tukey bisquare loss. As shown by Boente et al. [2], the S-scale estimator satisfies (A2) under mild conditions. The first part of (A4) imposes the almost sure boundedness of the functional covariate when viewed as an element of $\mathcal{L}^2([0, 1])$. This assumption has been used extensively in the asymptotics of the functional linear regression model, see for example Cardot et al. [4], Zhao et al. [52] and Boente et al. [2]. The second part of (A4) ensures that X is not concentrated on any subspace of $\mathcal{L}^2([0, 1])$, which is the case whenever X possesses a Karhunen-Loève decomposition consisting of infinitely many non-zero terms [16, Chapter 7]. Equivalently, the null-space of its covariance operator Γ should only consist of the zero element.

Assumptions (A5) and (A7) are mild smoothness conditions on the coefficient function. For $\mathcal{B}([0, 1])$ to be embeddable in $C([0, 1])$ it suffices to have $\mathcal{B}([0, 1]) \subset C([0, 1])$ and a constant $c_0 > 0$ such that

$$\|f\|_\infty \leq c_0\|f\|_{\mathcal{B}}, \quad f \in \mathcal{B}([0, 1]). \tag{5}$$

Equivalently, the identity operator between these two spaces should be bounded. Furthermore, the unit ball in $\mathcal{B}([0, 1])$ should be compact, when merged with $C([0, 1])$. Both parts of assumption (A5) are satisfied by many interesting spaces of functions. Consider, for example, the Sobolev space $\mathcal{W}^{1,p}([0, 1])$ defined as

$$\mathcal{W}^{1,p}([0, 1]) = \left\{f : [0, 1] \to \mathbb{R} : \left\{\int_0^1 |f(x)|^p dx\right\}^{1/p} + \left\{\int_0^1 |f'(x)|^p dx\right\}^{1/p} < \infty\right\},$$

with $p > 1$. It can be shown that $\mathcal{W}^{1,p}([0, 1])$ is complete when endowed with the norm $\|f\|_{\mathcal{W}^{1,p}} = \|f\|_p + \|f'\|_p$. The mean-value theorem and Hölder's inequality may be employed to show that (5) holds, while the unit ball $\{f \in \mathcal{W}^{1,p}([0, 1]) : \|f\|_{W^{1,p}} \leq 1\}$ is compact in the sup-norm by virtue of the Arzelà-Ascoli theorem, as this set of functions is equicontinuous. These observations may be naturally extended to higher-order Sobolev spaces, see Adams and Fournier [1].

Finally, assumption (A7) states that f_0 may be arbitrarily well-approximated by an element of Θ_K in the $\mathcal{L}^2([0, 1])$-norm when $K \to \infty$. This approximating sequence $\widetilde{f}_K$ should have finite roughness, as measured by $\mathcal{J}(\cdot)$, so that $\lambda\mathcal{J}(\widetilde{f}_K) \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$. In many cases we have $\mathcal{J}(\widetilde{f}_K) = O(1)$, hence if $\lambda \xrightarrow{\mathbb{P}} 0$ the assumption is satisfied. It is important to note that in this work we treat λ as a random quantity and not merely as a deterministic sequence, as is often the case in literature [4, 42, 48]. In our opinion, this constitutes an important generalization, as in most cases λ is selected by a data-driven procedure and thus is random rather than fixed.

Comparing the above assumptions with the assumptions required in the non-parametric setting, see e.g., Kalogridis and Van Aelst [20, 21], reveals that in the more complex functional regression setting considered here our theoretical development requires considerably more sophisticated assumptions on the predictor X and the coefficient function f_0. Our first result extends Theorem 3.1 of Boente et al. [2] for the unpenalized B-spline estimator to the more general setting considered herein. It ensures that the penalized sieve estimators converge uniformly to the target coefficient function f_0. By (A4), uniform convergence also implies convergence with respect to prediction error.

Theorem 1. *Suppose that assumptions (A1)–(A7) hold. Furthermore, let*

$$M(f,\sigma) = \mathbb{E}\left\{\rho\left(\frac{Y-\langle X,f\rangle}{\sigma}\right)\right\},$$

and assume that $M(f_0,\sigma_0) = b < 1$ *and* $c < 1-b$ *in (A6). Then,* $\|\widehat{f_n} - f_0\|_{\mathcal{B}} = O_{\mathbb{P}}(1)$ *and* $\|\widehat{f_n} - f_0\|_\infty = o_{\mathbb{P}}(1)$*, as* $n \to \infty$.

The condition $M(f_0,\sigma_0) = b < 1$ required by Theorem 1 serves to avoid boundary solutions, in which $|\epsilon/\sigma_0|$ is so large that $\rho(\epsilon/\sigma_0) = 1$ almost surely (recall that ρ is even and $\|\rho\|_\infty = 1$). The second condition $c < 1-b$ parallels condition (A3) in Yohai [50] and may be viewed as a compatibility condition, see also Smucler and Yohai [43] for a similar use.

3.2. Rates of convergence

The result of Theorem 1 covers estimators based on many different basis systems and penalties, which all converge under suitable assumptions. To illustrate potential differences among estimators, we go one step further and investigate their respective rates of convergence in Theorem 2 below, which is based on the following development. We begin by defining the finite-sample version of $M(f,\sigma)$, that is,

$$M_n(f,\sigma) = \frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{Y_i-\langle X_i,f\rangle}{\sigma}\right).$$

Our objective function is $M_n(f,\widehat{\sigma}_n) + \lambda\mathcal{J}(f)$ and the minimization is over all $f \in \Theta_K$. By the projection theorem, $\widetilde{f_K} \in \Theta_K$ and therefore

$$M_n(\widehat{f_n},\widehat{\sigma}_n) + \lambda\mathcal{J}(\widehat{f_n}) \le M_n(\widetilde{f_K},\widehat{\sigma}_n) + \lambda\mathcal{J}(\widetilde{f_K}). \tag{6}$$

Adding $M(\widehat{f_n},\widehat{\sigma}_n) - M(\widetilde{f_K},\widehat{\sigma}_n)$ on both sides of (6), moving $M_n(\widehat{f_n},\widehat{\sigma}_n)$ to the right-hand side and noting that $\lambda\mathcal{J}(\widehat{f_n}) \ge 0$ yields

$$M(\widehat{f_n},\widehat{\sigma}_n) - M(\widetilde{f_K},\widehat{\sigma}_n) \le U_n(\widetilde{f_K},\widehat{f_n},\widehat{\sigma}_n) + \lambda J(\widetilde{f_K}), \tag{7}$$

where $U_n(f,g,\sigma) : \Theta_K \times \Theta_K \times \mathbb{R}_+ \to \mathbb{R}$ denotes the mean-centered process given by

$$U_n(f,g,\sigma) = \{M_n(f,\sigma) - M(f,\sigma)\} + \{M_n(g,\sigma) - M(g,\sigma)\}.$$

Now, under our assumptions it can be shown that there exist strictly positive constants η and L such that

$$M(\widehat{f_n},\widehat{\sigma}_n) - M(\widetilde{f_K},\widehat{\sigma}_n) \ge \eta|\pi(\widehat{f_n},\widetilde{f_K})|^2 - L\|\widetilde{f_K} - f_0\|\pi(\widehat{f_n},\widetilde{f_K}), \tag{8}$$

for all large n with high probability. The regularity of the process $U_n(f,g,\sigma)$ determines the asymptotic variance, cf. Lemma 3.2 in van de Geer [45]. In particular, we show that

$$U_n(\widetilde{f_K},\widehat{f_n},\widehat{\sigma}_n) = O_{\mathbb{P}}(1)\{\gamma_n\pi(\widehat{f_n},\widetilde{f_K}) \vee \gamma_n^2\}, \tag{9}$$

where $\gamma_n = K\log n/n$ and $a \vee b = \max(a,b)$. Rearranging, we obtain

$$\eta|\pi(\widehat{f_n},\widetilde{f_K})|^2 \le O_{\mathbb{P}}(1)\{\gamma_n\pi(\widehat{f_n},\widetilde{f_K}) \vee \gamma_n^2\} + O_{\mathbb{P}}(1)\|\widetilde{f_K} - f_0\|\pi(\widehat{f_n},\widetilde{f_K}) + \lambda\mathcal{J}(\widetilde{f_K}). \tag{10}$$

This inequality involving the square of $\pi(\widehat{f_n},\widetilde{f_K})$ in the left-hand side and $\pi(\widehat{f_n},\widetilde{f_K})$ in the right-hand side is key to Theorem 2 below, see the appendix for a detailed derivation.

Theorem 2. *Suppose that assumptions (A1)–(A7) hold, $M(f_0, \sigma_0) < b$ and $c < 1 - b$ in (A6). Then,*

$$|\pi(\widehat{f}_n, f_0)|^2 = O_{\mathbb{P}}\left(\frac{K \log n}{n}\right) + O_{\mathbb{P}}(\|\widetilde{f}_K - f_0\|^2) + O_{\mathbb{P}}(\lambda \mathcal{J}(\widetilde{f}_K)), \qquad \text{as } n \to \infty.$$

Theorem 2 presents the prediction error as a decomposition into three terms, which represent the variance, the modelling bias and the regularization bias respectively. The variance term depends only on the dimension of the sieve and not on its type. This situation has a well-known parallel in non-parametric regression, [see, e.g., 9, Chapter 15]. The $\log n$-term appearing in our decomposition is non-standard and results from slightly imprecise local entropy calculations [cf. 44, Chapter 9]. An intuitive explanation is that it reflects the difficulty of inference whenever the predictor variable is an infinite-dimensional object.

The second term in the decomposition of Theorem 2 is the bias stemming from the approximation of a generic $\mathcal{L}^2([0,1])$-function with a Θ_K-function. To ensure that this approximation error decreases fast as $K = K_n \to \infty$, we need to select a sieve that approximates well the class of functions to which f_0 belongs. Lastly, the penalization bias is reflected by the term $\lambda \mathcal{J}(f_K)$. This term suggests that to obtain high rates of convergence other than an appropriate basis system, one also needs an appropriate measure of roughness $\mathcal{J}(\cdot)$ on Θ_K. For example, given λ selecting the wavelet subspace of Example 3 combined with $J(f) = \|f^{(q)}\|^2$ would most likely lead to large values of $\mathcal{J}(\widetilde{f}_K)$ thereby diminishing the asymptotic performance of the estimator. Let us now revisit the previous examples and see how the prediction error behaves for some standard choices of $\mathcal{J}(\cdot)$.

Example 1 (Cont.)**.** Assume that f_0 has uniformly bounded derivatives up to order $r \geq 1$ with rth derivative satisfying a Lipschitz condition of order $v \in (0, 1)$. Note that this space of functions also satisfies (A5) under its usual norm. Then, for $p > r$ and equidistant interior knots we find $\|\widetilde{f}_K - f_0\| = O(K^{-r-v})$, [see 7, p.149]. At the same time, $\|\widetilde{f}^{(q)}\|^2 = O(1)$ for all $q < p$ [see, e.g., 20] leading to

$$|\pi(\widehat{f}_n, f_0)|^2 = O_{\mathbb{P}}\left(\frac{K \log n}{n}\right) + O_{\mathbb{P}}\left(\frac{1}{K^{2r+2v}}\right) + O_{\mathbb{P}}(\lambda).$$

For $K \asymp n^{1/\{2(r+v)+1\}}$ and $\lambda = O_{\mathbb{P}}(n^{-\gamma})$ with $\gamma \geq 2(r+v)/\{2(r+v)+1\}$ we obtain $|\pi(\widehat{f}_n, f_0)|^2 = O_P(n^{-2(r+v)/\{2(r+v)+1\}} \log n)$. This is a much higher rate of convergence than the $n^{-2(r+v)/\{4(r+v)+1\}}$-rate obtained by Cardot et al. [4] for the penalized least squares estimator, which is a consequence of our use of modern empirical process methodology to derive the result. Note that the rate of convergence obtained here is very different from the rate of convergence obtained in the context of robust nonparametric regression with penalized splines [see e.g., 21], due to the infinite-dimensional predictor process in functional regression. In fact, in the functional setting we are only able to reproduce the "small number of knots" asymptotic scenario.

Example 2 (Cont.)**.** Under similar assumptions on f_0 as in the previous example we have $\|\widetilde{f}_K - f_0\| = O(K^{-r-v})$, as seen from DeVore and Lorentz [8, Corollary 7.2.4]. At the same time [Theorem 7.2.7 of DeVore and Lorentz [8] implies $\|\widetilde{f}^{(q)}\|^2 = O(1)$ for $q \leq r$, whence

$$|\pi(\widehat{f}_n, f_0)|^2 = O_{\mathbb{P}}\left(\frac{K \log n}{n}\right) + O_{\mathbb{P}}\left(\frac{1}{K^{2r+2v}}\right) + O_{\mathbb{P}}(\lambda).$$

For similar choices of K and λ as in the spline setting, we are again led to $|\pi(\widehat{f}_n, f_0)|^2 = O_P(n^{-2(r+v)/\{2(r+v)+1\}} \log n)$. The same conclusion holds for the harmonic acceleration penalty, provided that $r \geq 3$. The fact that many different sieves yield exactly the same rate of convergence for smooth functions is well-known in classical nonparametric regression [41].

Example 3 (Cont.)**.** For a demonstration of a different flavour consider the Sobolev space $\mathcal{W}^{m,2}([0,1])$, which satisfies (A5) for any $m \geq 1$. If $f_0 \in \mathcal{W}^{m,2}([0,1])$, under the assumptions of Zhao et al. [52] we find $\|\widetilde{f}_K - f_0\| = O(K^{-m})$ and for the ℓ_1-penalty on the wavelet coefficients we have $\lambda \mathcal{J}(\widetilde{f}_K) = \lambda^{1-r/2}$ for some $r \in (0, 2)$ leading to

$$|\pi(\widehat{f}_n, f_0)|^2 = O_{\mathbb{P}}\left(\frac{K \log n}{n}\right) + O_{\mathbb{P}}\left(\frac{1}{K^{2m}}\right) + O_{\mathbb{P}}\left(\lambda^{1-r/2}\right).$$

The regularization bias is now different from the two previous examples because of the thresholded wavelet coefficients, see Zhao et al. [52] for more details.

3.3. Generalization to higher dimensions

For clarity, we focused on the case of $\mathcal{I} = [0,1]$, i.e. a stochastic process $X : [0,1] \times \Omega \to \mathbb{R}$. However, our estimation method and theoretical development permit important extensions to random fields. In particular, let $\mathcal{I}$ now denote a subset of $\mathbb{R}^d$ with $d \geq 1$ and consider the multidimensional extension of (1) given by

$$Y = \alpha_0 + \int_{\mathcal{I}} X(\mathbf{t}) f_0(\mathbf{t}) d\mathbf{t} + \sigma_0 \epsilon,$$

for some $f_0 \in \mathcal{L}^2(\mathcal{I})$.

For $\mathcal{I} = [0,1]^d$ an approximating subspace may be easily constructed by taking tensor products of univariate approximating subspaces, that is, we consider the subspace $\Theta_K = \Theta_{K_1} \otimes \ldots \otimes \Theta_{K_d}$. A multivariate penalized MM-estimator may now be defined as follows

$$(\widehat{\alpha}_n, \widehat{f}_n) = \underset{\alpha \in \mathbb{R}, f \in \Theta_{\mathbf{K}}}{\operatorname{argmin}} \left[\frac{1}{n} \sum_{i=1}^{n} \rho\left(\frac{Y_i - a - \langle X_i, f \rangle}{\widehat{\sigma}_n} \right) + \mathcal{J}_{\boldsymbol{\lambda}}(f) \right], \tag{11}$$

where for $f, g \in \mathcal{L}^2(\mathcal{I})$, $\langle f, g \rangle = \int_{\mathcal{I}} f(\mathbf{t}) g(\mathbf{t}) d\mathbf{t}$ and $\mathcal{J}_{\boldsymbol{\lambda}} : \Theta_{\mathbf{K}} \to \mathbb{R}_+$ is an appropriate penalty functional that depends on a vector of smoothing parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)$.

Inspection of the proofs of Theorems 1 and 2 above reveals that the these theorems carry over to multivariate MM-estimators without additional difficulty, under straightforward adaptations of assumptions (A4), (A5), (A6) and (A7). A uniform law of large numbers as in Lemma 2 of the Appendix (with $\Theta_{\mathbf{K}}$ replacing Θ_K) can then be derived, provided that $\prod_{j=1}^{d} K_j = o(n)$. Combined with the adapted assumptions, this law allows to show that Theorem 1 remains valid in the multivariate setting, i.e., $\sup_{\mathbf{x} \in \mathcal{I}} |\widehat{f}_n(\mathbf{x}) - f_0(\mathbf{x})| \xrightarrow{\mathbb{P}} 0$. Furthermore, the bracketing integral of the class of functions of $(X, y) \in \mathcal{L}^2(\mathcal{I}) \times \mathbb{R}$ given by

$$\mathcal{G}_{n,c,\delta} = \left\{ \rho\left(\frac{y - \langle X, f \rangle}{\sigma} \right) - \rho\left(\frac{y - \langle X, \widetilde{f}_{\mathbf{K}} \rangle}{\sigma} \right),\ f \in \Theta_{\mathbf{K}},\ \|f - \widetilde{f}_{\mathbf{K}}\| \leq c,\ |\sigma - \sigma_0| \leq \delta \right\}$$

from 0 to every small $\epsilon > 0$ behaves like $C_0 \epsilon \log^{1/2}\left(\frac{1}{\epsilon}\right) \prod_{j=1}^{d} K_j$ for some constant C_0. Therefore, the argument in the proof of Theorem 2 yields

$$|\pi(\widehat{f}_n, f_0)|^2 = O_{\mathbb{P}}\left(\frac{\log n \prod_{j=1}^{d} K_j}{n} \right) + O_{\mathbb{P}}(\|\widetilde{f}_{\mathbf{K}} - f_0\|^2) + O_{\mathbb{P}}(\mathcal{J}_{\boldsymbol{\lambda}}(\widetilde{f}_{\mathbf{K}})).$$

The inflation of the variance term is a manifestation of the curse of dimensionality and translates into comparatively lower rates of convergence for large d. Appropriate roughness penalties on $\Theta_{\mathbf{K}}$ are thin-plate and tensor product penalties, see [47, Chapter 5] for more details.

4. A Monte-Carlo study

In our simulation scenarios we examine the effects of the shape of the true coefficient function and outlying observations on four functional regression estimators. The first estimator we consider is the proposed penalized MM-estimator, abbreviated with PMM, based on a spline subspace with penalty and settings described in Section 2. We compare this estimator to the following estimators.

- The least-squares penalized spline variant of the proposed estimator abbreviated as PLS.
- The robust $\mathcal{W}^{2,2}([0,1])$-estimator of Shin and Lee [42] abbreviated as RKHS.
- The robust unpenalized cubic B-spline estimator of Boente et al. [2] abbreviated as MM.
- The robust smoothing spline type estimator of Maronna and Yohai [26] abbreviated as MMSS.

For all four robust estimators, PMM, RKHS, MM and MMSS, we used the Tukey-bisquare loss with tuning constant equal to 4.685, corresponding to 95% efficiency in the location model under Gaussian errors. The unpenalized B-spline estimator of Boente et al. [2] is based on equidistant knots whose number is selected through a robust BIC-type criterion as proposed by Boente et al. [2]. The penalty parameter for RKHS is selected through a robust generalized cross-validation criterion. To compute the MMSS estimator the optimization problem is first reduced to an MM ridge regression as outlined in Maronna and Yohai [26], which is then solved by the pensem_cv function of the R-package pense [5]. The penalty parameter is selected through the robust cross-validation procedure using the default candidate values of the R function.

In order to compare the five competing estimators we have generated curves according to the truncated Karhunen-Loève decomposition given by

$$X(t) = \sum_{j=1}^{50} j^{-1} Z_j \sqrt{2} \cos((j-1)\pi t), \tag{12}$$

where the Z_j are random variables whose distribution is varied according to the scenarios outlined below. These curves are combined with each of the following four coefficient functions:

1. $f_1(t) = \sin(2\pi t)$
2. $f_2(t) = t^2 \phi(t, 0, 0.1)$
3. $f_3(t) = 1/(1 + e^{-20(t-0.5)})$
4. $f_4(t) = -\phi(t, 0.2, 0.03) + 3\phi(t, 0.5, 0.03) + \phi(t, 0.75, 0.04)$,

where $\phi(t, \mu, \sigma)$ denotes the Gaussian density with mean μ and standard deviation σ. These regression functions represent a variety of different characteristics: f_1 is a sinusoid, f_2 is almost a straight line with some curvature near the boundaries, f_3 is a sigmoid and f_4 is bumpy. Due to its local characteristics, f_4 is much more difficult to estimate precisely than the other functions.

We set $\sigma_0 = 1$ in (1) and consider the following scenarios for the scores Z_j in (12) and errors ϵ_i in (1):

Scen. 1: The Z_j and ϵ_i both follow standard Gaussian distributions.

Scen. 2: The Z_j follow a standard Gaussian distribution and the ϵ_i follow a t_3-distribution.

Scen. 3: The Z_j follow standard Gaussian distributions and the ϵ_i follow a Gaussian mixture distribution with density $0.9\phi(t, 0, 1) + 0.1\phi(t, 14, 1)$.

Scen. 4: The Z_j follow a t_3-distribution and the ϵ_i follow the same Gaussian mixture distribution as in the previous scenario.

These scenarios reflect increasingly severe contamination. The first scenario portrays the ideal situation of light-tailed predictors and error, the second scenario introduces mild vertical outliers and the third scenario yields more severe contamination. Lastly, by perturbing the distribution of the Z_j, the fourth scenario combines vertical outliers and leverage points. For a better appreciation of the effect of the distribution of the Z_j on the shape of the curves, Fig. 1 plots two representative samples of curves with the Z_j following Gaussian and t_3 distributions.

To handle the curves practically we have discretized them in 100 equidistant points $t_1, \ldots, t_{100}$, within the $[0, 1]$-interval and computed all related inner products using Riemann approximations. To evaluate the performance of the estimators we calculate their mean-square error (MSE) given by

$$\text{MSE} = 100^{-1} \sum_{j=1}^{100} |\widehat{f}(t_j) - f_0(t_j)|^2.$$

This statistic is an approximation to the squared $\mathcal{L}^2([0, 1])$-distance $\|\widehat{f} - f_0\|^2$. Table 1 below presents average and median MSEs for all of our settings for $n = 150$ and 1000 replications (500 for MMSS).

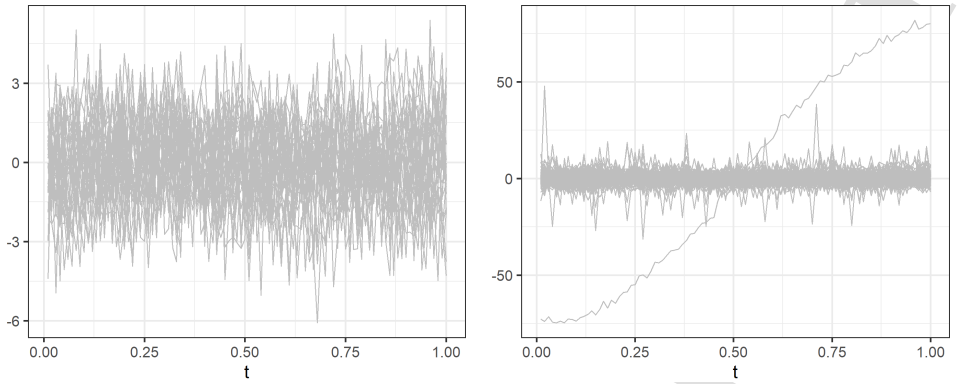

Fig. 1: Two representative samples of simulated curves with $Z_j \sim \mathcal{N}(0,1)$ (Scen. 3) and $Z_j \sim t_3$ (Scen. 4) on the left and right respectively.

There are several interesting conclusions that emerge from this study. First, the performance of the least-squares estimator PLS heavily depends on the distribution of the scores and errors. The estimator performs best under light-tailed distributions, but quickly loses ground when faced with slightly heavier tails, e.g., with a t_3-distribution in the errors. This performance is in line with expectations regarding least-squares estimators which are known to be very sensitive to even a small number of mildly outlying observations. In contrast, the robust estimators MM and PMM maintain a much steadier performance. Comparing the performance of the robust estimators in more detail reveals that MM and PMM outperform RKHS and MMSS in all settings except with respect to f_2, where MMSS performs up to twice as well as MM and PMM. However, in the other settings MMSS and RKHS are outperformed by MM and PMM, often by a large margin.

The estimators MM and PMM perform comparably with respect to f_1, f_2 and f_3, but MM performs poorly with respect to f_4. The reason that MM performs well for the first three coefficient functions but not for f_4 is that f_1, f_2 and f_3 are simple enough to allow for effective approximation with a spline basis defined on a handful of equidistant interior knots, whereas f_4 possesses local characteristics that make such an approximation difficult. To illustrate the key differences between MM and PMM with respect to f_4, Fig. 2 presents the 1000 estimates for f_4 obtained under Scenario 1 along with the true coefficient function in solid black. From Fig. 2 it may be seen that PMM correctly identifies the bumps of f_4 whereas MM produces more variable estimates that often lack the correct amount of smoothness. The performance of MM could be improved by a more careful selection of the location of the knots, but this would inevitably lead to a much increased computational burden. Overall, these simulation results indicate that PMM is a viable alternative to PLS in clean data and remains reliable in a wider range of contaminated data settings than its unpenalized alternative MM.

5. Real data example: archaeological glass vessels

In this section we apply the proposed penalized estimator to the popular glass dataset. This dataset contains measurements for 180 archaeological glass vessels (15th to 17th century) that were excavated from the old city of Antwerp, which prior to the tumultuous 17th century was one of the largest ports in Europe with extensive ties to commercial centres all over the continent, see [18] for more background. The dataset is freely available in R-packages `chemometrics` [12] and `cellWise` [35].

For each of the vessels we are in possession of near-infrared spectra with 750 wavelengths, along with the values of 13 chemical compounds which are crucial for the determination of the type of glass as well as its origin. A reduced form of this dataset with only the non-null spectra was analyzed by Maronna and Yohai [26]. However, here we avoid any preprocessing of the data. A plot of the spectra and the histogram of one of the chemical compounds are given in

Table 1: Mean and median of the MSE (×1000) for the competing estimators over 1000 datasets of size $n = 150$. Best median performances are in bold. PMM is the penalized MM-estimator proposed in Section 2. The other estimators are outlined at the beginning of Section 4.

		PLS		RKHS		MM		MMSS		PMM	
		Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median
f_1	Scen. 1	0.152	**0.125**	0.537	0.502	0.183	0.159	1.886	0.723	0.164	0.131
	Scen. 2	0.399	0.301	0.947	0.690	0.235	**0.201**	1.878	1.133	0.254	0.210
	Scen. 3	2.282	1.598	1.353	1.211	0.191	**0.165**	16.72	0.935	0.252	0.184
	Scen. 4	1.624	1.103	1.348	1.135	0.151	0.128	14.94	0.445	0.178	**0.121**
f_2	Scen. 1	0.103	0.073	0.082	0.075	0.143	0.098	0.078	**0.042**	0.109	0.075
	Scen. 2	0.205	0.118	0.107	0.079	0.166	0.119	0.079	**0.045**	0.138	0.090
	Scen. 3	1.093	0.528	0.126	0.084	0.119	0.085	0.104	**0.042**	0.134	0.074
	Scen. 4	0.454	0.224	0.118	0.057	0.058	0.051	0.041	**0.025**	0.063	0.048
f_3	Scen. 1	0.195	**0.167**	0.378	0.356	0.260	0.241	27.52	27.37	0.217	0.178
	Scen. 2	0.478	0.359	0.973	0.538	0.368	0.318	28.90	28.60	0.320	**0.267**
	Scen. 3	2.386	1.833	17.69	25.39	0.281	0.252	37.70	36.35	0.295	**0.231**
	Scen. 4	0.996	0.762	26.08	25.76	0.105	**0.087**	36.34	35.17	0.123	0.094
f_4	Scen. 1	10.78	**10.69**	89753	89322	9023	8976	25037	25088	11.02	10.90
	Scen. 2	13.29	12.78	90015	88921	9033	8986	24834	24892	12.42	**12.32**
	Scen. 3	32.26	31.25	89934	88875	9025	8981	24587	24627	11.76	**11.45**
	Scen. 4	18.31	17.88	88326	84843	9054	8995	28750	28541	12.45	**11.48**

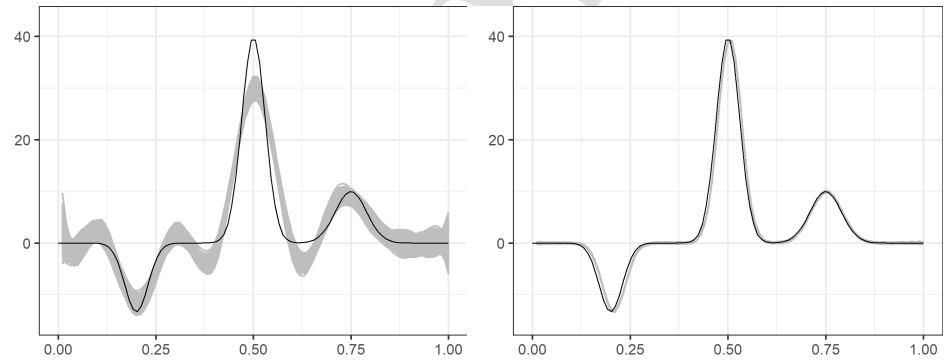

Fig. 2: The unpenalized MM-estimates $\widehat{f}_{MM}$ proposed by Boente et al. [2] (left) and the penalized MM-estimates $\widehat{f}_{PMM}$ proposed in Section 2 (right) for f_4 under Scenario 1. The solid black line in each plot corresponds to the true coefficient function.

Fig. 3. By examining the heights of the peaks in the spectra it may be conjectured that there are three types of glass in the sample, which is indeed the key finding of Janssens et al. [18]. The histogram of the chemical compounds further indicates that the distributions of these responses are right-skewed with several potential outliers.

As the overall best performing estimators in clean and contaminated data respectively, we compare the predictive performance of PLS and PMM for each of the 13 responses in this dataset. To measure the prediction performance of the methods, we apply 5-fold cross-validation. For each chemical compound we then compute the 10% trimmed

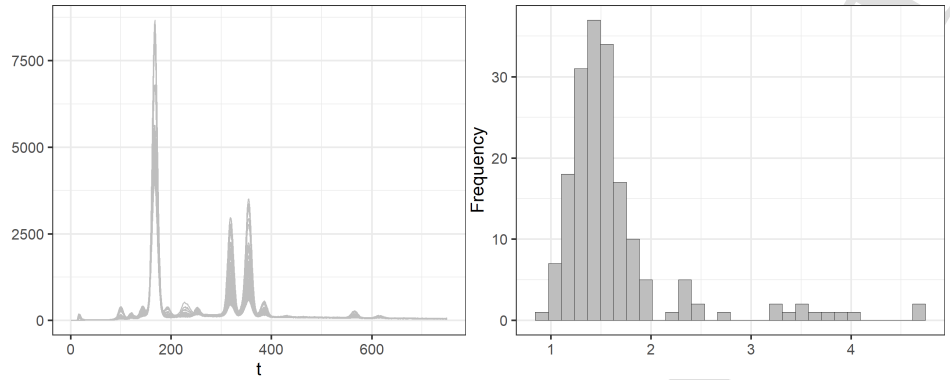

Fig. 3: Visualization of the glass dataset [18] containing measurements for 180 archaeological glass vessels and 13 associated chemical compounds. The left plot contains the near-infrared spectra with 750 wavelengths while the right plot shows a histogram of the chemical compound Al2O3.

Table 2: Prediction performance of penalized functional regression least squares estimates (PLS) and MM-estimates (PMM) proposed in Section 2 for each of the 13 chemical compounds in the glass dataset. Performance is measured by the 10% trimmed root mean squared error of the predictions obtained by 5-fold cross-validation, denoted by RMSPE(0.9). Best performances are in bold.

	Na_2O	MgO	Al_2O_3	SiO_2	P_2O_5	SO_3	Cl	K_2O	CaO	MnO	Fe_2O_3	BaO	PbO
PLS	0.587	0.142	0.069	0.480	**0.042**	**0.039**	0.014	0.147	0.186	0.019	**0.013**	0.018	**0.089**
PMM	**0.513**	**0.138**	**0.064**	**0.458**	0.043	0.040	**0.013**	**0.129**	**0.176**	**0.018**	0.014	**0.017**	0.093

root mean squared error of the predictions, denoted by RMSPE(0.9). This trimming is essential to measure prediction performance of the regular data because some of the left-out observations can be outliers [19]. To reduce variability we repeat the procedure 30 times and report the average RMSPE(0.9) for each compound.

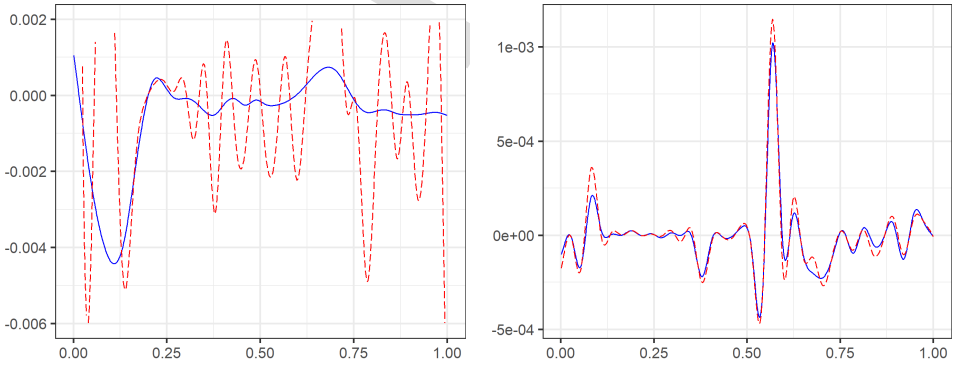

Fig. 4: Penalized functional regression least squares estimates (PLS) and MM-estimates (PMM) proposed in Section 2 for the chemical compounds SiO_2 and BaO in the glass dataset. The lines (——, – –) correspond to PMM and PLS estimates, respectively.

The results are summarized in Table 2. It can be seen that PMM outperforms PLS in 9 out of the 13 components.

In practice, this means that in these cases PMM provides better fits for the majority of the observations, thereby leading to better predictions of the observations following the model. The estimates of the coefficient functions for SiO_2 and BaO are shown in Fig. 4. It is interesting to observe that for these two components PLS produces more wiggly estimates than PMM, but these more complex estimates do not translate into better predictive performance. This suggests that the smoothness of PLS is influenced by outlying observations leading to worse predictive ability.

6. Concluding remarks

We have shown that lower-rank penalized estimators based on bounded loss functions possess good theoretical properties, are computationally efficient and are capable of handling a diversity of complex problems, such as estimation of coefficient functions with local characteristics based on data with atypical observations. Moreover, these important properties almost seamlessly extend to the case of scalar-on-function regression with random functions defined on multidimensional sets, such as images. To select the penalty parameter we proposed to minimize a robust cross-validation criterion. Alternatively, a slightly larger λ may be selected, e.g., by using the one standard error rule as in Maronna and Yohai [26] thereby increasing the smoothness of the fit.

In future work we aim to further relax the assumptions underpinning the present theoretical development to take into account the often discrete sampling of functional data. This often neglected aspect of functional data has important practical and theoretical consequences, particularly when the discretization grid is small, see, e.g., Kalogridis and Van Aelst [22] for the case of location estimation. A robust lower rank penalized regression estimator in this setting would constitute an effective and computationally efficient alternative to the smoothing spline estimators of Crambes et al. [6] and Maronna and Yohai [26].

7. Acknowledgements

The authors are grateful to two anonymous reviewers and the editor-in-chief, Professor Dietrich von Rosen, whose remarks greatly improved the paper both with respect to accessibility and content. I. Kalogridis gratefully acknowledges support by grant C16/15/068 of Internal Funds KU Leuven and by the Research Foundation-Flanders (project 1221122N).

8. Appendix: Proofs of the theorems

Lemma 1 (Fisher consistency)**.** *Assume that assumptions (A1), (A3) and (A6) hold. Then, for any* $\sigma > 0$ *and* $f \in \mathcal{B}([0,1])$*,* $M(f_0, \sigma) < M(f, \sigma)$*, where*

$$M(f,\sigma) = \mathbb{E}\left\{\rho\left(\frac{Y - \langle X, f\rangle}{\sigma}\right)\right\}.$$

Proof. The proof is an adaptation of the corresponding proofs in Yohai [50] and Boente et al. [2]. First, Lemma 3.1 of Yohai [49] in combination with (A3) shows that the function $g(\alpha) = \mathbb{E}\{\rho(\epsilon\sigma_0/\sigma - \alpha)\}$ has a unique minimum at zero, viz, for any $\alpha \neq 0$

$$\mathbb{E}\left\{\rho\left(\epsilon\frac{\sigma_0}{\sigma} - \alpha\right)\right\} > \mathbb{E}\left\{\rho\left(\epsilon\frac{\sigma_0}{\sigma}\right)\right\}. \tag{13}$$

Fix $f \in \mathcal{B}([0,1])$, set $\mathcal{A}_0 = \{X : \Phi(X) = \langle X, f - f_0\rangle = 0\}$ and $\alpha(X) = \Phi(X)/\sigma$. Then, using the independence of ϵ and X, it is not difficult to show that

$$M(f,\sigma) = \mathbb{E}\left\{\rho\left(\epsilon\frac{\sigma_0}{\sigma}\right)\right\}\mathbb{P}(\mathcal{A}_0) + \mathbb{E}\left\{\mathbb{E}\left\{\rho\left(\epsilon\frac{\sigma_0}{\sigma} - \alpha(X)\right)|X\right\}\mathcal{I}_{\mathcal{A}_0^c}(X)\right\},$$

where $\mathcal{I}_B(\cdot)$ denotes the indicator function for a set B. Similarly, by the independence of ϵ and X and (13), for any $X \notin \mathcal{A}_0$,

$$\mathbb{E}\left\{\rho\left(\epsilon\frac{\sigma_0}{\sigma} - \alpha(X)\right)|X = X_0\right\} = \mathbb{E}\left\{\rho\left(\epsilon\frac{\sigma_0}{\sigma} - \alpha(X_0)\right)\right\} > \mathbb{E}\left\{\rho\left(\epsilon\frac{\sigma_0}{\sigma}\right)\right\}.$$

Hence, since, by (A6), $\mathbb{P}(\mathcal{A}_0) > 0$, we find

$$M(f,\sigma) > \mathbb{E}\left\{\rho\left(\epsilon\frac{\sigma_0}{\sigma}\right)\right\}\mathbb{P}(\mathcal{A}_0) + \mathbb{E}\left\{\mathbb{E}\left\{\rho\left(\epsilon\frac{\sigma_0}{\sigma}\right)\right\}\mathcal{I}_{\mathcal{A}_0^c}(X)\right\} = \mathbb{E}\left\{\rho\left(\epsilon\frac{\sigma_0}{\sigma}\right)\right\}\mathbb{P}(\mathcal{A}_0) + \mathbb{E}\left\{\rho\left(\epsilon\frac{\sigma_0}{\sigma}\right)\right\}\mathbb{P}(\mathcal{A}_0^c) = M(f_0,\sigma),$$

where the strict inequality follows from the fact that $\mathcal{A}_0$ has strictly positive probability. □

Lemma 2. *Let ρ satisfy (A1) and K satisfy (A7). Then the following uniform law holds*

$$\sup_{\sigma>0, f\in\Theta_K} |M_n(f,\sigma) - M(f,\sigma)| \xrightarrow{a.s.} 0, \qquad as\ n\to\infty.$$

Proof. The proof may be deduced from the proof of Lemma A.1.2 in Boente et al. [2], we omit the details. □

Lemma 3. *Suppose that assumptions (A1), (A2), (A4) and (A7) hold. Then, $M(\widehat{f_n},\sigma_0) \xrightarrow{\mathbb{P}} M(f_0,\sigma_0)$, as $n\to\infty$.*

Proof. By Lemma 1, f_0 is the unique minimizer of $M(f,\sigma_0)$ over all $f\in\mathcal{B}([0,1])$ and, by (A7), $\widehat{f_n}\in\mathcal{B}([0,1])$. Therefore,

$$0 \le M(\widehat{f_n},\sigma_0) - M(f_0,\sigma_0) = I + II + III, \tag{14}$$

with

$$I = M(\widehat{f_n},\sigma_0) - M_n(\widehat{f_n},\sigma_0),\ II = M_n(\widehat{f_n},\sigma_0) - M_n(\widehat{f_n},\widehat{\sigma}_n),\ III = M_n(\widehat{f_n},\widehat{\sigma}_n) - M(f_0,\sigma_0).$$

Since $\widehat{f_n}\in\Theta_K$ by definition of the estimator, Lemma 2 yields

$$|I| \le \sup_{\sigma>0, f\in\Theta_K} |M(f,\sigma) - M_n(f,\sigma)| \xrightarrow{a.s.} 0. \tag{15}$$

Moreover, a first order Taylor expansion immediately yields

$$|II| \le \sup_{x\in\mathbb{R}} |x\psi(x)| \frac{|\widehat{\sigma}_n - \sigma_0|}{\widehat{\xi}_n}, \tag{16}$$

where $\widehat{\xi}_n$ is an intermediate value in the linear segment joining $\widehat{\sigma}_n$ and σ_0. By (A2), $\widehat{\xi}_n > 0$ for all large n with high probability and $|\sigma_n - \sigma_0| \xrightarrow{\mathbb{P}} 0$ as $n\to\infty$ leading to $|II| \xrightarrow{\mathbb{P}} 0$.

To complete the proof we now treat III. Note that $\widehat{f_n}$ minimizes $M_n(f,\widehat{\sigma}_n) + \lambda\mathcal{J}(f)$ over all $f\in\Theta_K$ and, by construction, $\widetilde{f_K}\in\Theta_K$. Therefore, by (A7)

$$M_n(\widehat{f_n},\widehat{\sigma}_n) \le M_n(\widehat{f_n},\widehat{\sigma}_n) + \lambda\mathcal{J}(\widehat{f_n}) \le M_n(\widetilde{f_K},\widehat{\sigma}_n) + \lambda\mathcal{J}(\widetilde{f_K}) = M_n(\widetilde{f_K},\widehat{\sigma}_n) + o_{\mathbb{P}}(1).$$

Hence,

$$III \le M_n(\widetilde{f_K},\widehat{\sigma}_n) - M(f_0,\sigma_0) + o_{\mathbb{P}}(1) = \{M_n(\widetilde{f_K},\widehat{\sigma}_n) - M_n(f_0,\sigma_0)\} + \{M_n(f_0,\sigma_0) - M(f_0,\sigma_0)\} + o_{\mathbb{P}}(1).$$

By the law of large numbers, $M_n(f_0,\sigma_0) \xrightarrow{\mathbb{P}} M(f_0,\sigma_0)$. At the same time, by assumption (A4) and the Schwarz inequality,

$$\begin{aligned}|M_n(\widetilde{f_K},\widehat{\sigma}_n) - M_n(f_0,\sigma_0)| &\le |M_n(\widetilde{f_K},\widehat{\sigma}_n) - M_n(\widetilde{f_K},\sigma_0)| + |M_n(\widetilde{f_K},\sigma_0) - M_n(f_0,\sigma_0)| \\ &\le \sup_{x\in\mathbb{R}} |x\psi(x)| \frac{|\widehat{\sigma}_n - \sigma_0|}{\widehat{\xi}_n} + \frac{C\|\psi\|_\infty}{\sigma_0}\|\widetilde{f_K} - f_0\|,\end{aligned}$$

with probability one, where $\widehat{\xi}_n$ is an intermediate point. By assumptions (A2) and (A7) both terms tend to zero in probability, hence $III \le o_{\mathbb{P}}(1)$ which in combination with (15) and (16) completes the proof. □

Proof of Theorem 1. The first part of the Theorem follows from a simple adaptation of Lemma A.1.4 in Boente et al. [2], because by assumption (A5), $\{f \in \mathcal{B}([0,1]) : \|f\|_{\mathcal{B}} \leq 1\}$ is compact in the $\|\cdot\|_\infty$-topology.

The first result of the theorem implies that for every $\epsilon > 0$ there exists a $L = L_\epsilon$ such that $\mathbb{P}(\|\widehat{f_n} - f_0\|_{\mathcal{B}} > L) < \epsilon$ for all large n. Thus, it suffices to restrict attention to the set $\{\|\widehat{f_n} - f_0\|_{\mathcal{B}} \leq L\}$. To prove uniform convergence it suffices to show that

$$\inf_{f \in \mathcal{B}([0,1]):\|f-f_0\|_{\mathcal{B}} \leq L, \|f-f_0\|_\infty \geq \epsilon} M(f, \sigma_0) > M(f_0, \sigma_0). \tag{17}$$

This is sufficient, because by Lemma 3, $M(\widehat{f_n}, \sigma_0) \xrightarrow{\mathbb{P}} M(f_0, \sigma_0)$. Hence, (17) would imply that with high probability $\|\widehat{f_n} - f_0\|_\infty < \epsilon$ for all large n. To establish (17), let $\{f_k\}_k$ denote a minimizing sequence, i.e., a sequence satisfying $\|f_k - f_0\|_{\mathcal{B}} \leq L$, $\|f_k - f_0\|_\infty \geq \epsilon$ and

$$\lim_{k\to\infty} M(f_k, \sigma_0) = \inf_{f \in \mathcal{B}([0,1]):\|f-f_0\|_{\mathcal{B}} \leq L, \|f-f_0\|_\infty \geq \epsilon} M(f, \sigma_0).$$

Such a sequence exists, because ρ is nonnegative and therefore the infimum is bounded from below by 0. By compactness in assumption (A5), there exists a subsequence $f_{k_j} - f_0$ which converges uniformly to a function $f^\star \in C([0,1])$. By continuity of the norm, this implies that $\lim_{j\to\infty} \|f_{k_j} - f_0\|_\infty = \|f^\star\|_\infty$ and since $\|f_k - f_0\|_\infty \geq \epsilon$ for all $k \in \mathbb{N}$, we must also have $\|f^\star\|_\infty \geq \epsilon$. By the bounded convergence theorem it now follows that

$$\inf_{f \in \mathcal{B}([0,1]):\|f-f_0\|_{\mathcal{B}} \leq L, \|f-f_0\|_\infty \geq \epsilon} M(f, \sigma_0) = \lim_{j\to\infty} M(f_{k_j}, \sigma_0) = M(f_0 + f^\star, \sigma_0).$$

Moreover, since, by (5),

$$\|f_0 + f^\star - f_0\|_{\mathcal{B}} = \|f^\star\|_{\mathcal{B}} \geq c_0^{-1}\|f^\star\|_\infty \geq c_0^{-1}\epsilon > 0,$$

where c_0 is the embedding constant, Lemma 1 yields $M(f_0 + f^\star, \sigma_0) > M(f_0, \sigma_0)$ which completes the proof. □

We now introduce some useful notation. Let $\mathcal{G}$ denote a class of real-valued functions on $\mathcal{L}^2([0,1]) \times \mathbb{R}$. For $g \in \mathcal{G}$ we define

$$\|g\|_\infty = \sup_{x \in \mathcal{L}^2([0,1]), y \in \mathbb{R}} |g(x, y)|.$$

The covering number in this uniform metric, $\mathcal{N}_\infty(\epsilon, \mathcal{G})$, is defined as the smallest value of $N \in \mathbb{N}$ such that there exists a sequence $\{g_j\}_{j=1}^N$ with the property that

$$\sup_{g \in \mathcal{G}} \min_{j=1,\ldots,N} \|g - g_j\|_\infty \leq \epsilon.$$

The corresponding entropy, $\mathcal{H}_\infty(\epsilon, \mathcal{G})$, is the logarithm of the covering number, i.e., $\mathcal{H}_\infty(\epsilon, \mathcal{G}) = \log \mathcal{N}_\infty(\epsilon, \mathcal{G})$.

For a probability measure $\mathbb{P}$ we also define the bracketing number in the $\mathcal{L}^2(\mathbb{P})$-metric, $\mathcal{N}_B(\epsilon, \mathcal{G}, \mathbb{P})$, as the smallest value of $N \in \mathbb{N}$ for which there exist N pairs of functions $\{[g_j^L, g_j^U]\}$ such that $\|g_j^U - g_j^L\|_{\mathcal{L}^2(\mathbb{P})} \leq \epsilon$ for all $j = 1, \ldots, N$, and such that for each $g \in \mathcal{G}$, there is a $j = j(g) \in \{1, \ldots, N\}$ such that

$$g_j^L(x, y) \leq g(x, y) \leq g_j^U(x, y).$$

The corresponding bracketing entropy, $\mathcal{H}_B(\epsilon, \mathcal{G}, \mathbb{P})$, is defined as the logarithm of the bracketing number, i.e., $\mathcal{H}_B(\epsilon, \mathcal{G}, \mathbb{P}) = \log \mathcal{N}_B(\epsilon, \mathcal{G}, \mathbb{P})$.

Lemma 4 (Bracketing entropy)**.** *Suppose that (A1) and (A4) hold. For* $(x, y) \in \mathcal{L}^2([0,1]) \times \mathbb{R}$ *define*

$$g_{f,\sigma}(x, y) = \rho\left(\frac{y - \langle x, f\rangle}{\sigma}\right) - \rho\left(\frac{y - \langle x, \widetilde{f_K}\rangle}{\sigma}\right)$$

and the class of functions

$$\mathcal{G}_{n,c,\delta} = \left\{ g_{f,\sigma}(x,y),\ f \in \Theta_K,\ \|f - \widetilde{f}_K\| \leq c,\ |\sigma - \sigma_0| \leq \delta \right\}.$$

Let $\mathbb{P}$ denote the probability measured induced by (X, y). Then, there exists a constant $A > 0$ depending only on c and δ such that

$$H_B(\epsilon, \mathcal{G}_{n,c,\delta}, \mathbb{P}) \leq 3K \log\left(1 + \frac{A}{\epsilon}\right).$$

Proof. Let us begin by observing that, by Lemma 2.1 of van de Geer [44], we have

$$\mathcal{H}_B(\epsilon, \mathcal{G}_{n,c,\delta}, \mathbb{P}) \leq \mathcal{H}_\infty(\epsilon/2, \mathcal{G}_{n,c,\delta}),$$

so that it suffices to bound the covering number of $\mathcal{G}_{n,c,\delta}$ in the uniform metric. Applying the triangle inequality twice now yields

$$\begin{aligned} |g_{f_1,\sigma_1}(x,y) - g_{f_2,\sigma_2}(x,y)| &\leq |g_{f_1,\sigma_1}(x,y) - g_{f_1,\sigma_2}(x,y)| + |g_{f_1,\sigma_2}(x,y) - g_{f_2,\sigma_2}(x,y)| \\ &\leq \frac{2}{\sigma_0 - \delta} \sup_{x \in \mathbb{R}} |x\psi(x)| \|\sigma_1 - \sigma_2| + \frac{C}{\sigma_0 - \delta} \|f_1 - \widetilde{f}_K\| + \frac{C}{\sigma_0 - \delta} \|f_2 - \widetilde{f}_K\|, \end{aligned}$$

where we have used (A4). This implies that modulo some constants the covering number in the uniform metric may be bounded by the covering number of a Euclidean ball with radius δ and the square of the covering number of a set of functions in $\mathcal{L}^2([0,1])$ with radius c, viz,

$$\mathcal{N}_\infty(\epsilon, \mathcal{G}_{n,c,\delta}) \leq \mathcal{N}(c_1\epsilon, \mathcal{V}_{\sigma_0}) \times \mathcal{N}^2(c_2\epsilon, \{f \in \Theta_K : \|f - \widetilde{f}_K\| \leq c\}),$$

for $V_{\sigma_0} = [\sigma_0 - \delta, \sigma_0 + \delta]$, $c_1 = (\sigma_0 - \delta)/(8 \sup_x |x\psi(x)|)$ and $c_2 = (\sigma_0 - \delta)/(4C)$. By Lemma 2.5 and Corollary 2.6 of van de Geer [44] respectively, these covering numbers may be bounded by

$$\mathcal{N}_\infty(\epsilon, \mathcal{G}_{n,c,\delta}) \leq \left(\frac{2\sigma_0}{c_1\epsilon} + 1\right) \times \left(\frac{4c}{c_2\epsilon} + 1\right)^{2K} \leq \left(\frac{A'}{\epsilon} + 1\right)^{2K+1},$$

for $A' = \max\{2\sigma_0/c_1, 4c/c_2\}$. Now take logarithms, put $A = 2A'$ and use that $K \geq 1$. □

Proof of Theorem 2. To establish Theorem 2 we fill in the details of the development in Section 3.2. In particular, we first establish (8), then we prove (9) and finally we deduce Theorem 2 from (10).

First, note that by Theorem 1, $\|\widehat{f}_n - f_0\|_\infty \xrightarrow{\mathbb{P}} 0$ and by assumption (A2), $\widehat{\sigma}_n \xrightarrow{\mathbb{P}} \sigma_0$. Therefore, we may restrict attention to the set

$$F_n = \{\|\widehat{f}_n - f_0\|_\infty < \delta \wedge |\widehat{\sigma}_n - \sigma_0| < \delta\} \tag{18}$$

for some small $\delta > 0$ to be chosen later.

To prove (8), it suffices to show that, for all $f \in \Theta_K$ and $\sigma > 0$ satisfying $\|f - f_0\|_\infty < \delta$ and $|\sigma - \sigma_0| < \delta$ respectively, we have

$$M(f,\sigma) - M(\widetilde{f}_K, \sigma) \geq \eta |\pi(f, \widetilde{f}_K)|^2 - L\|\widetilde{f}_K - f_0\| \pi(f, \widetilde{f}_K), \tag{19}$$

for some $\eta \geq 0$ and $L > 0$ with high probability. To see that this is sufficient, let us assume without loss of generality that $\eta|\pi(\widehat{f}_n, \widetilde{f}_K)|^2 - L\|\widetilde{f}_K - f_0\|\pi(\widehat{f}_n, \widetilde{f}_K) > 0$ for all large n (if that were not true for some n, then $\eta|\pi(\widehat{f}_n, \widetilde{f}_K)| \leq L\|\widetilde{f}_K - f_0\|$ and there is nothing to prove). Then,

$$\begin{aligned} M(\widehat{f}_n, \widehat{\sigma}_n) - M(\widetilde{f}_K, \widehat{\sigma}_n) &= \frac{M(\widehat{f}_n, \widehat{\sigma}_n) - M(\widetilde{f}_K, \widehat{\sigma}_n)}{\eta|\pi(\widehat{f}_n, \widetilde{f}_K)|^2 - L\|\widetilde{f}_K - f_0\|\pi(\widehat{f}_n, \widetilde{f}_K)} \{\eta|\pi(\widehat{f}_n, \widetilde{f}_K)|^2 - L\|\widetilde{f}_K - f_0\|\pi(\widehat{f}_n, \widetilde{f}_K)\} \\ &\geq \inf_{\substack{\|f - f_0\|_\infty < \delta, |\sigma - \sigma_0| < \delta \\ \eta|\pi(f, \widetilde{f}_K)|^2 - L\|\widetilde{f}_K - f_0\|\pi(f, \widetilde{f}_K) > 0}} \left[\frac{M(f,\sigma) - M(\widetilde{f}_K, \sigma)}{\eta|\pi(f, \widetilde{f}_K)|^2 - L\|\widetilde{f}_K - f_0\|\pi(f, \widetilde{f}_K)} \right] \{\eta|\pi(\widehat{f}_n, \widetilde{f}_K)|^2 - L\|\widetilde{f}_K - f_0\|\pi(\widehat{f}_n, \widetilde{f}_K)\} \\ &\geq \eta|\pi(\widehat{f}_n, \widetilde{f}_K)|^2 - L\|\widetilde{f}_K - f_0\|\pi(\widehat{f}_n, \widetilde{f}_K), \end{aligned}$$

since the infimum is ≥ 1 according to (19).

We thus have to prove inequality (19) to establish (8). First, write $Y - \langle X, f\rangle = \sigma_0\epsilon + R + \langle X, \widetilde{f}_K - f\rangle$, where $R = \langle X, f_0 - \widetilde{f}_K\rangle$. A first order Taylor expansion with Lagrange remainder yields

$$M(f,\sigma) - M(\widetilde{f}_K,\sigma) = \mathbb{E}\left\{\psi\left(\frac{\sigma_0\epsilon + R}{\sigma}\right)\frac{\langle X, \widetilde{f}_K - f\rangle}{\sigma}\right\} + \frac{1}{2}\mathbb{E}\left\{\psi'\left(\frac{\sigma_0\epsilon + R + \xi}{\sigma}\right)\left|\frac{\langle X, \widetilde{f}_K - f\rangle}{\sigma}\right|^2\right\}, \tag{20}$$

for some random variable ξ satisfying $|\xi| \leq |\langle X, \widetilde{f}_K - f\rangle|$. Applying the mean-value theorem on the first term of the rhs of (20) we also find that there exists a random variable χ such that $|\chi| \leq |R|$ and

$$\begin{aligned}\mathbb{E}\left\{\psi\left(\frac{\sigma_0\epsilon + R}{\sigma}\right)\frac{\langle X, \widetilde{f}_K - f\rangle}{\sigma}\right\} &= \mathbb{E}\left\{\psi\left(\frac{\sigma_0\epsilon}{\sigma}\right)\frac{\langle X, \widetilde{f}_K - f\rangle}{\sigma}\right\} + \mathbb{E}\left\{R\psi'\left(\frac{\sigma_0\epsilon + \chi}{\sigma}\right)\frac{\langle X, \widetilde{f}_K - f\rangle}{\sigma}\right\} \\ &= \mathbb{E}\left\{R\psi'\left(\frac{\sigma_0\epsilon + \chi}{\sigma}\right)\frac{\langle X, \widetilde{f}_K - f\rangle}{\sigma}\right\}. \end{aligned} \tag{21}$$

To see why the first term vanishes, note that $\mathbb{E}\{\psi(\sigma_0\epsilon/\sigma)\} = 0$ for any $\sigma > 0$ because Lemma 1 shows that f_0 minimizes $M(f,\sigma)$. For the remaining term in (21), by noting again that $\sigma > \sigma_0 - \delta$ we obtain

$$\begin{aligned}\left|\mathbb{E}\left\{R\psi'\left(\frac{\sigma_0\epsilon + \chi}{\sigma}\right)\frac{\langle X, \widetilde{f}_K - f\rangle}{\sigma}\right\}\right| &\leq (\sigma_0 - \delta)^{-1}\|\psi'\|_\infty \mathbb{E}\{|\langle X, \widetilde{f}_K - f_0\rangle\langle X, \widetilde{f}_K - f\rangle|\} \\ &\leq C(\sigma_0 - \delta)^{-1}\|\psi'\|_\infty \|\widetilde{f}_K - f_0\|\left[\mathbb{E}\{|\langle X, f - \widetilde{f}_K\rangle|^2\}\right]^{1/2} = L_\delta\|\widetilde{f}_K - f_0\|\pi(f, \widetilde{f}_K),\end{aligned}$$

for $L_\delta = C(\sigma_0 - \delta)^{-1}\|\psi'\|_\infty$. This is exactly the second term in the right hand side of (8).

The last part of the proof establishes a strictly positive lower bound on the second term of (20) involving $|\pi(f, \widetilde{f}_K)|^2$. Note that for all $f \in \Theta_K$ satisfying $\|f - f_0\|_\infty < \delta$ we have

$$\|\widetilde{f}_K - f\|_\infty \leq \|\widetilde{f}_K - f_0\|_\infty + \|f - f_0\|_\infty < 2\delta,$$

for all large n, by virtue of (A7). Since X is bounded by (A4), for all large n, $|\xi| \leq 2C\delta$. Assumption (A7) in combination with (A4) also implies $|R| \leq C\delta$ for every δ, for sufficiently large n. By (A1) ψ' is continuous and bounded, and by (A4), $\mathbb{E}\{\psi'(\epsilon)\} > 0$. Hence, $m(t,\sigma) := \mathbb{E}\{\psi((\sigma_0\epsilon + t)/\sigma\}$ is continuous at $(0,\sigma_0)$ and $m(0,\sigma_0) = \mathbb{E}\{\psi'(\epsilon)\} > 0$. This observation now leads to

$$\inf_{|b|<3C\delta, |\sigma-\sigma_0|<\delta} \mathbb{E}\left\{\psi'\left(\frac{\sigma_0\epsilon + b}{\sigma}\right)\right\} \geq \frac{\mathbb{E}\{\psi'(\epsilon)\}}{2} > 0,$$

for all sufficiently small $\delta > 0$. Setting $\eta = (\sigma_0 + \delta)^{-2}\mathbb{E}\{\psi'(\epsilon)\}/4$, we finally have

$$\frac{1}{2}\mathbb{E}\left\{\psi'\left(\frac{\sigma_0\epsilon + R + \xi}{\sigma}\right)\left|\frac{\langle X, \widetilde{f}_K - f\rangle}{\sigma}\right|^2\right\} \geq \eta\mathbb{E}\{|\langle X, f - \widetilde{f}_K\rangle|^2\} = \eta|\pi(f, \widetilde{f}_K)|^2,$$

for all large n, completing the first part of the proof.

The second step in our proof is the establishment of (9). Recall that by assumption (A2) and Theorem 1, we may restrict attention to the set F_n in (18). As previously remarked, in this set we also have $\|\widehat{f}_n - \widetilde{f}_K\|_\infty \leq 2\delta$ for all large n. It then also follows that $\|\widehat{f}_n - \widetilde{f}_K\| \leq 2\delta$ because the uniform norm dominates the $\mathcal{L}^2([0,1])$-norm. Thus, in the notation of Section 3,

$$\left|\frac{U_n(\widehat{f}_n, \widetilde{f}_K, \widehat{\sigma}_n)}{\gamma_n\pi(\widehat{f}_n, \widetilde{f}_K) \vee \gamma_n^2}\right| \leq \sup_{\substack{f\in\Theta_K : \|f - \widetilde{f}_K\| \leq 2\delta \\ |\sigma - \sigma_0| \leq \delta}} \left|\frac{U_n(f, \widetilde{f}_K, \sigma)}{\gamma_n\pi(f, \widetilde{f}_K) \vee \gamma_n^2}\right|, \tag{22}$$

for all large n. To prove (9) it suffices to show that the random variable in the rhs of (22) is bounded in probability. For convenience, let $\Phi_{K,\delta} = \{f \in \Theta_K : \|f - \widetilde{f}_K\| \le 2\delta\}$ and $V_{\sigma_0,\delta} = [\sigma_0 - \delta, \sigma_0 + \delta]$, then we equivalently need to show that

$$\lim_{T\to\infty} \limsup_{n\to\infty} \mathbb{P}\Big(\sup_{\substack{f\in\Phi_{K,\delta},\sigma\in V_{\sigma_0,\delta} \\ \pi(f,\widetilde{f}_K)\le\gamma_n}} \Big| U_n(f, \widetilde{f}_K, \sigma) \Big| \ge T\gamma_n^2 \Big) = 0, \tag{23}$$

as well as

$$\lim_{T\to\infty} \limsup_{n\to\infty} \mathbb{P}\Big(\sup_{\substack{f\in\Phi_{K,\delta},\sigma\in V_{\sigma_0,\delta} \\ \pi(f,\widetilde{f}_K)>\gamma_n}} \Big| \frac{U_n(f, \widetilde{f}_K, \sigma)}{\pi(f, \widetilde{f}_K)} \Big| \ge T\gamma_n \Big) = 0. \tag{24}$$

First, observe that for all $\epsilon > 0$ sufficiently small, say $\epsilon \le \epsilon_0$, there exists a constant $B > 0$ such that

$$\int_0^\epsilon \log^{1/2}\left(1 + \frac{1}{u}\right) du \le B\epsilon \log^{1/2}\left(\frac{1}{\epsilon}\right). \tag{25}$$

This inequality will be useful in the derivation of both (23) and (24). To show (23), we aim to apply Theorem 5.11 of van de Geer [44] on this mean-centered process. Let us rewrite $U_n(f, \widetilde{f}_K, \sigma)$ in terms of the empirical process $U_n(f, \widetilde{f}_K, \sigma) = n^{-1/2} v_n(g_{f,\sigma})$, where, as in Lemma 4,

$$g_{f,\sigma}(X, y) = \rho\left(\frac{y - \langle X, f\rangle}{\sigma}\right) - \rho\left(\frac{y - \langle X, \widetilde{f}_K\rangle}{\sigma}\right),$$

and $v_n(g_{f,\sigma}) = \int g_{f,\sigma} d(\mathbb{P}_n - \mathbb{P})$, with $\mathbb{P}_n$ the empirical measure. By assumption (A1),

$$\sup_{\substack{f\in\Phi_{K,\delta},\sigma\in V_{\sigma_0,\delta} \\ \pi(f,\widetilde{f}_K)\le\gamma_n}} |g_{f,\sigma}| \le 2, \quad \{g_{f,\sigma}, f \in \Phi_{K,\delta}, \sigma \in V_{\sigma_0,\delta}, \pi(f, \widetilde{f}_K) \le \gamma_n\} \subset \mathcal{G}_{n,2\delta,\delta}.$$

Thus, by Lemma 5.10 of van de Geer [44], the generalized entropy with bracketing in the Bernstein norm $\mathcal{H}_{B,8}(\epsilon, \mathcal{G}_{n,2\delta,\delta}, \mathbb{P})$ may be bounded by the bracketing entropy, i.e.,

$$\mathcal{H}_{B,8}\left(\epsilon, \mathcal{G}_{n,2\delta,\delta}, \mathbb{P}\right) \le \mathcal{H}_B\left(\epsilon/\sqrt{2}, \mathcal{G}_{n,2\delta,\delta}, \mathbb{P}\right) \le 3K \log\left(1 + \frac{\sqrt{2}A}{\epsilon}\right), \tag{26}$$

where the last inequality follows from Lemma 4. Furthermore, by (A1) we have

$$|g_{f,\sigma}(X, y)| \le \sigma^{-1} \|\psi\|_\infty |\langle X, \widetilde{f}_K - f\rangle|.$$

Consequently, by definition of $\pi(f, g)$,

$$\sup_{\substack{f\in\Phi_{K,\delta},\sigma\in V_{\sigma_0,\delta} \\ \pi(f,\widetilde{f}_K)\le\gamma_n}} \mathbb{E}\{|g_{f,\sigma}(X, y)|^2\} \le \frac{\|\psi\|_\infty^2}{(\sigma_0 - \delta)^2} \gamma_n^2$$

It follows by Lemma 5.8 of van de Geer [44] that we may take $R = c'\gamma_n$ with $c' = \|\psi\|_\infty/(\sigma_0 - \delta)$ in Theorem 5.11 of van de Geer [44]. We proceed to check the conditions of the theorem. By (A7) we have that $\gamma_n \to 0$ as $n \to \infty$. Hence, by a change of variables and (26), we find

$$\int_0^R \mathcal{H}_{B,8}^{1/2}(u, \mathcal{G}_{n,2\delta,\delta}, \mathbb{P}) du \le C_0 K^{1/2} \gamma_n \log^{1/2} n,$$

for some $C_0 > 0$. By taking $T = C_0$ and $C_1 = 8C_0/c_0^2$ in the theorem, it may be seen that conditions (5.31)–(5.34) in van de Geer [44] are satisfied for $\alpha = Tn^{1/2}\gamma_n^2$ and sufficiently large C_0. Thus, applying Theorem 5.11 of van de Geer [44], there exists a universal constant $C > 0$ such that

$$\mathbb{P}\Big(\sup_{\substack{f\in\Phi_K,\sigma\in V_\sigma\\ \pi(f,\widetilde{f}_K)\le\gamma_n}}\Big|U_n(f,\widetilde{f}_K,\sigma)\Big| \ge T\gamma_n^2\Big) = \mathbb{P}\Big(\sup_{\substack{f\in\Phi_K,\sigma\in V_\sigma\\ \pi(f,\widetilde{f}_K)\le\gamma_n}}\big|v_n(g_{f,s})\big| \ge Tn^{1/2}\gamma_n^2\Big) \le C\exp\Big[-\frac{T^2K\log n}{C^2(C_1+1)}\Big].$$

Since $K \to \infty$ as $n \to \infty$, the exponential tends to zero as $n \to \infty$ and, since this holds for all T sufficiently large, (23) now follows.

To show (24) we modify the peeling argument given in Lemma 5.13 of [44]. First, note that, by (A4), $\pi(f, \widetilde{f}_K) \le C\|f - \widetilde{f}_K\|$ and $\|f - \widetilde{f}_K\| \le 2\delta$ for all $f \in \Phi_{K,\delta}$. By choosing $\delta \le \epsilon_0/(2Cc')$, with ϵ_0 determined in (25) and $c' = \|\psi\|_\infty/(\sigma_0 - \delta)$, we may assume without loss of generality that $\pi(f, \widetilde{f}_K) \le \epsilon_0/c'$ for all $f \in \Phi_{K,\delta}$. Thus, to prove (24), it suffices to prove

$$\lim_{T\to\infty}\limsup_{n\to\infty}\mathbb{P}\Big(\sup_{\substack{f\in\Phi_{K,\delta},\sigma\in V_{\sigma_0}\\ \gamma_n<\pi(f,\widetilde{f}_K)\le\epsilon_0/c'}}\Big|\frac{U_n(f,\widetilde{f}_K,\sigma)}{\pi(f,\widetilde{f}_K)}\Big| \ge T\gamma_n\Big) = 0. \tag{27}$$

Now, let $S = \min\{s > 1 : 2^{-s}\epsilon_0/c' < \gamma_n\}$. Since, by assumption (A7), $K \asymp n^\beta$ for $\beta \in (0,1)$ we clearly have $S \le [c\log_2 n + 1]$ for some $c > 0$. Using Boole's inequality we obtain

$$\begin{aligned}\mathbb{P}\Big(\sup_{\substack{f\in\Phi_{K,\delta},\sigma\in V_{\sigma_0,\delta}\\ \gamma_n<\pi(f,\widetilde{f}_K)\le\epsilon_0/c'}}\Big|\frac{U_n(f,\widetilde{f}_K,\sigma)}{\pi(f,\widetilde{f}_K)}\Big| \ge T\gamma_n\Big) &\le \sum_{s=1}^{S}\mathbb{P}\Big(\sup_{\substack{f\in\Phi_{K,\delta},\sigma\in V_{\sigma_0,\delta}\\ 2^{-s}\epsilon_0/c'<\pi(f,\widetilde{f}_K)\le 2^{-s+1}\epsilon_0/c'}}\Big|\frac{U_n(f,\widetilde{f}_K,\sigma)}{\pi(f,\widetilde{f}_K)}\Big| \ge T\gamma_n\Big)\\ &\le \sum_{s=1}^{S}\mathbb{P}\Big(\sup_{\substack{f\in\Phi_{K,\delta},\sigma\in V_{\sigma_0,\delta}\\ \pi(f,\widetilde{f}_K)\le 2^{-s+1}\epsilon_0/c'}}\Big|U_n(f,\widetilde{f}_K,\sigma)\Big| \ge T2^{-s}\frac{\epsilon_0}{c'}\gamma_n\Big).\end{aligned}$$

We bound each one of these summands through individual application of Theorem 5.11 of van de Geer [44] (see also the proof of (23)). Rewriting in terms of the empirical process we have

$$\mathbb{P}\Big(\sup_{\substack{f\in\Phi_{K,\delta},\sigma\in V_{\sigma_0,\delta}\\ \pi(f,\widetilde{f}_K)\le 2^{-s+1}\epsilon_0/c'}}\Big|U_n(f,\widetilde{f}_K,\sigma)\Big| \ge T2^{-s}\frac{\epsilon_0}{c'}\gamma_n\Big) = \mathbb{P}\Big(\sup_{\substack{f\in\Phi_{K,\delta},\sigma\in V_{\sigma_0,\delta}\\ \pi(f,\widetilde{f}_K)\le 2^{-s+1}\epsilon_0/c'}}|v_n(g_{f,\sigma})| \ge T2^{-s}\frac{\epsilon_0}{c'}n^{1/2}\gamma_n\Big)$$

Clearly, $2^{-s+1}c'\epsilon_0/c' \le \epsilon_0$ for all $1 \le s \le S$. Hence, for all sufficiently large C_0 the bracketing integral for each one of these classes may be bounded by

$$\int_0^{2^{-s+1}\epsilon_0}\mathcal{H}_{B,8}^{1/2}(u,\mathcal{G}_{n,2\delta,\delta},\mathbb{P})du \le C_0K^{1/2}2^{-s+1}\epsilon_0\log n,$$

for all large n, by the construction of S, i.e., $2^S \le 2cn$ for large n. The conditions of Theorem 5.11 in van de Geer [44] are satisfied for sufficiently large C_0, $C_1 = 8C_0c'$ and $T = C_0$ since by definition of S, we have $\gamma_n \le 2^{-s+1}\epsilon_0/c'$ for all $1 \le s \le S$. Thus, this theorem yields

$$\mathbb{P}\Big(\sup_{\substack{f\in\Phi_{K,\delta},\sigma\in V_{\sigma_0,\delta}\\ \pi(f,\widetilde{f}_K)\le 2^{-s+1}\epsilon_0/c'}}\Big|U_n(f,\widetilde{f}_K,\sigma)\Big| \ge T2^{-s}\frac{\epsilon_0}{c'}\gamma_n\Big) \le C\exp\Big[-\frac{T^2K\log n}{4C^2(C_1+1)|c'|^2}\Big],$$

for the same universal constant $C > 0$. None of these terms depend on s, hence after summing over $s \in \{1,\dots,S\}$ and recalling that $S \le [c\log_2 n + 1]$ we obtain

$$\mathbb{P}\Big(\sup_{\substack{f\in\Phi_{K,\delta},\sigma\in V_{\sigma_0,\delta}\\ \gamma_n<\pi(f,\widetilde{f}_K)\le\epsilon_0/c'}}\Big|\frac{U_n(f,\widetilde{f}_K,\sigma)}{\pi(f,\widetilde{f}_K)}\Big| \ge T\gamma_n\Big) \le C'[c\log_2 n + 1]\exp\Big[-\frac{T^2K\log n}{4C^2(C_1+1)|c'|^2}\Big],$$

for some $C' > 0$ and all large n. We have thus established (24) and part (ii) now follows.

To complete the proof we now deduce Theorem 2 from (10). First, note that by (A4) we have

$$|\pi(\widehat{f}_n, f_0)|^2 \leq 2|\pi(\widehat{f}_n, \widetilde{f}_K)|^2 + 2|\pi(\widetilde{f}_K), f_0)|^2 \leq 2|\pi(\widehat{f}_n, \widetilde{f}_K)|^2 + 2C^2\|\widetilde{f}_K - f_0\|^2. \tag{28}$$

Hence, we need to study $|\pi(\widehat{f}_n, \widetilde{f}_K)|^2$. We only have to handle the case $\gamma_n\pi(\widehat{f}_n, \widetilde{f}_K) > \gamma_n^2$, or, equivalently $\pi(\widehat{f}_n, \widetilde{f}_K) > \gamma_n$, since for $\gamma_n\pi(\widehat{f}_n, \widetilde{f}_K) \leq \gamma_n^2$, the theorem clearly holds. From parts (i) and (ii) we have

$$\eta|\pi(\widehat{f}_n, \widetilde{f}_K)|^2 \leq U_n(\widetilde{f}_K, \widehat{f}_n, \widehat{\sigma}_n) + L\|\widetilde{f}_K - f_0\|\pi(\widehat{f}_n, \widetilde{f}_K) + \lambda\mathcal{J}(\widetilde{f}_K) = O_{\mathbb{P}}(1)\gamma_n\pi(\widehat{f}_n, \widetilde{f}_K) + L\|\widetilde{f}_K - f_0\|\pi(\widehat{f}_n, \widetilde{f}_K) + \lambda\mathcal{J}(\widetilde{f}_K).$$

Equivalently, since $\eta > 0$,

$$|\pi(\widehat{f}_n, \widetilde{f}_K)|^2 \leq O_{\mathbb{P}}(1)\gamma_n\pi(\widehat{f}_n, \widetilde{f}_K) + O_{\mathbb{P}}(1)\|\widetilde{f}_K - f_0\|\pi(\widehat{f}_n, \widetilde{f}_K) + \lambda\mathcal{J}(\widetilde{f}_K)/\eta.$$

Now, this is an inequality of the form $x_0^2 \leq bx_0 + c$ with $x_0 = \pi(\widehat{f}_n, \widetilde{f}_K) \geq 0, b = O_{\mathbb{P}}(1)(\|\widetilde{f}_K - f_0\| + \gamma_n)$ and $c = \lambda\mathcal{J}(\widetilde{f}_K)/\eta$. This means that x_0 must be less than or equal to the positive root of $x^2 - bx - c = 0$, that is,

$$0 \leq x_0 \leq \frac{b + \sqrt{b^2 + 4c}}{2} \leq b + \sqrt{c},$$

and after substituting the expressions of x_0, b and c, we obtain

$$\pi(\widehat{f}_n, \widetilde{f}_K) \leq O_{\mathbb{P}}(1)(\|\widetilde{f}_K - f_0\| + \gamma_n) + O_{\mathbb{P}}\left(\sqrt{\lambda\mathcal{J}(\widetilde{f}_K)}\right).$$

Squaring and using the inequality $(x + y)^2 \leq 2x^2 + 2y^2$ twice yields

$$|\pi(\widehat{f}_n, \widetilde{f}_K)|^2 \leq O_{\mathbb{P}}(1)\gamma_n^2 + O_{\mathbb{P}}(1)\|\widetilde{f}_K - f_0\|^2 + O_{\mathbb{P}}(1)\lambda\mathcal{J}(\widetilde{f}_K).$$

The result of the theorem now follows easily from (28) which completes the proof. □

9. Supplementary material

The accompanying supplementary material contains additional simulation results.

References

[1] R. A. Adams, J. J. F. Fournier, Sobolev Spaces, Elsevier/Academic Press, Amsterdam, The Netherlands, second edition, 2003.
[2] G. Boente, M. Salibián-Barrera, P. Vena, Robust estimation for semi-functional linear regression models, Comp. Statist. Data Anal. 152 (2020).
[3] H. Cardot, F. Ferraty, P. Sarda, Functional linear model, Statist. Probab. Lett. 45 (1999) 11–22.
[4] H. Cardot, F. Ferraty, P. Sarda, Spline estimators for the functional linear model, Statist. Sinica 13 (2003) 571–591.
[5] G. V. Cohen Freue, D. Kepplinger, M. Salibián Barrera, E. Smucler, Robust Elastic Net Estimators for Variable Selection and Identification of Proteomic Biomarkers, Ann. Appl. Stat. 13 (2019) 2065–2090.
[6] C. Crambes, A. Kneip, P. Sarda, Smoothing splines estimators for functional linear regression, Ann. Statist. 37 (2009) 35–72.
[7] C. de Boor, A Practical Guide to Splines, Springer, New York, revised edition, 2001.
[8] R. A. DeVore, G.G. Lorentz, Constructive Approximation, Springer, Berlin, 1993.
[9] P. P. B. Eggermont, V. N. LaRiccia, Maximum Penalized Likelihood Estimation, Volume II: Regression, Springer, New York, 2009.
[10] P. H. C. Eilers, B. D. Marx, Flexible smoothing with B-splines and penalties, Statist. Sci. 11 (1996) 89–102.
[11] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis: Theory and Practice, Springer, New York, 2006.
[12] P. Filzmoser, K. Varmuza, `Chemometrics`: R companion to the book "Introduction to Multivariate Statistical Analysis in Chemometrics", K. Varmuza and P. Filzmoser (2009), 2017.
[13] J. Goldsmith, J. Bobb, C. M. Crainiceanu, B. Caffo, D. Reich, Penalized functional regression, Comput. Statist. Data Anal. 20 (2011) 830–851.
[14] J. Goldsmith, F. Scheipl, Estimator selection and combination in scalar-on-function regression, Comput. Statist. Data Anal. 70 (2011) 362–372.
[15] P. Hall, J. L. Horowitz, Methodology and convergence rates for functional linear regression, Ann. Statist. 35 (2007) 70–91.
[16] T. Hsing, R. Eubank, Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators, Wiley, New York, 2007.

[17] G.M. James, J. Wang, J. Zhu, Functional linear regression that's interpretable, Ann. Statist. 37 (2009) 271–293.
[18] K. Janssens, I. Deraedt, O. Schalm, J. Veeckman, Composition of 15–17th century archaeological glass vessels excavated in Antwerp, Belgium, in: G. Love, W.A.P. Nicholson, A. Armigliato (Eds.), Modern Developments and Applications in Microbeam Analysis, Springer, Vienna 253–267, 1998.
[19] J. A. Khan, S. Van Aelst, R. H. Zamar, Fast Robust Estimation of Prediction Error Based on Resampling, Comput. Statist. Data Anal. 54 (2010) 3121-3130.
[20] I. Kalogridis, S. Van Aelst, M-type penalized splines with auxiliary scale estimation, J. Statist. Plann. Inference 212 (2021) 97–113.
[21] I. Kalogridis, S. Van Aelst, Robust penalized spline estimation with difference penalties, Econom. Stat. appeared online (2021).
[22] I. Kalogridis, S. Van Aelst, Robust optimal estimation of location from discretely sampled functional data, Scand. J. Stat., appeared online (2022).
[23] Y. Li, T. Hsing, On rates of convergence in functional linear regression, J. Multivariate Anal. 98 (2007) 1782–1804.
[24] E. Mammen, S. van de Geer, Locally adaptive regression splines, Ann. Statist. 25 (1997) 387–413.
[25] R. A. Maronna, Robust ridge regression for high-dimensional data, Technometrics 53 (2011) 44–53.
[26] R. A. Maronna, V. J. Yohai, Robust functional linear regression based on splines, Comput. Statist. Data Anal. 65 (2013) 46–55.
[27] R. A. Maronna, D. Martin, M. Salibián-Barrera, V. J. Yohai, Robust Statistics: Theory and Methods, Wiley, Chichester, second edition, 2019.
[28] J. S. Morris, Functional regression, Annu. Rev. Stat. Appl. 2 (2015) 321–359.
[29] J. Nocedal, S. J. Wright, Numerical Optimization Springer, New York, second edition, 2006.
[30] F. O'Sullivan, A statistical perspective of ill-posed problems, Statist. Sci. 1 (1986) 502–518.
[31] T. Qingguo, M-estimation for functional linear regression, Comm. Statist. Theory Methods 46 (2017) 3782–3800.
[32] J. O. Ramsay, When the data are functions, Psychometrika 47 (1982) 379–396.
[33] J. O. Ramsay., C. J. Dalzell, Some Tools for Functional Data Analysis, J. R. Stat. Soc. Ser. B. Stat. Methodol 53 (1991) 539–561.
[34] J. O. Ramsay, B. W. Silverman, Functional Data Analysis, Wiley, New York, 2005.
[35] J. Raymaekers, P. J. Rousseeuw, W. Van den Bossche, M. Hubert, `CellWise`: Analyzing Data with Cellwise Outliers, 2019.
[36] P. T. Reiss, R. T. Ogden, Functional principal component regression and functional partial least squares, J. Amer. Statist. Assoc. 102 (2007) 984–996.
[37] P. T. Reiss, J. Goldsmith, H. L. Shang, R. T. Ogden Methods for scalar-on-function regression, Int. Stat. Rev. 85 (2017) 228–249.
[38] P. J. Rousseeuw, V. J. Yohai, Robust regression by means of S-estimators, in: Franke, J., Härdle, W.K., Martin D. (Eds.), Robust and Nonlinear Time Series Analysis, Springer, Berlin, 256–272, 1984.
[39] D. Ruppert, M. P. Wand, R. J. Carroll, Semiparametric Regression, Cambridge, New York, 2003.
[40] M. Salibian-Barrera, V. J. Yohai, A Fast Algorithm for S-Regression Estimates, J. Comput. Graph. Statist. 15 (2006) 414–427.
[41] X. Shen, W. H. Wong, Convergence Rate of Sieve Estimates, Ann. Statist. 22 (1994) 580–615.
[42] H. Shin, S. Lee, An RKHS approach to robust functional linear regression, Statist. Sinica 26 (2016) 255–272.
[43] E. Smucler, V. J. Yohai Robust and sparse estimators for linear regression models, Comput. Statist. Data Anal. 111 (2017) 116–130.
[44] S. van de Geer, Empirical Processes in M-Estimation, Cambridge University Press, New York, NY, 2000.
[45] S. van de Geer, M-estimation using penalties or sieves, J. Stat. Plann. Infer. 108 (2002) 55-–69.
[46] G. Wahba, Spline models for observational data, Siam, Philadelphia, Pen., 1990.
[47] S. Wood, Generalized Additive Models, CRC Press, Boca Raton, FL., second edition, 2017.
[48] M. Yuan, T. T. Cai, A reproducing kernel Hilbert space approach to functional linear regression, Ann. Statist. 38 (2010) 3412–3444.
[49] V. J. Yohai, High breakdown-point and high efficiency robust estimates for regression, Technical report No. 66 , Dept. Statistics, Univ. Washington, Seattle, 1985.
[50] V. J. Yohai, High breakdown-point and high efficiency robust estimates for regression, Ann. Statist. 15 (1987) 642–656.
[51] V. J. Yohai, R. H. Zamar, High Breakdown-Point Estimates of Regression by Means of the Minimization of an Efficient Scale, J. Amer. Statist. Assoc. 83 (1988) 406–413.
[52] Y. Zhao, T. R. Ogden, P. T. Reiss, Wavelet-based LASSO in functional linear regression, J. Comput. Graph. Statist. 21 (2012) 600–617.
[53] J. Zhou, M. Chen, Spline estimators for semi-functional linear model, Statist. Probab. Lett. 82 (2012) 505–513.

Author statement

Ioannis Kalogridis: Conceptualization, methodology, investigation, software, writing- original draft

Stefan Van Aelst: supervision, writing- review & editing