Electronic Journal of Statistics
Vol. 16 (2022) 5751–5778
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS2067

Estimating the conditional distribution in functional regression problems*

Siegfried Hörmann†

Institute of Statistics,
Graz University of Technology, Austria
e-mail: shoermann@tugraz.at

Thomas Kuenzer

Institute of Statistics,
Graz University of Technology, Austria
e-mail: kuenzer@tugraz.at

and

Gregory Rice

Department of Statistics and Actuarial Science,
University of Waterloo, Canada
e-mail: grice@uwaterloo.ca

Abstract: We consider the problem of estimating the conditional distribution $\mathrm{P}(Y \in A|X)$ of a functional data object $Y = (Y(t) : t \in [0,1])$ in the space of continuous functions, given covariates X in a general space and assuming that Y and X are related by a functional linear regression model. Two estimation methods are proposed, based on either the empirical distribution of the estimated model residuals, or fitting functional parametric models to the model residuals. We show that consistent estimation can be achieved under relatively mild assumptions. We exemplify a general class of sets A specifying path properties of Y that are of interest in applications. The proposed methods are studied in several simulation experiments, and data analyses of electricity price and pollution curves.

MSC2020 subject classifications: 62G05, 62G20, 62J05.
Keywords and phrases: Functional regression, functional quantile regression, functional time series, empirical distribution, prediction sets.

Received June 2022.

1. Introduction

We suppose we have observed data $(Y_1, X_1), \ldots, (Y_n, X_n)$ from a strictly stationary process $(Y_k, X_k)_{k\in\mathbb{Z}}$ that are assumed to follow a general functional regression model with additive noise of the form

$$Y_k = \varrho X_k + \varepsilon_k. \tag{1}$$

*This research was partly funded by the Austrian Science Fund (FWF) [P 35520], and the Natural Science and Engineering Research Council of Canada [RGPIN-03723].
†Corresponding author

Here $Y_k = (Y_k(t)\colon t \in [0,1])$ is a curve in the space $C[0,1]$ of continuous functions on the unit interval. Assuming the functions are defined over $[0,1]$ is done simply for convenience, and $[0,1]$ could be replaced easily by a general interval. The covariates X_k take values in a normed space H and are distributed so that X_k is independent of the model error ε_k, which is assumed to stem from a zero-mean i.i.d. sequence. The operator ϱ is assumed to be a bounded linear mapping from H to $C[0,1]$ satisfying suitable conditions (see, e.g., Assumption 5 (b)). If the covariate X_k is a single curve, then (1) describes a linear function-on-function regression. This setting also includes functional autoregressive models [6] when $X_k = Y_{k-1}$. We may also have multiple functional predictors (see e.g. Ivanescu et al. [32]), in which case X_k is a vector of curves. Generally though, X_k might be also comprised of a mixture of curves and scalar covariates, etc.

Suppose (Y, X) is a generic pair following (1). *The primary goal of this paper is to introduce and study methods to consistently estimate the conditional distribution of Y given X, i.e.* $\mathrm{P}(Y \in A|X)$*, for sets $A \subset C[0,1]$.* The set A can be used to describe a property of the response curve. One property of special interest in this article will be the following.

Example 1 (Level sets). Let $\alpha \in \mathbb{R}$ and $z \in [0,1]$, and define

$$A_{\alpha,z} := \big\{y \in C[0,1]\colon \lambda(t\colon y(t) > \alpha) \le z\big\},$$

where again λ denotes the standard Lebesgue measure on $[0,1]$. The set $A_{\alpha,z}$ contains curves that stay a limited amount of time z above a threshold α.

The time a curve spends in a certain range can be viewed as a simple scalar summary of the curve, and often sets of interest A can be expressed in terms of such scalar summaries. We call a scalar transformation $T(Y)$ of the response a *curve feature*. Evidently, methods to generally estimate $\mathrm{P}(Y \in A|X)$ may also be used to estimate the conditional distribution of curve features $T(Y)$, i.e. to estimate the probability that Y will lie in sets of the form $A = \{y\colon T(y) \le z\}$, conditional on the covariate X.

When considering a curve feature $Z = T(Y)$ of the response, it is natural to simply model Z directly. For example, in order to estimate the conditional distribution of Z, one might compute the scalar responses $Z_i = T(Y_i)$, and then fit a scalar-on-function regression using the data $(Z_1, X_1), ..., (Z_n, X_n)$. This is the approach undertaken most frequently in the literature related to this problem, which we review below.

An issue that arises here though is that it is difficult to determine a suitable parametric model for Z in terms of X. We note that even if the curves Y and X are related by model (1), the relationship between the amount of time the response spends in a certain range $T(Y) = \lambda(t : Y(t) \in [a,b])$ and X is complex and non-linear. One of the main strengths of the approach we pursue, which is distinct from competitive methods, is that we model the entire response curve before extracting features of interest. When the relationship between the response curve and the covariates Y and X can be approximated by (1), it is advantageous to incorporate the full information contained in the functional

responses and regression model to approximate the potentially complex relationship between $T(Y)$ and X.

An important related problem to estimating $\mathrm{P}(Y \in A|X)$ for a given A is to determine for a given $p \in (0,1)$ a suitable *prediction set* $A_p = A_p(X)$, such that $\mathrm{P}(Y \in A_p|X) \geq p$. Estimating the conditional distribution is then needed to appropriately calibrate A_p. See Goldsmith, Greven and Crainiceanu [23], Choi and Reimherr [12], Liebl and Reimherr [35], Hyndman and Shang [29], and Paparoditis and Shang [40] for a review of methods for constructing prediction sets for functional responses and parameters.

Most often, prediction sets for functional responses are given in terms of prediction bands. Prediction bands are related to the quantiles of the curve features $Z_1 = \max_{t\in[0,1]} Y(t)$ and $Z_2 = \min_{t\in[0,1]} Y(t)$. E.g., for a one-sided prediction band, one has to determine q_p such that $\mathrm{P}(Z_1 \leq q_p|X) = p$. This problem then can be seen as a special case of the more general setting where we estimate the *conditional quantile function* of a curve feature $Z = T(Y)$.

A second goal of this paper is to explore how our proposed estimation scheme can be used for the construction of prediction sets and conditional quantile functions of curve features.

Aside from our interest in these general problems, this work was primarily motivated by the statistical challenge of forecasting aspects of response curves Y_k describing daily electricity prices. The specific data that we consider consists of hourly electricity prices, demand, and wind energy production in Spain over the period from 2014 to 2019, which includes observations from 2191 days (the data are available at `www.esios.ree.es`). We project the hourly data onto a basis of 18 twice differentiable B-splines to construct daily price, demand, and wind energy production curves, as illustrated in Figure 1. The price of electricity naturally fluctuates based on supply and demand, and exhibits daily, weekly, and yearly seasonality. The rather predictable variation in demand does not influence the price as much as surges in wind energy production, especially if they occur on days with weak demand. Letting Y_k denote the price curves and X_k the vector of the demand and wind curves, both adjusted for yearly seasonality and trends, we then model Y_k using an FAR(7) model with exogenous variables

$$Y_k = \sum_{i=1}^{7} \Psi_i Y_{k-i} + \varrho X_k + \varepsilon_k, \tag{2}$$

where $\Psi_1, ..., \Psi_7$ denote autoregressive operators; see González, Muñoz San Roque and Pérez [24]. The details of this are explained in Section 6, but for now it suffices to acknowledge that this is a regression model of the form (1). For such electricity price curves, their likelihood of falling within sets as given in Example 1 is of particular interest, as forecasting whether price or demand curves will spend prolonged periods of time above certain levels is useful in anticipating volatility in continuous intraday electricity markets, and planning for peak loads [48].

The literature on functional regression models of the form (1) is vast. The questions that have been most investigated are (i) how to find a consistent es-

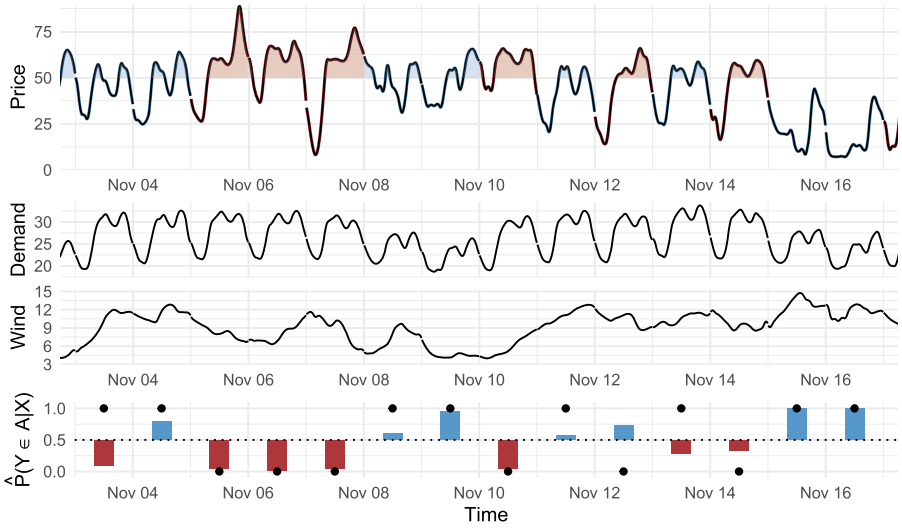

FIG 1. *Spanish electricity data on price, demand, and wind energy production during two weeks in November 2014. Price curves are colored blue or red according to whether or not they lie in the level set* $\{y \in C[0,1] : \lambda(t : y(t) > 50) \leq 0.5\}$. *The bar plot on the bottom shows the estimated conditional probability for* Y_k *to lie in this set, with the decision threshold* $1/2$, *indicated by a dotted line, whether the event occurred is indicated by black dots.*

timator $\hat{\varrho}_n$ of ϱ, and (ii) how to forecast consistently, i.e. to guarantee that $\hat{\varrho}_n X - \varrho X \to 0$ suitably in probability. Moreover, the majority of the literature devoted to linear regression with functional responses concentrates on curves in $L^2[0,1]$, the separable Hilbert space of square integrable functions on the unit interval. Ramsay and Silverman [43] for example proposes a double truncation scheme based on functional principal component analysis to estimate ϱ in this setting, and Mas [36], Imaizumi and Kato [31] derive a convergence rate for $\|\hat{\varrho}_n - \varrho\|_{\mathcal{S}}$ in a "single-truncation" estimation scheme based on an increasing (in the sample size) number of principal components, where $\|\cdot\|_{\mathcal{S}}$ denotes the Hilbert–Schmidt norm. Similar consistency results for the resulting forecasts in functional linear regression can be found in Crambes and Mas [14], and under general stationarity conditions and in the FAR setting in Hörmann and Kidziński [26] and Aue, Norinho and Hörmann [1]. Estimating the operator ϱ can be viewed as a special case of estimating the conditional mean $\mathrm{E}[Y|X]$, and this general problem has also been extensively considered; see Chiou, Müller and Wang [11], Ferraty, Van Keilegom and Vieu [18], and Wang, Chiou and Müller [49].

The problem of estimating the conditional distribution of Y given X has been comparatively far less studied. Numerous methods have been proposed to estimate the conditional distribution of a scalar response Y with a functional covariate X, including Chen and Müller [8], Kato [33], Yao, Sue-Chee and Wang [51], Wang, Chiou and Müller [49], and Sang and Cao [45], who propose esti-

mators based on quantile regression, and Ferraty and Vieu [20], who propose Nadaraya–Watson style kernel-smoothed estimators. Estimating the conditional distribution of Y when Y takes values in a function space is largely unexplored to our knowledge, even in the context of model (1). Fernández de Castro, Guillas and González Manteiga [17] and Paparoditis and Shang [40] develop bootstrap procedures based on functional principal component analysis to produce prediction sets in the context of forecasting with Hilbertian FAR models, which can be viewed as a special case of this problem. Franke and Nyarige [21] develop a related residual-based bootstrap for the purpose of estimating the sampling distribution of statistics based on Hilbertian FAR series. For functional data taking values in $L^2[0,1]$, Chen and Müller [9] and Fan and Müller [16] develop methods for estimating the conditional distribution of Y given X assuming that X and Y are jointly Gaussian, and that the conditional distribution of the response has sample paths satisfying differentiability conditions.

In this paper, we propose natural procedures to estimate $\mathrm{P}(Y \in A|X)$, in which we first estimate ϱ with a suitably consistent estimator $\hat{\varrho}_n$, and then either (i) use the estimated residuals $\hat{\varepsilon}_{k,n} = Y_k - \hat{\varrho}_n X_k$ to estimate $\mathrm{P}(Y \in A|X)$ with the empirical distribution of $\hat{\varrho}_n X + \hat{\varepsilon}_{k,n}$, or (ii) assuming Gaussianity of the model errors ε_k, we estimate $\mathrm{P}(Y \in A|X)$ using simulation by modelling Y conditioned on X as a Gaussian process with mean $\hat{\varrho}_n X$, and covariance estimated from the residual sequence $\hat{\varepsilon}_{k,n}$. We establish general conditions on the estimator $\hat{\varrho}_n$ in the setting when the response space is $C[0,1]$ such that our proposed algorithms will lead to consistent estimation of $\mathrm{P}(Y \in A|X)$. Subsequent to this, we define an estimator $\hat{\varrho}_n$ and show that it satisfies these conditions under regularity assumptions on the operator ϱ, and the process $(Y_k, X_k)_{k\in\mathbb{Z}}$. In particular, the conditions we assume allow for serial dependence of both the response and covariates.

In principle, the algorithms we propose can be applied to estimate the conditional distribution $\mathrm{P}(Y \in A|X)$ for any arbitrary set A. However, and perhaps not surprisingly, consistent estimation can only be expected for continuity sets of the distribution of the response, i.e. sets A for which $\mathrm{P}(Y \in \partial A) = 0$. This property evidently depends strongly on the space the response curve lies in, as well as the norm that the space is equipped with. We remark that for many interesting examples, the metric on the space $L^2[0,1]$ is too weak to allow for meaningful continuity sets A. An illustrative example is uniform prediction band sets of the form $A = \{y\colon \lambda(t : a(t) < y(t) < b(t)) = 1\}$, where a and b are continuous functions on $[0,1]$, in which case $\partial A = A$ when A is viewed as a subset of $L^2[0,1]$. To handle many interesting examples in functional data analysis involving path properties of the response, such as level sets, the space $C[0,1]$ is much more natural. The literature on estimating and consistently forecasting with ϱ outside of the $L^2[0,1]$-framework is sparse and limited in specialized settings; see Pumo [41], Ruiz-Medina and Álvarez-Liébana [44], and Bosq [6] in the context of FAR estimation. Furthermore, the problem of consistently estimating the conditional distribution $\mathrm{P}(Y \in A|X)$ in these settings has not been studied, to our knowledge. We also refer the reader to Dette, Kokot and Aue [15] for a review of functional data analysis methods in $C[0,1]$.

The rest of the paper is organized as follows. In Section 2, we formally introduce the methods described above to estimate $\mathrm{P}(Y \in A|X)$. In Section 3 we present results on their consistency, including results on uniform consistency over monotone families of sets A that are relevant in constructing prediction sets with a specified coverage and quantile function estimates. We also briefly discuss the problem of establishing convergence rates of our estimators. These results depend on the properties of the estimator $\hat{\varrho}_n$, and in Section 4 we define an estimator $\hat{\varrho}_n$ based on functional principal component analysis and a single truncation regularization scheme, and establish that it leads to consistent estimation of $\mathrm{P}(Y \in A|X)$. Due to the central role that level-sets play in our real data example, we provide in Section 5 a discussion on the boundaries with respect to the sup-norm in $C[0,1]$. A number of competing methods are introduced in Section 6, and these are compared with the proposed methods in several simulations studies and real data illustrations. The proofs of all results can be found in Supplementary Material [28].

We conclude this introduction by providing some notation that will be used throughout this article. We consider the response space $C[0,1]$ equipped with the supremum norm $\|f\|_\infty = \sup_{t\in[0,1]}|f(t)|$ and the covariate space H, which is assumed to be a Hilbert space with inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|$. In order to lighten the notation, and when it is clear from context, we write $\|\cdot\|$ also for the norm on $C[0,1]$. We use the tensor product notation $\otimes$ to denote the operator $a\otimes b(\cdot) = a\langle\cdot, b\rangle$ if b is viewed as an element of a Hilbert space and the kernel integral operator with kernel $a\otimes b(t,s) = a(t)b(s)$ if b is viewed as an element of $C[0,1]$.

2. Estimation procedures

We assume throughout that the covariates X_k and the model errors ε_k satisfy the following independence condition, which we do not explicitly state in the below results, but take as granted.

Assumption 1. In model (1), $(\varepsilon_k)_{k\in\mathbb{Z}}$ is an i.i.d. sequence in $C[0,1]$ with $\mathrm{E}\,\varepsilon_k = 0$, and ε_k is independent from $(X_j)_{j\le k}$ for all $k\in\mathbb{Z}$.

In order to formally describe the methods we use to estimate $\mathrm{P}(Y \in A|X)$, we assume for the moment that we may consistently estimate ϱ with an estimator $\hat{\varrho}_n$ based on the sample $\mathcal{S}_n = \{(Y_1,X_1),\ldots,(Y_n,X_n)\}$. Specific conditions on this estimator, and an estimator that satisfies those conditions, will follow. The first method we describe directly uses the estimated residuals as a surrogate for the noise.

Algorithm 1 (empirical distribution, abbreviated **empir**).

1. Estimate ϱ in (1) with $\hat{\varrho}_n$.
2. Calculate the model residuals $\hat{\varepsilon}_{k,n} = Y_k - \hat{\varrho}_n X_k$.

3. Define the estimator of $\mathrm{P}(Y \in A|X)$ as

$$\widehat{P}_n^{\mathrm{E}}(Y \in A|X) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}\{\hat{\varrho}_n X + \hat{\varepsilon}_{k,n} \in A\}.$$

We show below that this estimator is weakly consistent under mild conditions (see, e.g., Assumption 2 and Theorem 1, or Proposition 2 under Assumptions 4 and 5 (a)). The algorithm **empir** can be applied without specific distributional assumptions on the errors. It is not uncommon, though, that model errors ε_k are thought to be Gaussian processes. As the model errors have mean zero by assumption, in this special case their distribution is determined by their covariance

$$\Gamma = \mathrm{E}\big[\varepsilon_k \otimes \varepsilon_k\big], \quad k \in \mathbb{Z}.$$

The above algorithm may then be adapted as follows:

Algorithm 2 (Gaussian process estimation, abbreviated **Gauss**).

1. Estimate the model residuals $\hat{\varepsilon}_{k,n}$ as in Algorithm 1.
2. Estimate the covariance operator of the noise by

$$\widehat{\Gamma}_{\varepsilon,n} = \frac{1}{n}\sum_{k=1}^{n}(\hat{\varepsilon}_{k,n} - \bar{\varepsilon}_{\cdot,n}) \otimes (\hat{\varepsilon}_{k,n} - \bar{\varepsilon}_{\cdot,n}), \tag{3}$$

   where $\bar{\varepsilon}_{\cdot,n} = \frac{1}{n}\sum_{k=1}^{n} \hat{\varepsilon}_{k,n}$.
3. Conditional on the sample $\mathcal{S}_n$, let $\varepsilon^{(n)}$ be a Gaussian variable with zero mean and covariance $\widehat{\Gamma}_{\varepsilon,n}$. Then set

$$\widehat{P}_n^{\mathrm{G}}(Y \in A|X) = \mathrm{P}\big(\hat{\varrho}_n X + \varepsilon^{(n)} \in A\big|X, \mathcal{S}_n\big). \tag{4}$$

   The latter probability can be approximated by Monte-Carlo simulation, see Remark 2.

Remark 1. The scaling $1/n$ in the definition of $\widehat{\Gamma}_{\varepsilon,n}$ does not take into account the degrees of freedom T_n lost in the estimation of the regression operator ϱ. It has thus been advocated, for example in [13], to instead divide by $n - T_n$, where T_n is related to the dimension of the dimensionality reduction technique used in estimating $\hat{\varrho}_n$. If $ET_n = o(n)$, as is the case for most estimation approaches, the resulting scaling difference is asymptotically negligible. Some authors also propose splitting the sample and estimating the regression operator and the noise covariance operator on separate parts of the sample in order to reduce the bias of the estimator $\widehat{\Gamma}_{\varepsilon,n}$; see [14].

Remark 2. In order to determine $\mathrm{P}\big(\hat{\varrho}_n X + \varepsilon^{(n)} \in A\big|X, \mathcal{S}_n\big)$, we proceed as follows: let $(\hat{\nu}_{j,n})_{1\le j\le n}$ denote the eigenvalues of $\widehat{\Gamma}_{\varepsilon,n}$, with corresponding eigenfunctions $(\hat{\psi}_{j,n})_{1\le j\le n}$ satisfying $\widehat{\Gamma}_{\varepsilon,n}\hat{\psi}_{j,n} = \hat{\nu}_{j,n}\hat{\psi}_{j,n}$ and $\langle\hat{\psi}_{i,n}, \hat{\psi}_{j,n}\rangle = \delta_{i\,j}$. Let $\{Z_i,\ 1 \le i \le n\}$ denote a sequence of i.i.d. standard normal random variables,

independent of the sample $\mathcal{S}_n$, and define

$$\varepsilon^{(n)} = \sum_{j=1}^{n} \hat{\nu}_{j,n}^{1/2} Z_j \hat{\psi}_{j,n}.$$

Conditionally on the sample, in particular on $\widehat{\Gamma}_{\varepsilon,n}$, $\varepsilon^{(n)}$ is a Gaussian process with mean zero and covariance operator $\widehat{\Gamma}_{\varepsilon,n}$. Now generate in the same way i.i.d. copies $(\varepsilon_k^{(n)})_{k\geq 1}$ and estimate the right-hand side of (4), for a large M, by

$$\frac{1}{M}\sum_{k=1}^{M} \mathbb{1}\big\{\hat{\varrho}_n X + \varepsilon_k^{(n)} \in A\big\}.$$

Remark 3. Algorithm **Gauss** can be extended to other parametric distributions of the noise. A notable example for this are infinite-dimensional elliptical distributions, where $\varepsilon_k = \Xi_k \varepsilon'_k$ with two independent random variables $\varepsilon'_k \in C[0,1]$, which is Gaussian, and $\Xi_k \geq 0$ is a scalar random variable from a known univariate parametric distribution, for instance the Pareto distribution. The following investigation can easily be adapted to this setting. For details on elliptical distributions of functional data, we refer to [5].

Remark 4. An advantage, which derives from the simple form of both our estimators, is that they satisfy the basic properties of a probability measure, e.g. they are monotone in A. While this may seem like an obvious requirement, it is not always fulfilled by competing approaches; see Section 6.1.

3. Consistency results

3.1. Estimation algorithms

We now aim to establish consistency results for the proposed algorithms. In order to disentangle how the estimation of $\mathrm{P}(Y \in A|X)$ depends on the estimation of ϱ, and hence allow for different estimators, these results are initially stated in terms of the following basic consistency properties of $\hat{\varrho}_n$. A specific estimator satisfying these conditions is provided in Section 4.

Assumption 2. The estimator $\hat{\varrho}_n$ is such that

(a) its out-of-sample prediction is consistent, i.e. if X is independent from the sample and $X \overset{d}{=} X_1$, then

$$\|\hat{\varrho}_n X - \varrho X\| \xrightarrow{\mathrm{P}} 0, \quad n \to \infty.$$

(b) its in-sample prediction is consistent, i.e. let K_n be independent from the sample and uniformly distributed on $\{1,\dots,n\}$, then

$$\|\hat{\varrho}_n X_{K_n} - \varrho X_{K_n}\| \xrightarrow{\mathrm{P}} 0, \quad n \to \infty.$$

In order to establish the consistency of the algorithm **Gauss**, we need additional conditions on the model errors and the estimated covariance operator (3).

Assumption 3.

(a) The estimator $\hat{\varrho}_n$ is such that

$$\sup_{t,s\in[0,1]} \left| \frac{1}{n} \sum_{i=1}^{n} (\hat{\varrho}_n - \varrho)(X_k) \otimes (\hat{\varrho}_n - \varrho)(X_k)(t,s) \right| \xrightarrow{\mathrm{P}} 0, \quad n \to \infty.$$

(b) The estimated variance of the model error increments,

$$V_n^2(t,s) = \mathrm{Var}\big(\varepsilon^{(n)}(t) - \varepsilon^{(n)}(s) \,\big|\, \widehat{\Gamma}_{\varepsilon,n}\big) = \widehat{\Gamma}_{\varepsilon,n}(t,t) - 2\widehat{\Gamma}_{\varepsilon,n}(t,s) + \widehat{\Gamma}_{\varepsilon,n}(s,s)$$

satisfies the Hölder condition

$$V_n^2(t,s) < M_V^2 \, |t-s|^{2\alpha}, \quad t,s \in [0,1],$$

for some $0 < \alpha \leq 1$, where M_V is a random variable with $\mathrm{E}\, M_V < \infty$.

Assumption 3 implicitly demands a degree of continuity of ϱX_k and the model errors ε_k. Under the above assumptions, we can now formulate our main consistency results.

Theorem 1. *Suppose that Assumption 2 holds and that* $\mathrm{P}(Y \in \partial A) = 0$. *Then* $\widehat{P}_n^E(Y \in A|X) \xrightarrow{\mathrm{P}} \mathrm{P}(Y \in A|X)$ *as* $n \to \infty$.

Theorem 2. *Suppose that Assumptions 2 (a) and 3 hold. Assume that* $(\varepsilon_k)_{k\geq 1}$ *are i.i.d. Gaussian random variables in* $C[0,1]$, *and that* $\mathrm{P}(Y \in \partial A) = 0$. *Then* $\widehat{P}_n^G(Y \in A|X) \xrightarrow{\mathrm{P}} \mathrm{P}(Y \in A|X)$ *as* $n \to \infty$.

Theorems 1 and 2 show that consistent estimation of $\mathrm{P}(Y \in A|X)$ is achieved by both **empir** and **Gauss** under natural consistency conditions on $\hat{\varrho}_n$, and when A is a continuity set of the response Y. The convergence results above can be related via a version of the Portmanteau theorem to the notion of "weak convergence in probability" to the true conditional distribution of the response, see [50]. This also gives a hint as to why consistency holds only for continuity sets of the limiting distribution. Note that up to this stage we do not even require ϱ to be linear.

3.2. Prediction sets

We note that the results of Theorems 1 and 2 can be readily extended to sets A that, rather than being fixed, are dependent on the predictor X and the estimator $\hat{\varrho}_n$, so long as there is a certain degree of continuity in relating $\{Y \in A\}$ to $\hat{\varrho}_n X$. This is of interest when constructing prediction sets for the response Y, as in the following examples.

Example 2 (Pointwise and uniform prediction sets). Suppose a and b are positive functions in $C[0,1]$. Given a covariate X, define, for $s \in [0,1]$, the *point prediction sets*

$$\hat{A}_{a,b}^{(n)}(s) = \big\{y \in C[0,1]\colon \hat{\varrho}_n X(s) - a(s) \le y(s) \le \hat{\varrho}_n X(s) + b(s)\big\},$$

and the *uniform prediction sets*

$$\hat{U}_{a,b}^{(n)} = \big\{y \in C[0,1]\colon \lambda\big(t : \hat{\varrho}_n X(t) - a(t) \le y(t) \le \hat{\varrho}_n X(t) + b(t)\big) = 1\big\}.$$

These approximate the true sets $A_{a,b}(s)$ and $U_{a,b}$ where the estimator $\hat{\varrho}_n$ is replaced by ϱ.

Corollary 1. *For some $s \in [0,1]$, let $\hat{A}_{a,b}^{(n)}(s)$, $\hat{U}_{a,b}^{(n)}$, $A_{a,b}(s)$, and $U_{a,b}$ be defined in Example 2. Suppose that $\mathrm{P}(Y \in \partial A_{a,b}(s)) = 0$. If Assumption 2 holds, then*

$$\widehat{P}_n^E(Y \in \hat{A}_{a,b}^{(n)}(s)|X) \xrightarrow{\mathrm{P}} \mathrm{P}(Y \in A_{a,b}(s)|X), \quad \text{as } n \to \infty. \tag{5}$$

If Assumptions 2 (a) and 3 hold, then (5) holds with $\widehat{P}_n^G$ instead of $\widehat{P}_n^E$. Under $\mathrm{P}(Y \in \partial U_{a,b}) = 0$, the analogue results hold with the sets $\hat{U}_{a,b}^{(n)}$ and $U_{a,b}$.

3.3. Monotone families of sets and conditional quantiles

For a potentially unbounded interval $[a,b] \subset \overline{\mathbb{R}}$, we call a family $\mathcal{A} = \{A_\xi\colon \xi \in [a,b]\}$ of measurable sets monotone if the sets A_ξ are increasing or decreasing in ξ. Suppose that A_ξ is increasing, the decreasing case can be handled similarly, and that we are interested in finding

$$\xi_p(X) = \inf\big\{\xi \in [a,b]\colon \mathrm{P}(Y \in A_\xi|X) \ge p\big\}, \quad p \in (0,1).$$

Note that on a technical level, this definition only makes sense for p that a.s. satisfy $\mathrm{P}(Y \in A_a|X) < p \le \mathrm{P}(Y \in A_b|X)$. As an example where this problem is relevant, consider a scalar transformation of the response $Z = T(Y)$, and suppose we wish to estimate the conditional quantile of Z given the covariate X,

$$\begin{aligned} q_p(Z|X) &= \inf\big\{\xi \in [a,b]\colon \mathrm{P}(Z \le \xi|X) \ge p\big\} \\ &= \inf\big\{\xi \in [a,b]\colon \mathrm{P}\big(Y \in T^{-1}\big([a,\xi]\big)|X\big) \ge p\big\}. \end{aligned}$$

The sets $A_\xi := T^{-1}\big([a,\xi]\big)$ evidently define a monotone family. Consistent scalar-on-function quantile regression can hence be cast as the problem of consistently estimating $\xi_p(X)$, which can be done using $\widehat{P}_n^{\mathrm{E}}$ or $\widehat{P}_n^{\mathrm{G}}$. To this end, we consider the estimator

$$\hat{\xi}_p^{\mathrm{E}}(X) := \inf\big\{\xi \in [a,b]\colon \widehat{P}_n^{\mathrm{E}}(Y \in A_\xi|X) \ge p\big\}. \tag{6}$$

We note that based on the definition of $\widehat{P}_n^{\mathrm{E}}$, $p \mapsto \hat{\xi}_p^{\mathrm{E}}(X)$ is a non-decreasing function. The same holds for $\hat{\xi}_p^{\mathrm{G}}(X)$, which is defined using $\widehat{P}_n^{\mathrm{G}}$. While this

observation is rather trivial, in other approaches to scalar-on-function quantile regression one often has to take special care in order to guarantee monotonicity of estimators of $q_p(Z|X)$, see e.g. [33].

The goal is now to show that $\hat{\xi}_p^{\mathrm{E}}(X) \xrightarrow{\mathrm{P}} \xi_p(X)$ and $\hat{\xi}_p^{\mathrm{G}}(X) \xrightarrow{\mathrm{P}} \xi_p(X)$. In order to do so, we need the following uniform convergence result for the estimated conditional probabilities.

Proposition 1. *Let $\{A_\xi : \xi \in [a,b]\}$ be a monotone family of sets such that $\mathrm{P}(Y \in A_\xi|X)$ is a.s. continuous in ξ. Suppose the estimator $\widehat{P}_n(Y \in A_\xi|X)$ is non-decreasing, right-continuous, and satisfies $\widehat{P}_n(Y \in A_\xi|X) \xrightarrow{\mathrm{P}} \mathrm{P}(Y \in A_\xi|X)$ for all $\xi \in [a,b]$. Then*

$$\sup_{\xi\in[a,b]} \left|\widehat{P}_n(Y \in A_\xi|X) - \mathrm{P}(Y \in A_\xi|X)\right| \xrightarrow{\mathrm{P}} 0, \quad n \to \infty.$$

We note that both $\widehat{P}_n^{\mathrm{E}}(Y \in A_\xi|X)$ and $\widehat{P}_n^{\mathrm{G}}(Y \in A_\xi|X)$ satisfy the conditions of Proposition 1 under the conditions of Theorems 1 and 2.

Corollary 2. *Let $p \in (0,1)$, and assume that the interval $[a,b]$ is such that a.s. $\mathrm{P}(Y \in A_a|X) < p \le \mathrm{P}(Y \in A_b|X)$. Define $\hat{\xi}_p(X)$ as in (6) for a general estimator $\widehat{P}_n(Y \in A_\xi|X)$. Under the assumptions of Proposition 1 with increasing sets A_ξ, we have that $\mathrm{P}(Y \in A_{\hat{\xi}_p(X)}|X) \xrightarrow{\mathrm{P}} p$. If the true probability $\mathrm{P}(Y \in A_\xi|X)$ is strictly increasing in ξ, then $\hat{\xi}_p(X) \xrightarrow{\mathrm{P}} \xi_p(X)$.*

3.4. Rates of convergence

The results above confirm consistency of the proposed algorithms. From a theoretical point of view, it appears interesting to obtain corresponding rates of convergence. However, given the challenge of even establishing estimation rates for the point estimator of the linear regression operator, it is clear that a comprehensive answer to this problem requires its own devoted article and is outside the scope of this paper. In this section, we give a short glance to this problem on the basis of Algorithm 1. Below we use the notation

$$\partial_{(\delta)}A = \big\{x\colon\ \exists\, z_1 \in A, z_2 \in A^c \text{ with } \max\big(\|x - z_1\|,\ \|x - z_2\|\big) \le \delta\big\}.$$

and

$$h(\delta, n) = \mathrm{P}\big(\|(\hat{\varrho}_n - \varrho)X\| > \delta/2\big) + \frac{1}{n}\sum_{k=1}^{n} P\big(\|(\hat{\varrho}_n - \varrho)X_k\| > \delta/2\big).$$

Here X is as in Assumption 2.

Theorem 3. *Consider a non-negative sequence (δ_n) such that $h(\delta_n, n) = o(1)$. Let (b_n) be such that $\mathrm{P}(Y \in \partial_{(\delta_n)}A) = o(b_n)$. Then*

$$\big|\widehat{P}_n^{E}\big(Y \in A\big|X\big) - \mathrm{P}\big(Y \in A\big|X\big)\big| = \mathcal{O}_P\Big(b_n \vee n^{-1/2}\Big).$$

It remains open if such a rate can be shown to be optimal, i.e. if a corresponding lower bound can also be established. The bound involves two rate conditions that appear rather intuitive: δ_n reflects the convergence rate of the regression operator estimator, and $\mathrm{P}(Y \in \partial_{(\delta_n)} A) = o(b_n)$ is a quantitative version of the condition $\mathrm{P}(Y \in \partial A) = 0$ that also takes into account the convergence rate of $\hat{\varrho}_n$ reflected in δ_n. Importantly, this convergence rate depends not only on the distribution of the process Y at hand, and the estimator $\hat{\varrho}_n$, but also on the set A considered. Thus it does not appear possible to obtain a universal rate over all A, unless A is restricted to certain classes of sets. In general for any rate δ_n, one can construct examples of processes Y and sets A for which $\mathrm{P}(Y \in \partial_{(\delta_n)} A) > b_n$ for any b_n tending to zero. As for the estimation of the regression operator, a polynomial rate of convergence is established for a bounded linear operator ϱ in Example 5 in the next section. Assuming a polynomial estimation rate for $\hat{\varrho}_n$, let us consider two special cases.

Example 3. Suppose $A = \{y : \max_u y(u) \leq \alpha\}$. This *extremal set* corresponds to a level set with $z = 0$. If Y is such that its maximum has a bounded probability density function (e.g. this is the case for the Brownian motion), then $\mathrm{P}(Y \in \partial_{(\delta_n)} A) = \mathcal{O}(\delta_n)$. Suppose we have a regression operator estimator that is consistent with a prediction error of $o_P\big(n^{-\gamma}\big)$, with $0 < \gamma \leq 1/2$, then $\widehat{P}^{\mathrm{E}}$ attains a convergence rate of $\mathcal{O}_P\big(n^{-\gamma}\big)$.

Example 4. If we consider a level set $A_{\alpha,z} = \big\{y\colon\ \lambda\big(u\colon\ y(u) > \alpha\big) \leq z\big\}$ with $z \in (0, 1]$, we need to inspect the set

$$\partial_{(\delta)} A_{\alpha,z} = \Big\{y\colon\ \lambda\big(u\colon\ y(u) > \alpha + \delta\big) \leq z < \lambda\big(u\colon\ y(u) > \alpha - \delta\big)\Big\}. \tag{7}$$

It can be shown that for the Brownian motion, it holds that for any $\epsilon > 0$, we have $\mathrm{P}\big(W \in \partial_{(\delta)} A_{\alpha,z}\big) = o\big(\delta^{1-\epsilon}\big)$, implying that in this setting we may choose $b_n = \delta_n^{1-\epsilon}$ in Theorem 3. The proof for this can be found in Supplementary Material; see Lemma 4. Suppose, as in the previous example, that we have an estimator $\hat{\varrho}_n$ that attains a prediction error of $\mathcal{O}_P\big(n^{-\gamma}\big)$, with $0 < \gamma \leq 1/2$, then $\widehat{P}^{\mathrm{E}}$ again attains a convergence rate of $\mathcal{O}_P\big(n^{-\gamma+\epsilon}\big)$ for any $\epsilon > 0$.

4. Estimation of the regression operator

In this section we aim to define an estimator $\hat{\varrho}_n$ that satisfies the consistency conditions detailed in Assumptions 2 and 3 when ϱ in model (1) is linear. In order to do so, we make the following assumptions on model (1).

Assumption 4. (a) H is a separable Hilbert space.

(b) The process $(X_k)_{k\in\mathbb{Z}}$ has mean zero, and is L^4-m-approximable in H.

(c) The operator $\varrho\colon H \to C[0, 1]$ is a bounded linear operator.

(d) The sequence $(\varepsilon_k)_{k\in\mathbb{Z}}$ is a mean zero, i.i.d. sequence in $C[0, 1]$, and satisfies $\mathrm{E}\,\|\varepsilon_k\|^4 < \infty$.

Assumption 4 (b) implies that X_k is a (strongly) stationary and ergodic sequence with $\mathrm{E}\,\|X_k\|^4 < \infty$, and allows the X_k to be weakly serially dependent

in a certain sense, see [27]. Hörmann and Kokoszka [27] show that many commonly studied stationary time series in function space, like FAR processes or functional analogs of GARCH processes, are L^4-m-approximable under suitable moment conditions.

The estimator that we consider is a truncated (functional) principal components–based estimator. Let the empirical covariance operator of X_k, and the empirical cross-covariance operator between Y_k and X_k, be denoted as

$$\widehat{C}_{XX} = \frac{1}{n}\sum_{k=1}^{n} X_k \otimes X_k, \quad \text{and} \quad \widehat{C}_{YX} = \frac{1}{n}\sum_{k=1}^{n} Y_k \otimes X_k.$$

In order to lighten the notation, the empirical covariance operator of some sequence Z_k will be denoted by $\widehat{C}_Z = \widehat{C}_{ZZ}$, and analogously for the population version of this. We note that $\widehat{C}_X$ defines a non-increasing sequence of eigenvalues $\hat{\lambda}_i \geq 0$, and eigenfunctions $\hat{v}_i$, satisfying $\widehat{C}_X(\hat{v}_i) = \hat{\lambda}_i \hat{v}_i$, and $\langle \hat{v}_i, \hat{v}_j \rangle = \delta_{ij}$.We then define

$$\hat{\varrho}_n(x) := \sum_{i=1}^{T_n} \frac{1}{\hat{\lambda}_i} \widehat{C}_{YX}\, \hat{v}_i \otimes \hat{v}_i(x), \tag{8}$$

The estimator (8) only truncates the covariance operator of X in order to obtain a feasible approximation to $\widehat{C}_X^{-1}$, yielding a so-called "single-truncated" estimator. The asymptotic properties of these estimated operators have, e.g., been studied in [36] and [26]. In order to select the truncation parameter T_n in such a way that leads to asymptotic consistency of $\hat{\varrho}_n$, we use the following criterion:

$$T_n = \max\{j \geq 1 \colon \hat{\lambda}_j \geq m_n^{-1}\}, \quad \text{with } \; m_n \to \infty. \tag{9}$$

Here m_n is a tuning parameter, tending to infinity at a rate specified in the results below. We refer to [26] for details on the consistency properties of this choice of T_n.

We note that another standard way to select T_n is to use the percentage of variance explained (PVE) approach, which entails taking

$$T_n = \min\left\{ d : \frac{\sum_{j=1}^{d} \hat{\lambda}_j}{\sum_{j=1}^{n} \hat{\lambda}_j} \geq v \right\}, \tag{10}$$

where v is a user specified percentage treated as a tuning parameter. While the criterion in (9) is more transparent in terms of describing the asymptotic consistency of $\hat{\varrho}_n$, since it gives a direct description of the decay rate of the sequence of eigenvalues $\hat{\lambda}_j$, in applications the PVE criterion is prevailing, due to its ease of interpretation. By choosing the associated tuning parameters appropriately, the two criteria may be made comparable.

Now we present results which imply Assumptions 2 and 3, and hence the consistency of the estimators in the Algorithms **empir** and **Gauss**.

We add the following assumption in addition to Assumption 4, supposing a degree of smoothness to ϱX and ε_k:

Assumption 5. For some $0 < \alpha \leq 1$,

(a) The model errors ε_k a.s. satisfy the Hölder condition

$$\left|\varepsilon_k(t) - \varepsilon_k(s)\right| < M_k \left|t - s\right|^{\alpha} \tag{11}$$

where M_k is a random variable independent from X_k, with $\mathrm{E}\, M_k^2 < \infty$.

(b) For a finite constant M_ϱ and all $x \in H$, the regression operator ϱ satisfies

$$\left|\varrho x(t) - \varrho x(s)\right| \leq M_\varrho \left\|x\right\| \left|t - s\right|^{\alpha}.$$

Assumption 5 (a) is fulfilled by a wide range of stochastic processes, most notably the Brownian motion and the fractional Brownian motion. Since ϱ is linear, Assumption 5 (b) is a natural formulation of the Hölder condition for the conditional mean of the response. In particular, this implies that ϱ is a compact operator.

Remark 5. Suppose $H = L^2[0,1]$, so that model (1) describes function-on-function regression. A frequently employed class of operators ϱ in this setting are kernel integral operators, defined by a continuous kernel $\rho \in C[0,1]^2$ as $\varrho x(t) = \int \rho(t,u)\, x(u) du$. If there exists an $a \in H$ such that almost everywhere

$$\left|\rho(t,u) - \rho(s,u)\right| < a(u) \left|t - s\right|^{\alpha},$$

then Assumption 5 (b) is easily verified since $\left|\varrho x(t) - \varrho x(s)\right| \leq \left\|a\right\| \left\|x\right\| \left|t - s\right|^{\alpha}$.

Proposition 2. *Suppose that Assumptions 4 and 5 (a) hold, and we define* $\hat{\varrho}_n$ *as in* (8) *with* $m_n = o\big(n^{\alpha/2}\big)$. *Then Assumption 2 holds. If we additionally assume Assumption 5 (b), then Assumption 3 also holds.*

We conclude this section with some technical discussion. We begin by noting that the sequence m_n, which controls how many principal components of X_k are used in forming $\hat{\varrho}_n$, can be of asymptotically higher order if α is larger, meaning that the paths of the noise process are less rough. In the case where the Hölder exponent in Assumption 5 is $\alpha = 1$, the responses Y_k are Lipschitz continuous, which implies that they are weakly differentiable. The order $n^{\alpha/2}$ is sufficient but not sharp. In fact, for $\alpha < 1/2$, a different proof shows that $m_n = o\big(n^{1/(2+\alpha^{-1})}\big)$ also leads to consistency. In the case of the Brownian motion, the order condition demands that $m_n = \mathcal{O}\big(n^{1/4-\epsilon}\big)$ for some $\epsilon > 0$. This is still of higher order than that suggested by [26] for consistent estimation of the regression operator in Hilbert spaces.

Our second technical remark concerns the choice of H. Assumption 4 (a) requires H to be a Hilbert space. Typically, this is not a restriction, but some care needs to be taken in the case of an FAR model. While it is natural to assume that the covariate and response space coincide for an FAR (i.e. requiring $H = C[0,1]$, too), this is not necessarily the case. For example, when we consider a kernel integral operator ϱ with a continuous kernel, then we may still use $H = L^2[0,1]$ because of the natural embedding of $C[0,1]$ in $L^2[0,1]$.

The following example is only of illustrative nature and complements the discussion and examples presented in Section 3.4.

Example 5. If X is confined to a finite-dimensional space H, then the estimation simplifies greatly and we readily obtain a rate of $\mathcal{O}_P\big(n^{-\alpha/2}\big)$ for the prediction error. For the general, infinite-dimensional case, deriving explicit rates for $\hat{\varrho}_n$ as we defined it in Section 4 is inconvenient. For the sake of illustration, suppose we instead choose the truncation parameter $T_n =: d_n \to \infty$ deterministically. An adaptation of the proof to Proposition 2 then yields the following rate of convergence for the squared point prediction error:

$$\big\|(\hat{\varrho}_n - \varrho)X\big\|^2 + \big\|(\hat{\varrho}_n - \varrho)X_K\big\|^2 = \mathcal{O}_P\Big(\lambda_{d+1} + n^{-1}d\,(\lambda_d - \lambda_{d+1})^{-2} + n^{-\alpha}\lambda_d^{-2}\Big).$$

Here $K = K_n$ is as in Assumption 2. These prediction errors converge to zero for a suitable choice of d_n. Suppose, e.g., that the eigenvalues $\lambda_i \sim i^{-2}$ and we choose the asymptotically optimal truncation parameter d_n depending on α. Then the prediction error is $\mathcal{O}_P\big(n^{-\frac{\alpha\wedge 2/3}{6}}\big)$. An illustrative case is the Brownian motion, which satisfies the Hölder condition for all $\alpha < 1/2$. In such a case, $\delta_n = n^{-1/12+\epsilon}$ for some $\epsilon > 0$ is a possible choice in Theorem 3. Together with Examples 3 and 4, it follows that when considering level sets with Y distributed as a Brownian motion, convergence rates of $\mathcal{O}_P\big(n^{-1/12+\epsilon}\big)$ for any $\epsilon > 0$ are attainable.

5. The boundary condition

A crucial condition in Theorems 1 and 2 is that $\mathrm{P}(Y \in \partial A) = 0$. Below we exemplify this condition in the case of level sets (Example 1).

Proposition 3. *Let $\alpha \in \mathbb{R}$ and $z \in [0,1)$. Consider the level set $A_{\alpha,z}$ defined as in Example 1. The following conditions are sufficient for* $\mathrm{P}(Y \in \partial A_{\alpha,z}) = 0$.

(A) If $z \in (0,1)$:

$$\text{(i)} \quad \mathrm{P}\big(\lambda(Y = \alpha) > 0\big) = 0 \quad \textit{and} \quad \text{(ii)} \quad \mathrm{P}\big(\lambda(Y > \alpha) = z\big) = 0$$

(B) If $z = 0$:

$$\text{(iii)} \quad \mathrm{P}\Big(\sup_{t\in[0,1]} Y(t) = \alpha\Big) = 0.$$

The conditions of Proposition 3 are satisfied by many well-known processes, including the Brownian motion. They are also generally satisfied by continuously differentiable Gaussian processes under standard non-degeneracy conditions. Such processes might be used to model functional data generated by applying standard smoothing operations, for instance using cubic splines or trigonometric polynomials, to raw discrete data. We note that comparable differentiability conditions are assumed in [16]. The following proposition describes these conditions.

Proposition 4. *Suppose that Y is a continuously differentiable Gaussian process with covariance kernel C_Y. If $C_Y(t,t) > 0$ for all $t \in [0,1]$, then* (i) *holds.*

For $\ell \in \mathbb{N}$, and $0 \le t_1 < \cdots < t_\ell \le 1$, let

$$r_Y(t,s) = \frac{C_Y(t,s)}{[C_Y(t,t)C_Y(s,s)]^{1/2}}, \quad \text{and} \quad R_{t_1,\ldots,t_\ell} = \{r_Y(t_i,t_j)\}_{1\le i,j\le \ell} \in \mathbb{R}^{\ell\times\ell}.$$

If, in addition, for all $\ell \in \mathbb{N}$ and $0 \le t_1 < \cdots < t_\ell \le 1$, there exist constants $c_1, c_2 > 0$ such that $\det(R_{t_1,\ldots,t_\ell}) \ge c_1 \min_{1\le i\ne j\le \ell} |t_i - t_j|^{c_2}$, then (ii) *holds. If Y is twice continuously differentiable, and*

$$(Y(t_1),\ldots,Y(t_\ell), Y'(t_1),\ldots,Y'(t_\ell), Y''(t_1),\ldots,Y''(t_\ell))$$

has a non-degenerate distribution, then (iii) *holds.*

6. Simulation experiments and data illustrations

In this section we present the results of simulation experiments and real data analyses that aimed to evaluate and compare the performance of our algorithms, and illustrate their application. We begin by defining some alternate methods that may be used to estimate $\mathrm{P}(Y \in A|X)$, and we describe two recent procedures proposed for construction of prediction sets in functional data forecasting and functional quantile regression, respectively.

6.1. Competing methods

A simple method to estimate $\mathrm{P}(Y \in A|X)$ is to employ functional binomial regression. This entails positing the model $\mathrm{P}(Y \in A|X = x) = g\big(\beta_0 + \langle x, \beta\rangle\big)$ for some $\beta_0 \in \mathbb{R}$ and $\beta \in L^2[0,1]$, and a link function g that can be chosen from a variety of possibilities, but is most often the logistic link function, or the cumulative distribution function of the standard normal distribution (the "probit link"). For more details of such models, we refer to [38] and [37]. One notable drawback of of this approach is that changing the set A necessitates refitting the model, which can be computationally cumbersome. A theoretical deficiency is that the resulting estimators of $\mathrm{P}(Y \in A|X)$ need not be monotone with respect to increasing sets A. An approach to adjust such estimators to restore monotonicity is to use rearrangement or isotonization, as discussed in [10].

Since the exact relationship between the function X and the event $\{Y \in A\}$ is unknown and difficult to describe in parametric terms, even under model (1), another promising approach is to use nonparametric techniques such as kernel estimators. Generalizing the method found in Section 5.4 of [20], the conditional distribution $\mathrm{P}(Y \in A|X = x) = \mathrm{E}[\mathbb{1}\{Y \in A\}|X = x]$ can be estimated by the functional extension of the Nadaraya–Watson estimator

$$\widehat{P}^{\mathrm{NW}}(Y \in A|X = x) = \frac{\sum_{i=1}^n K\big(h^{-1}d(x,X_i)\big)\,\mathbb{1}\{Y_i \in A\}}{\sum_{i=1}^n K\big(h^{-1}d(x,X_i)\big)}, \tag{12}$$

where K is a kernel function on the nonnegative real numbers, d is a distance measure on H, and $h > 0$ is a smoothing parameter corresponding to the bandwidth of the kernel. While the choice of K is typically unproblematic, the choice of d is more intricate and is often taken to depend on the data. The bandwidth h represents the trade-off between bias (oversmoothing) and error (undersmoothing), and is normally taken to decrease with the sample size n. Ferraty and Vieu [20] establish consistency conditions for the estimator (12) in the case when the sequence $\{(X_k, Y_k)\colon k \geq 1\}$ is α-mixing and Y is a scalar. When we apply this method below, we take K to be the standard Gaussian kernel, d to be the norm on H, and select h using cross-validation. We note that similarly to functional logistic regression–based estimators, a drawback of these estimators is that if one changes the set A, then the bandwidth h in general should be recalibrated, and the resulting estimators need not be monotone in A if the bandwidth h is not held fixed for all sets A.

Similar options may be derived from the local linear functional estimator, which improves upon the Nadaraya–Watson estimator by including a linear term of the form $\langle x - X_i, \beta\rangle$ into the computation of the weights; see [3] and [19]. The k-nearest neighbors (kNN) functional estimator is a variation on the Nadaraya–Watson estimator with adaptive bandwidth, i.e. h is the smallest number such that $\bigl|\{X_i\colon d(x, X_i) \leq h\}\bigr| = k$. The kNN estimator has been shown to be consistent for non-parametric regression in [34]. Since we found the performance of the kNN estimator to be comparable, we only report results of the simpler Nadaraya–Watson estimator (12).

In order to evaluate the proposed algorithms for the construction of prediction sets, we compare to the method of Paparoditis and Shang [40] in the setting of forecasting FAR(1) processes $Y_k - \mu = \varrho(Y_{k-1} - \mu) + \varepsilon_k$. We construct uniform prediction bands (see Example 2) for Y_{n+1} of the form

$$\bigl\{y \in C[0,1]\colon\ \widehat{Y}_{n+1}(t) + L\hat{\sigma}_{n+1}(t) \leq y(t) \leq \widehat{Y}_{n+1}(t) + U\hat{\sigma}_{n+1}(t),\ \text{for all } t \in [0,1]\bigr\},$$

where $\widehat{Y}_{n+1} = \hat{\mu} + \hat{\varrho}_n(Y_n - \hat{\mu})$ is the point prediction, $\hat{\sigma}^2_{n+1}(t) = \widehat{\operatorname{Var}}\bigl(\hat{\varepsilon}_{K,n}(t)\bigr) = \widehat{\Gamma}_{\varepsilon,n}(t,t)$ is the estimated pointwise variance of the response, with K uniformly distributed on $\{1, \dots, n\}$, and for a specified coverage level $1 - \alpha$,

$$M = \sup_{t\in[0,1]} \frac{\bigl|\hat{\varepsilon}_{K,n}(t)\bigr|}{\hat{\sigma}_{n+1}(t)}, \quad U = Q_{1-\alpha}(M), \quad \text{and} \quad L = -U.$$

Here $Q_{1-\alpha}(\cdot)$ denotes the $(1-\alpha)$-quantile of a random variable. In comparison to the proposed methods **empir** and **Gauss**, Paparoditis and Shang [40] do not simply use the empirical distribution of the (centered) residuals $\hat{\varepsilon}_{K,n}$. Instead, they employ a more involved bootstrapping procedure. They start with a given forecasting model that delivers point predictions and then use a sieve bootstrap procedure that mimics the dependence structure of the functional time series Y_k to estimate the model misspecification error of the forecasting model. Subsequently, they arrive at the bootstrapped prediction error distribution of the forecasting model point predictions, which they then use to construct the

prediction bands. The theoretical consistency results of Paparoditis and Shang [40] are derived under the L^2 metric, and therefore do not immediately imply the asymptotic consistency of the coverage probability for uniform prediction band-sets, since such a set is nowhere dense in $L^2[0,1]$. The same is true for pointwise prediction bands.

In the setting of scalar-on-function quantile regression, we compare to the method of Sang and Cao [45], which entails for a scalar response $T(Y)$ modelling

$$T(Y) = g\big(\beta_0 + \langle x, \beta\rangle\big) + \varepsilon,$$

where g is an assumed to be unknown link function. The link function g as well as the parameter function β are assumed to be linear combinations of splines, and estimated in order to estimate the level p quantile of $T(Y)$ by minimizing the check function loss

$$\rho_p(y) = \big(p - \mathbb{1}\{y \le 0\}\big)\, y, \tag{13}$$

subject also to a roughness penalty on the functions g and β.

6.2. *Comparison to functional GLM and Nadaraya–Watson estimation*

In this simulation experiment, we generated synthetic data under model (1) in such a way that it resembled a real functional time series derived from daily square-root transformed PM_{10} concentration curves constructed by smoothing half-hourly measurements of PM_{10}. This is done using the function `Data2fd` in the `fda` package with default settings; see [42]. PM_{10} concentration denotes the concentration in air of respirable coarse particles having a diameter less than 10 μm, and the data that we consider was collected in Graz, Austria over the period from October 1st, 2010 to March 31st, 2011. An illustration of these data is given in Figure 2, and they are available in the `ftsa` package in `R`; see [30].

We use these data as a means to devise a realistic data generating process. To this end, we first fit an FAR(1) model to the square-root transformed PM_{10} curves. The estimator of the FAR operator ϱ obtained in this way differs from operators typically used in simulation settings in that it is highly asymmetrical, as illustrated in the right-hand panel of Figure 2.

With the estimated sample mean and the fitted FAR operator, we then generate synthetic FAR(1) time series samples by drawing noise ε_k from a Gaussian distribution, with the covariance operator estimated from the residuals of the FAR(1) model fit to the original data. This can be done as in the algorithm **Gauss**. The first 30 observations are dropped as a burn-in phase. A snapshot of the raw data in comparison to the synthetic data can be seen in Figure 2. In this manner we may generate time series of arbitrary sample sizes that are similar to the original PM_{10} data. We generated 1000 independent samples for each sample size $n \in \{50, 100, 250, 1000\}$. Then, for 50 different values of predictors Y_0^* simulated independently from the stationary distribution of the data generating process, we estimated the conditional probability of Y_1^* lying in the level

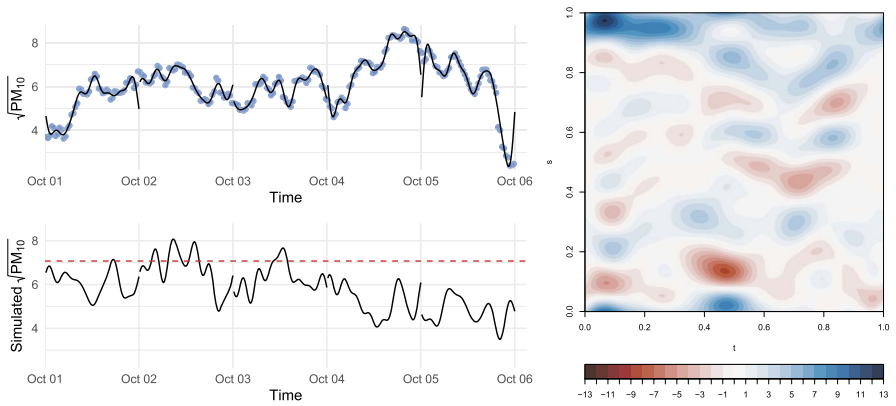

FIG 2. *Top left: the raw PM_{10} measurements (blue) with the fitted curves (black). Bottom left: simulated synthetic PM_{10} data (black) with $\alpha = \sqrt{50}$ (red) that we considered in the level set case. Right: the kernel operator $\varrho(t, s)$ used in the data generating process.*

set $\mathrm{P}\big(\lambda(Y_1^* > \sqrt{50}) \le 0.5 \big| Y_0^*\big)$ for each such sample. For each of the 50 predictors, we also approximated the true probability using Monte-Carlo simulation ($n_{\mathrm{MC}} = 10000$) from the data generating process.

We compared the estimators from algorithms **empir** and **Gauss**, as well as from a logistic functional GLM, and Nadaraya–Watson estimation. The number T_n of principal components used to estimate ϱ was chosen using criterion (9), so that

$$T_n = \max\big\{j \ge 1 \colon \hat{\lambda}_j \ge m_n^{-1} \hat{\lambda}_1\big\}, \quad \text{with } \; m_n = 5\, n^{0.45},$$

where the exponent is in line with $\alpha = 1$ corresponding to the smooth nature of the responses. We introduce $\hat{\lambda}_1$ into the definition of T_n so that the criterion does not depend on the scale of the eigenvalues, yielding a more practicable way of choosing T_n. For $n = 1000$, this approximately covers 98% of the variance of the simulated curves in the sense of the PVE criterion (10). Naturally, less variance is covered in smaller sample sizes. For the logistic GLM, we used the approach suggested by [38] and took the truncated Karhunen–Loève expansion as the predictor. In order to keep the methods comparable, we used the same number T_n of principal components for our algorithms and for the functional GLM. We calibrated the bandwidth h for the Nadaraya–Watson estimator using leave-one-out cross-validation on each generated sample. The results in terms of the root mean squared error (RMSE) over the 1000 simulations are displayed in Figure 3. Because it is difficult to visualize this for the 50 different predictors, we present boxplots summarizing the RMSE of each method over all predictors Y_0^*. More details on the results for a variety of specific values of Y_0^* can also be found in Table 4 in Supplementary Material.

We observed that algorithms **empir** and **Gauss** exhibited similar predictive performance in both examples and over all sample sizes. These methods clearly outperformed functional logistic regression and Nadaraya–Watson estimation

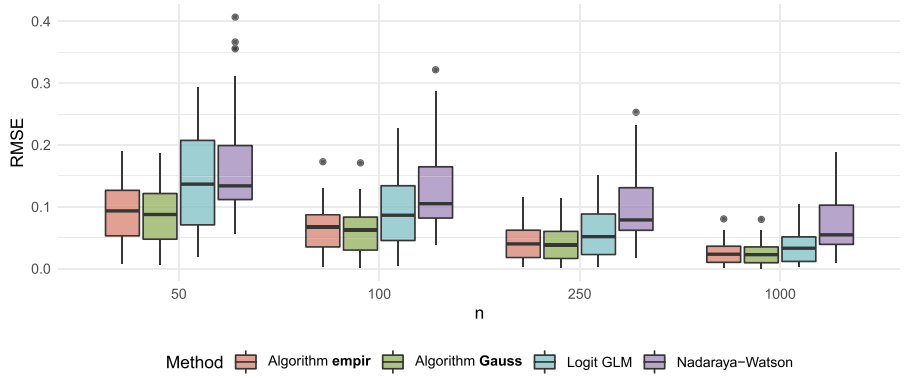

FIG 3. *RMSE of $\widehat{P}$ for 50 random predictors Y_0^* and 1000 independent simulations of samples of size $n \in \{50, 100, 250, 1000\}$ based on the estimators* ***empir***, ***Gauss***, *functional logistic regression, and Nadaraya–Watson estimators of the probability* $\mathrm{P}(\lambda(Y_1^* > \sqrt{50}) \leq 0.5|Y_0^*)$.

TABLE 1
RMSE for $\hat{\alpha}_p$, 5 different predictors and 1000 replications. We estimate $\hat{\alpha}_p$ such that $\mathrm{P}(\lambda(Y_1^* > \alpha_p) \leq 0.5|Y_0^*) = p$, where $p = 1 - n^{-1}$.

	$n = 50$ $p = 0.98$		$n = 100$ $p = 0.99$		$n = 250$ $p = 0.996$		$n = 1000$ $p = 0.999$	
Y_0^*	**empir**	**Gauss**	**empir**	**Gauss**	**empir**	**Gauss**	**empir**	**Gauss**
1	0.770	0.704	0.559	0.491	0.457	0.356	0.487	0.434
2	0.530	0.437	0.396	0.292	0.333	0.170	0.341	0.189
3	0.517	0.419	0.432	0.287	0.348	0.199	0.308	0.137
4	0.595	0.501	0.464	0.371	0.404	0.262	0.347	0.187
5	0.589	0.480	0.470	0.355	0.378	0.247	0.336	0.144

in estimating the conditional probability of level sets. The proposed methods achieved a similar mean squared error in this case to functional logistic regression with about a quarter of the sample size. The performance of the Nadaraya–Watson estimator was poor compared to the other methods considered in both cases and varied strongly depending on the predictor Y_0^*.

Although the estimator **Gauss** performs similarly to **empir** in the above example, it can be expected that **empir** runs into problems when $\mathrm{P}(Y \in A|X)$ is close to 0 or 1, since **empir** only uses the n estimated model residuals for estimation, whereas in producing the estimator **Gauss**, one can generate a sufficiently large Monte-Carlo sample to give a non-degenerate estimate of these probabilities, which can be expected to be accurate if the Gaussian assumption is plausible. To highlight this, we present the results of a short simulation study in which for a probability $p_n = 1 - 1/n$, we aimed to estimate α_p using **empir** and **Gauss** such that $\mathrm{P}(\lambda(Y_1^* > \alpha_{p_n}) \leq 0.5|Y_0^*) = p_n$. *This problem is hence related to the Value-at-Risk estimation.* We compared the RMSE of $\hat{\alpha}_{p_n}$ from the two algorithms for 50 different realizations of the predictor Y_0^* that were simulated from the same data generating process. We note that the value of α_p

varies between 7.26 and 11.77, depending on Y_0^* and p_n. In Table 1, we present the results from a subset of five predictors Y_0^* that were representative of the variability observed in the simulated series. The results show that **Gauss** outperforms **empir** in all cases, and the relative advantage increases with sample size, as expected. If we look at the results for all 50 predictors, RMSE of $\hat{\alpha}_p$ decreases by about 15% for $n = 50$, 22% for $n = 100$, 35% for $n = 250$ and 42% for $n = 1000$. This gives some indication of the difference in performance that can be expected between the two methods in forecasting extreme quantiles or events whenever the Gaussian assumption is plausible.

Following the suggestions of one reviewer, we also compared our results to a naive bootstrap approach. Here we use the estimator $\hat{\varrho}_n$ obtained from the original sample to generate bootstrap samples $S_n^{(b)} = \{X_i, Y_i^{(b)}\}_{1 \le i \le n}$, where $Y_i^{(b)} = \hat{\varrho}_n X_i + \delta_i^{(b)}$ with $\delta_i^{(b)}$ bootstrapped from the residuals $\{\hat{\varepsilon}_k\}_{1 \le k \le n}$. We apply Algorithm 1 to each bootstrap sample $S_n^{(b)}$, obtaining $\widehat{P}_n^{\mathrm{E}(b)}$. These estimators are then aggregated to calculate $\widehat{P}_n^{\mathrm{B}}(Y \in A|X) := \frac{1}{B}\sum_b \widehat{P}_n^{\mathrm{E}(b)}(Y \in A|X)$. When we do the simulation exercise related to Figure 3, we obtain results which are almost identical to **empir**. In the VaR scenario described in Table 1, the bootstrap is less accurate than **Gauss**, but it seems to be slightly favourable compared to **empir** at the price of much higher computational costs. A theoretical investigation of such a bootstrap approach is out of the scope of this paper.

6.3. Construction of prediction sets

To assess the performance of our method in the construction of prediction sets, we compare to Paparoditis and Shang [40]. We use the same data generating process as in the previous section and generate synthetic PM_{10} data. We fit an FAR(1) model to each simulated sample where we chose the truncation parameter T_n using the PVE criterion (10) with $v = 0.85$. This is the same value as used in [40]. Following the method proposed in [40] and as described above, we constructed uniform prediction sets to forecast each series 1-step ahead, with nominal coverage probabilities of 80% and 95%. The training set was initially set to be the first 80% of each sample, and then after each 1-step ahead prediction, the training set was increased by one observation, and subsequent 1-step ahead predictions were produced until the last 20% of the sample was exhausted. This process was then repeated independently for 1000 simulated samples, and for each sample size $n \in \{100, 200, 400, 800\}$. In line with our Remark 1, we used the normalization factor $n - 1 - T_n$ for $\widehat{\Gamma}_{\varepsilon,n}$ (one extra degree of freedom is subtracted since we also include an intercept term).

The results of this simulation are summarized in Table 2 in terms of empirical coverage probabilities, as well as the mean interval scores as computed in [40]. The mean interval scores take into account the *area* of the prediction band, which should be small, as well as the *excess* of functions that do not stay within the bounds of the band, with smaller scores suggesting better prediction bands. We refer to [22] for details.

TABLE 2
*Empirical coverage probabilities and mean interval score $\bar{S}_{1-\alpha}$ of uniform prediction bands for the data generating process in Section 6.2 calculated via **empir**, **Gauss**, as well as using the method of Paparoditis and Shang [40].*

Nominal coverage	n	**empir**		**Gauss**		*P., S. (2021)*	
		coverage	score	coverage	score	coverage	score
80%	100	0.7267	6.0832	0.7534	6.2094	0.8325	6.0685
	200	0.7649	6.1513	0.7784	6.2168	0.8522	6.1236
	400	0.7872	6.1772	0.7936	6.2116	0.8625	6.1477
	800	0.7892	6.1896	0.7928	6.2074	0.8623	6.1551
95%	100	0.9004	7.4971	0.9270	7.7644	0.9471	7.5692
	200	0.9290	7.6262	0.9396	7.7698	0.9574	7.6527
	400	0.9403	7.6817	0.9454	7.7607	0.9604	7.7009
	800	0.9444	7.7069	0.9475	7.7529	0.9603	7.7151

In general, all methods produced quite similar score values, but we see that **empir** and Paparoditis and Shang [40] performed slightly better than **Gauss**. In turn, the coverage rates of **Gauss** were overall closest to the nominal level. The method of Paparoditis and Shang [40] performed best for the smallest considered sample size. However, if we implement the method as suggested in [40], it exhibited notable bias for larger sample sizes, which is particularly pronounced when the nominal coverage is 80%. This bias did not seem to decrease when the sample size was increased. Overall, the method **empir** appeared to be the most reliable in this simulation when n is not too small. The methods **empir** and **Gauss** also enjoy vastly improved run-times compared to Paparoditis and Shang [40]. For example, when $n = 100$, using version 4.1.2 of `R` on a computer with Intel Core i5-7500 processor, calculating a single uniform prediction band took 8 ms with **empir**. With **Gauss** and using a Monte-Carlo sample of $M = 10\,000$ to compute the Gaussian quantiles, it took about 240 ms. The method of Paparoditis and Shang [40] took roughly 240 seconds. This significant difference in runtime may be explained by the fact that the algorithm of Paparoditis and Shang [40] draws $B = 1000$ bootstrap samples of the entire time series to construct the prediction band while we avoid the bootstrap completely.

6.4. Functional quantile regression

In this application, we compare to the data analysis of [45]. As in our previous example, this involves the functional time series of daily square-root transformed PM_{10} concentration curves constructed by smoothing half-hourly measurements of PM_{10}.

The goal of the analysis is to compare forecasts of the quantiles of the maximum values $M_t = \max_{u\in[0,1]} Y_t(u)$ (this being a prime example of a curve feature), where $Y_t(u)$ is the transformed PM_{10} curve on day t at intraday time u. As the covariate the curve Y_{t-1} is used. Now we model the relationship between (Y_t, Y_{t-1}) by a FAR(1) process and apply the method **empir** to estimate the conditional quantile of M_t. We select the truncation parameter T_n

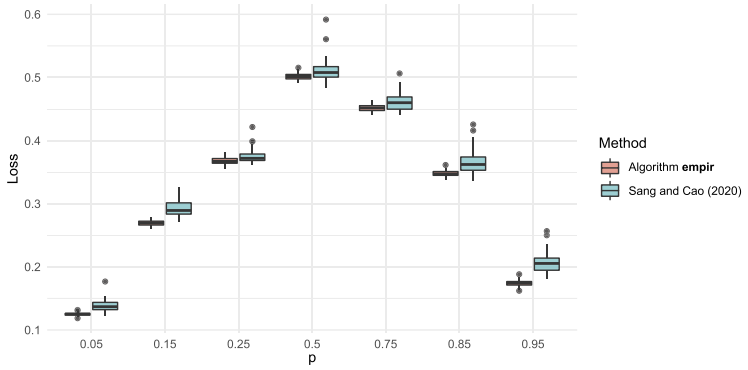

FIG 4. *Performance of* ***empir*** *compared to the functional single-index quantile regression model proposed by Sang and Cao [45]. The prediction error is compared using 5-fold cross-validation on 50 random splits of the* PM_{10} *data set.*

in order to explain 98% of the variance in the variables Y_t since for this fixed sample size, tuning T_n by an asymptotic criterion is not meaningful. At a quantile level τ, we compared these methods by 5-fold cross-validation, using the check function loss (13), and evaluate the error in the held out sample by the sum of $\rho_\tau\big(Y_i - \hat{Q}_\tau(Y|X = X_i)\big)$. We did this for seven different quantile levels $p \in \{0.05, 0.15, 0.25, 0.50, 0.75, 0.85, 0.95\}$. The experiment was repeated on 50 random splits of the data set. The results are displayed in Figure 4. In 87.4% of the cases, **empir** outperformed the functional single-index quantile regression model of [45] in terms of the loss considered. This advantage was smaller for central quantiles and became more apparent for the more extreme quantiles.

6.5. Spanish electricity price data

We now return to the Spanish electricity price data that we gave as an introductory example. Recall that these data are comprised of hourly electricity prices, demand, and wind energy production in Spain over the period from 2014 to 2019, which have been transformed into functional data objects by projection on 18 twice differentiable B-splines. These data are illustrated in Figure 1. We take as the goal of this analysis to compare estimates for the conditional probability that the price curves will lie in specified level sets, given the covariates of demand, and wind energy production. Since real data evidently do not exactly follow the model assumptions, the performance of the respective approaches can be used to compare their robustness to violations of the model assumptions.

In order to compare the various methods for doing this, we split the data into a training and testing set by randomly taking four months from each year, and assigning them to the test set, which created a 2:1 split between the training and testing set. Since we used 6 years of this data, the training set thus consists of 1453 days, and the test set consists of 731 days. Let Z_t denote one of

functional variables electricity price, demand or wind energy production. Then these variables were deseasonalized as follows: $\widetilde{Z}_t = Z_t - Z_t^{(Y)} - Z_t^{(W)}$, where $Z_t^{(Y)}$ is the yearly seasonality obtained by taking the mean for each day of the year and smoothing the result using a rolling mean with a window size of 21 days. $Z_t^{(W)}$ is the weekly seasonality that is estimated as the mean for each day of the week. For the wind curves, no weekly seasonality was removed. In order to employ the methods **empir** and **Gauss**, we fit the FARX(7) model described in (2) using the estimator introduced in Section 4 with the data in the training set. The truncation parameter T_n was again chosen in order to explain 98% of the variance of the covariates.

In order to compare the estimated conditional probabilities to the realized outcomes on the test set, we used the cross-entropy measure. This is a popular loss function in classification problems; see Murphy [39, Section 2.8]. Given the realizations $y_i = \mathbb{1}\{Y_i \in A\}$, $i \in \{1, ..., N\}$, and corresponding estimated conditional probabilities $\hat{p}_i = \widehat{P}(Y_i \in A|X_i)$ in the testing set of size N, empirical cross-entropy on the test set is defined as

$$-\frac{1}{N}\sum_{k=1}^{N}[y_k \log(\hat{p}_k) + (1 - y_k)\log(1 - \hat{p}_k)].$$

This is closely related to the deviance in logistic regression models. Accordingly, small values of the empirical cross-entropy indicate a good predictive power.

We considered level sets of the form $A_{\alpha,z}$ (as in Example 1) for various values of α and z. These values were chosen so that they reflect the range of the response Y. We calculated the cross-entropy on the test set of estimates of $\mathrm{P}(Y \in A|X)$ using the method **empir**, as well as for functional logistic regression, which was estimated using the same covariates (and PVE criterion) as those considered in generating the estimator in **empir**, as well as functional Nadaraya–Watson estimation with a Gaussian kernel with the predictors *demand*, *wind* and *lagged price*, and the bandwidth parameters were selected using leave-one-out cross-validation on the training set. We do not present the results for the method **Gauss**, as the results are again very similar to the method **empir**. This is in spite of the fact that the model residuals do not appear to be normally distributed, according to a Jarque–Bera type normality test for functional data; see, e.g., [25]. The estimated cross entropies on the test set for part of the sets A considered are presented in Table 3. The smallest value in each cell is marked in bold font. The full table can be found in Supplementary Material.

The method **empir** achieved lower values of cross-entropy on the test set compared to the competing methods for most combinations of α and z. **empir** had higher estimated cross-entropy in one case compared to functional logistic regression, and two cases compared to functional Nadaraya–Watson estimation.

Acknowledgments

Parts of the results presented in this work were computed using the high-performance-computing resources of the Graz University of Technology IT Ser-

TABLE 3

The cross-entropy of the estimated conditional probability $\mathrm{P}(\lambda(Y > \alpha) \leq z)$ *for different values* α *and* z*, evaluated on the test set. The comparison value GLM is a logit regression model with the same predictors. N–W is the Nadaraya–Watson estimator. The smallest value in each cell is marked in bold font, and any apparent ties are merely a result of the rounding to two digits.*

	α	30	35	40	45	50	55	60	65	70
	empir	**0.03**	**0.06**	**0.10**	**0.13**	**0.16**	**0.24**	0.23	**0.19**	**0.16**
$z = 0$	GLM	0.11	0.17	0.23	0.36	0.30	0.24	**0.22**	0.21	0.24
	N–W	0.05	0.10	0.20	0.21	0.22	0.31	0.29	0.27	0.26
	empir	**0.05**	**0.10**	**0.12**	**0.13**	**0.20**	**0.17**	**0.15**	**0.13**	**0.09**
$z = \frac{1}{3}$	GLM	0.25	0.15	0.26	0.23	0.23	0.17	0.23	0.20	0.39
	N–W	0.10	0.16	0.23	0.23	0.26	0.22	0.24	0.22	0.14
	empir	**0.09**	**0.11**	**0.14**	**0.19**	**0.22**	**0.18**	**0.12**	**0.08**	0.03
$z = \frac{2}{3}$	GLM	0.12	0.18	0.27	0.26	0.25	0.24	0.20	0.40	0.12
	N–W	0.16	0.20	0.23	0.26	0.28	0.25	0.19	0.13	**0.02**

vices. The authors wish to thank Mario Lang for his kind support, and Han Lin Shang for providing an implementation of the method described in Paparoditis and Shang [40].

Supplementary Material

Conditional distribution in functional regression problems
(doi: 10.1214/22-EJS2067SUPP; .pdf). Supplementary Material contains proofs of all the mathematical statements, along with additional simulation results from Section 6.

References

[1] AUE, A., NORINHO, D. D. and HÖRMANN, S. (2015). On the prediction of stationary functional time series. *Journal of the American Statistical Association* **110** 378–392. MR3338510

[2] AZAIS, J.-M. and WSCHEBOR, M. (2009). *Level sets and extrema of random processes and fields*. Wiley, Hoboken, NJ. MR2478201

[3] BERLINET, A., ELAMINE, A. and MAS, A. (2011). Local linear regression for functional data. *Annals of the Institute of Statistical Mathematics* **63** 1047–1075. MR2822967

[4] BILLINGSLEY, P. (1999). *Convergence of probability measures*. Wiley. MR1700749

[5] BOENTE, G., BARRERA, M. S. and TYLER, D. E. (2014). A characterization of elliptical distributions and some optimality properties of principal components for functional data. *Journal of Multivariate Analysis* **131** 254–264. MR3252648

[6] BOSQ, D. (2000). *Linear Processes in Function Spaces*. Springer. MR1783138

[7] Bulinskaya, E. V. (1961). On the Mean Number of Crossings of a Level by a Stationary Gaussian Process. *Theory of Probability & Its Applications* **6** 435–438.
[8] Chen, K. and Müller, H.-G. (2012). Conditional quantile analysis when covariates are functions, with application to growth data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74** 67–89. MR2885840
[9] Chen, K. and Müller, H.-G. (2014). Modeling Conditional Distributions for Functional Responses, With Application to Traffic Monitoring via GPS-Enabled Mobile Phones. *Technometrics* **56** 347–358. MR3238072
[10] Chernozhukov, V., Fernández-Val, I. and Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica* **78** 1093–1125. MR2667913
[11] Chiou, J.-M., Müller, H.-G. and Wang, J.-L. (2004). Functional response models. *Statistica Sinica* 675–693. MR2087968
[12] Choi, H. and Reimherr, M. (2016). A geometric approach to confidence regions and bands for functional parameters. arXiv:1607.07771. MR3744720
[13] Crambes, C., Hilgert, N. and Manrique, T. (2016). Estimation of the noise covariance operator in functional linear regression with functional outputs. *Statistics and Probability Letters* **113** 7–15. MR3480388
[14] Crambes, C. and Mas, A. (2013). Asymptotics of prediction in functional linear regression with functional outputs. *Bernoulli* **19** 2627–2651. MR3160566
[15] Dette, H., Kokot, K. and Aue, A. (2020). Functional data analysis in the Banach space of continuous functions. *The Annals of Statistics* **48** 1168–1192. MR4102692
[16] Fan, J. and Müller, H.-G. (2021). Conditional Distribution Regression For Functional Responses. *Scandinavian Journal of Statistics* **49** 502–524. MR4428494
[17] Fernández de Castro, B., Guillas, S. and González Manteiga, W. (2005). Functional Samples and Bootstrap for Predicting Sulfur Dioxide Levels. *Technometrics* **47** 212–222. MR2188081
[18] Ferraty, F., Van Keilegom, I. and Vieu, P. (2012). Regression when both response and predictor are functions. *Journal of Multivariate Analysis* **109** 10–28. MR2922850
[19] Ferraty, F. and Nagy, S. (2022). Scalar-on-function local linear regression and beyond. *Biometrika* **109** 439–455. MR4430967
[20] Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media. MR2229687
[21] Franke, J. and Nyarige, E. G. (2019). A residual-based bootstrap for functional autoregressions. arXiv:1905.07635.
[22] Gneiting, T. and Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* **102** 359–378. MR2345548
[23] Goldsmith, J., Greven, S. and Crainiceanu, C. (2013). Corrected con-

fidence bands for functional data using principal components. *Biometrics* **69** 41–51. MR3058050

[24] González, J. P., Muñoz San Roque, A. M. S. and Pérez, E. A. (2018). Forecasting Functional Time Series with a New Hilbertian ARMAX Model: Application to Electricity Price Forecasting. *IEEE Transactions on Power Systems* **33** 545–556. MR3714402

[25] Górecki, T., Hörmann, S., Horváth, L. and Kokoszka, P. (2018). Testing normality of functional time series. *Journal of Time Series Analysis* **39** 471–487. MR3819053

[26] Hörmann, S. and Kidziński, Ł. (2015). A note on estimation in Hilbertian linear models. *Scandinavian journal of statistics* **42** 43–62. MR3318024

[27] Hörmann, S. and Kokoszka, P. (2010). Weakly dependent functional data. *The Annals of Statistics* **38** 1845–1884. MR2662361

[28] Hörmann, S., Kuenzer, T. and Rice, G. (2022). Supplement to "Estimating the conditional distribution in functional regression problems". MR2662361

[29] Hyndman, R. J. and Shang, H. L. (2009). Forecasting functional time series. *Journal of the Korean Statistical Society* **38** 199–211. MR2750314

[30] Hyndman, R. J. and Shang, H. L. (2020). ftsa: Functional Time Series Analysis R package version 6.0.

[31] Imaizumi, M. and Kato, K. (2018). PCA-based estimation for functional linear regression with functional responses. *Journal of Multivariate Analysis* **163** 15–36. MR3732338

[32] Ivanescu, A., Staicu, A. M., Scheipl, F. and Greven, S. (2015). Penalized function-on-function regression. *Computational Statistics* **30** 539–568. MR3357075

[33] Kato, K. (2012). Estimation in functional linear quantile regression. *The Annals of Statistics* **40** 3108–3136. MR3097971

[34] Kudraszow, N. L. and Vieu, P. (2013). Uniform consistency of kNN regressors for functional variables. *Statistics & Probability Letters* **83** 1863–1870. MR3069890

[35] Liebl, D. and Reimherr, M. (2019). Fast and fair simultaneous confidence bands for functional parameters. arXiv:1910.00131. MR4365792

[36] Mas, A. (2007). Weak convergence in the functional autoregressive model. *Journal of Multivariate Analysis* **98** 1231–1261. MR2326249

[37] Mousavi, S. N. and Sørensen, H. (2017). Multinomial functional regression with wavelets and LASSO penalization. *Econometrics and Statistics* **1** 150–166. MR3669994

[38] Müller, H. G. and Stadtmüller, U. (2005). Generalized functional linear models. *The Annals of Statistics* **33** 774–805.

[39] Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective. Adaptive Computation and Machine Learning series*. MIT Press.

[40] Paparoditis, E. and Shang, H. L. (2021). Bootstrap Prediction Bands for Functional Time Series. *Journal of the American Statistical Association*. MR3010895

[41] Pumo, B. (1999). Prediction of Continuous Time Processes by C[0,1]-

Valued Autoregressive Process. *Statistical Inference for Stochastic Processes* **1** 297–309. MR2797138

[42] Ramsay, J. O., Graves, S. and Hooker, G. (2020). fda: Functional Data Analysis R package version 5.1.9.

[43] Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*, 2nd ed. ed. Springer. MR2168993

[44] Ruiz-Medina, M. D. and Álvarez-Liébana, J. (2019). Strongly consistent autoregressive predictors in abstract Banach spaces. *Journal of Multivariate Analysis* **170** 186–201. MR3913035

[45] Sang, P. and Cao, J. (2020). Functional single-index quantile regression models. *Statistics and Computing* 1–11. MR4108676

[46] Takács, L. (1995). On the Local Time of the Brownian Motion. *The Annals of Applied Probability* **5**. MR1359827

[47] Talagrand, M. (2014). *Upper and lower bounds for stochastic processes: modern methods and classical problems* **60**. Springer Science & Business Media. MR3184689

[48] Vilar, J. M., Cao, R. and Aneiros, G. (2012). Forecasting next-day electricity demand and price using nonparametric functional methods. *International Journal of Electrical Power & Energy Systems* **39** 48–55. MR2903632

[49] Wang, J.-L., Chiou, J.-M. and Müller, H.-G. (2016). Functional Data Analysis. *Annual Review of Statistics and Its Application* **3** 257–295.

[50] Xiong, S. and Li, G. (2008). Some results on the convergence of conditional distributions. *Statistics & Probability Letters* **78** 3249–3253. MR2479485

[51] Yao, F., Sue-Chee, S. and Wang, F. (2017). Regularized partially functional quantile regression. *Journal of Multivariate Analysis* **156** 39–56. MR3624684