# ON THE OPTIMAL RECONSTRUCTION OF PARTIALLY OBSERVED FUNCTIONAL DATA

By Dominik Liebl and Alois Kneip

*University of Bonn*

We propose a new prediction functional that aims to recover the missing parts of a function given the observed parts. We deviate from the approaches in the literature by building explicitly upon the very particular nature of this prediction problem: a response function and a predictor function that are both coming from the same function. It is shown that this new prediction functional is optimal among a very large class of functionals which contains the regression functionals considered so far, but also any other linear functionals that is reasonable from a statistical viewpoint. Our estimation theory allows for autocorrelated functional data and considers the practically relevant situation where the $n$ functions are only observed at $m$ discretization points. We derive uniform rates of consistency for our nonparametric estimation procedures using a double asymptotic ($n \to \infty, m \to \infty$) that allows investigate all data scenarios from almost sparse to dense functional data. The finite sample properties are investigated through simulations and a real data application.

**1. Introduction.** Our work is motivated by a data set from energy economics which is shown in Figure 1. The data consists of partially observed price functions. Practitioners use these functions, for instance, to do comparative statics, i.e., a ceteris-paribus analysis of the price effects with respect to changes in electricity demand (cf. Weigt 2009 and Hirth 2013). Though, the possibilities of such an analysis are limited by the extend to which we can observe the price functions. This motivates the aim of our work which is to develop a prediction procedure that allows to recover the total functions from their partial observations.

Let $X_1, \ldots, X_n$ be an iid sample of random functions, where each function $X_i$ is an element of the separable Hilbert space $\mathbb{L}^2([a,b])$ with $[a,b] \subset \mathbb{R}$. To save on notation, let us initially consider, without loss of generality, centered random functions, i.e., $\mathbb{E}(X_i(u)) = 0$ for all $u \in [a,b]$. We denote the observed

*AMS 2000 subject classifications:* Primary62M20, 62H25, 62G05, 62G08, 62M10

*Keywords and phrases:* functional data analysis, functional linear prediction, functional principal components, time series analysis

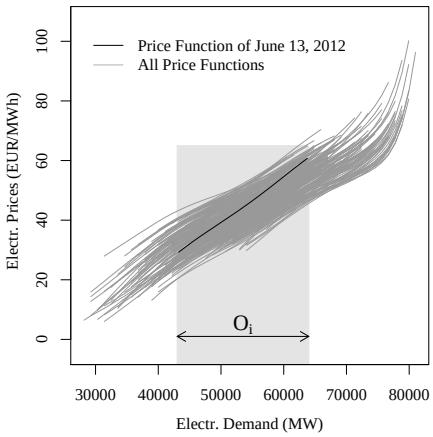


FIG 1. *Partially observed electricity price functions* $X_i^O(u)$ *with* $u \in O_i \subseteq [a,b]$.

and missing parts of $X_i$ by $X_i^{O_i}$ and $X_i^{M_i}$, where

$$\begin{aligned} X_i^{O_i}(u) &:= X_i(u) \quad \text{for} \quad u \in O_i \subseteq [a,b] \\ X_i^{M_i}(u) &:= X_i(u) \quad \text{for} \quad u \in M_i = [a,b] \setminus O_i, \end{aligned}$$

and where $O_i = [A_i, B_i] \subseteq [a,b]$, with $B_i - A_i > 0$ almost surely, is a random subinterval as illustrated in Figure 1. In what follows we write "$O$" and "$M$" in order to denote a given realization of $O_i$ and $M_i$. Furthermore, we use the following short hand notation for conditioning on $O_i$ and $M_i$:

$$\begin{aligned} X_i^O(u) &:= X_i^{O_i}(u) \,|(O_i = O) \\ X_i^M(u) &:= X_i^{M_i}(u)|(M_i = M) \end{aligned}$$

As usually, the inner product of $\mathbb{L}^2(O)$ is denoted by $\langle x, y\rangle_2 = \int_O x(u)y(u)du$ and the induced norm by $||x||_2^2 = \langle x, x\rangle_2$ for all $x, y \in \mathbb{L}^2(O)$.

Our object of interest is the following linear prediction problem:

$$X_i^M(u) = L_u(X_i^O) + Z_i(u), \quad u \in M, \tag{1}$$

which aims to reconstruct (i.e., predict) the unobserved value $X_i^M(u) \in \mathbb{R}$, for any $u \in M$, given the partial observation $X_i^O$. Our objective is to identify the optimal linear prediction functional $L_u : \mathbb{L}^2(O) \to \mathbb{R}$ which minimizes the mean squared prediction error $\mathbb{E}[(X_i^M(u) - L_u(X_i^O))^2]$ at any given $u \in M$.

Note that the prediction problem in Eq. (1) is a rather particular one, since the usually considered "linear regression functionals", i.e.,

$$L_u(X_i^O) = \int_O \beta_u(v) X_i^O(v) dv, \quad \text{with} \quad \beta_u \in \mathbb{L}^2(O) \tag{2}$$

do *not* belong to the optimal solution of the prediction problem. In order to see this, let us consider the "last observed"(="first missing") points, i.e, the so-called boundary points[1] $\vartheta \in \partial M$ of $M$. For any optimal prediction functional $L_u$, it must hold that the "first predicted value", $L_\vartheta(X_i^O)$, connects with the "last observed value", $X_i^O(\vartheta)$, i.e., that

$$X_i^O(\vartheta) = L_\vartheta(X_i^O) \quad \text{for all} \quad \vartheta \in \partial M.$$

Though, there is no hope to find a slope function $\beta_\vartheta \in \mathbb{L}^2(O)$ which fulfills the equation $X_i^O(\vartheta) = \int_O \beta_\vartheta(v) X_i^O(v) dv$ except in a degenerated manner, namely, if $\beta_\vartheta$ equals the Dirac-$\delta$ function $\delta_\vartheta(v) = 1_{(v=\vartheta)}$. However, Dirac-$\delta$ functions in $\mathbb{L}^2(O)$ are equivalent to the zero function $z(v) = 0$, since $||\delta_\vartheta - z||_2 = 0$, which makes it impossible to identify the optimal prediction functional $L_u$ within the class of all linear regression functionals.

This insight motivated our search for a new prediction functional beyond the commonly used regression functionals. By starting heuristically, we derive in Section 2 the following optimal linear prediction functional:

$$\mathcal{L}_u(X_i^O) = \sum_{k=1}^{\infty} \frac{\langle \phi_k^O, X_i^O \rangle_2 \, \langle \phi_k^O, \gamma_u \rangle_2}{\lambda_k^O}, \quad u \in M, \tag{3}$$

where $(\phi_k^O, \lambda_k^O)_{k \geq 1}$ denote the pairs of eigenfunctions and eigenvalues of the covariance operator $\Gamma^O(x)(u) = \int \gamma^O(u,v) x(v) dv$ with $x \in \mathbb{L}^2(O)$ and covariance function $\gamma^O(u,v) = \mathrm{Cov}(X_i^O(u), X_i^O(v))$, and where $\gamma_u(v) = \mathrm{Cov}(X_i^M(u), X_i^O(v))$.

The general structure of Eq. (3) is well known in the literature. For instance, the classic references for the functional linear regression model, Cai and Hall (2006) and Hall and Horowitz (2007), propose to use just the empirical and truncated version of Eq. (3) in order to estimate the regression slope function $\beta_u$. However, we derive Eq. (3) under a fundamentally different setup, since we do not start with assuming that $L_u(X_i^O)$ is a regression functional. By contrast to Cai and Hall (2006) and Hall and Horowitz (2007), it is therefore generally impossible to rewrite our prediction functional $\mathcal{L}_u$ in Eq. (3) in the form a regression functional $\int_O \beta_u(v) X_i^O(v) dv$ as in Eq. (2).

[1]The boundary $\partial M$ of a subset $M$ is defined as $\partial M := \overline{M} \cap \overline{O}$, where $\overline{M}$ and $\overline{O}$ denote the closures of the subsets $M$ and $O$.

In Section 2, we show that our proposed prediction functional is indeed the optimal prediction functional. The difficulty is to define a sensible and sufficiently large class of linear functionals against which we can compare our prediction functional. For this we define the class of "prediction functionals", which contains all linear functionals $L_u$ for which $\mathbb{V}(L_u(X_i^O)) < \infty$ for all $u \in M$. The class of prediction functionals is a much larger class than the class of regression functionals and contains the latter as a special case. Furthermore, we develop a framework which facilitates working with this new class of prediction functionals.

Besides our theoretical contributions on the optimal solution of the prediction problem in Eq. (1), we also address the problem of estimation. Here it is advantageous not to estimate the prediction functional $\mathcal{L}_u(X_i^O)$ from Eq. (3) directly, but to estimate its difference from the "last observed value", $X_i^O(\vartheta)$, i.e.,

$$\mathcal{D}_u(X_i^O) := (\mathcal{L}_u(X_i^O) - X_i^O(\vartheta_u)), \quad u \in M, \tag{4}$$

where $\vartheta_u$ denotes the boundary point $\vartheta \in \partial M$ that is closest to the considered $u \in M$, i.e., $\vartheta_u = A_i$ if $|A_i - u| < |B_i - u|$ and $\vartheta_u = B_i$ else.

By contrast to $\mathcal{L}_u(X_i^O)$, the difference functional $\mathcal{D}_u(X_i^O)$ can be a regression functional which implies the possibility faster convergence rates. Given an estimate, say $\hat{\mathcal{D}}_u(X_i^O)$, one can always compute the corresponding estimate for the prediction functional by $\hat{\mathcal{L}}_u(X_i^O) = X_i^O(\vartheta_u) + \hat{\mathcal{D}}_u(X_i^O)$.

Motivated by our real data application, we also consider the challenging case where none of observed functions $X_i^O$ cover the total domain such that only parts of the covariance function are estimable. To solve this problem, we suggest an iterative prediction algorithm that allows to reconstruct the functions under this challenging setup.

For estimating the mean and covariance functions – the basic ingredients of our prediction functional – we suggest using Local Linear Kernel (LLK) estimators. These LLK estimators are commonly used in the context of sparse functional data (see, e.g., Yao, Müller and Wang, 2005a), though, we do *not* consider the case of sparse functional data. In the context of partially observed functional data, it is advisable to use LLK estimators for the mean and covariance functions, since these will guarantee smooth estimation results, which is not the case when using the empirical moment estimators for partially observed functions suggested by Kraus (2015).

Again motivated by our real data application, our estimation theory allows for an autocorrelated time series of functional data and considers the practically relevant case where the functions $X_i^O$ are not directly observed, but have to be revealed from $m$ many data points $(Y_{i1}, U_{i1}), \ldots, (Y_{im}, U_{im})$

with $Y_{ij} = X_i^O(U_{ij}) + \varepsilon_{ij}$, $i = 1, \ldots, n$, and $j = 1, \ldots, m$. Inspired by the work of Zhang and Wang (2016), we derive uniform rates of convergence under a double asymptotic ($n \to \infty, m \to \infty$) which allows to investigate all data scenarios from almost sparse to dense functional data. This leads to different convergence rates depending on the relative relative order of $m$ and $n$.

The case of partially observed functional data was first considered in Goldberg, Ritov and Mandelbaum (2014) and Kraus (2015). Though, Goldberg, Ritov and Mandelbaum (2014) consider the case of finite dimensional functional data and their results have well known counterparts in multivariate statistics. Significant and compelling work for the case of infinite dimensional functional data is done by Kraus (2015). He uses a classic operator based approach which is, however, nested by our functional based approach, since our results apply to *every* $u \in M$. For showing consistency, Kraus (2015) imposes the classic Hilbert-Schmidt assumption which excludes the optimal solution of his prediction problem. Hilbert-Schmidt operators are the natural counterparts to regression functionals and cannot be optimal by essentially the same reasoning as used above for regression functionals.

We do not consider the case of so-called fragmentary observations, where the fragments are so short that one initially has to extend the fragments (by some appropriate procedure) before estimating the mean and covariance functions. Important references in this literature are Delaigle and Hall (2013) and Delaigle and Hall (2016). These approaches rely on more restrictive assumptions than we do in order to deal with the additional lack of information in the data.

The rest of this paper is structured as follows: Section 2 introduces our prediction functional and contains the optimality result. Section 3 comprises our estimation procedure. The asymptotic results are presented in Section 4. Section 5 contains the simulation study, and Section 6 describes the iterative prediction algorithm. The real data application can be found Section 7 and Section 8 concludes. All proofs can be found in Appendix A of the supplemental paper.

**2. Optimal prediction of partially observed functions.** Let our basic setup be as described in the second paragraph of Section 1. Let denote the mean function by $\mu(u) = \mathbb{E}(X_i(u))$ and the covariance function by $\gamma(u, v) = \mathrm{Cov}(X_i(u), X_i(v))$. In order to simplify the notation within this theoretical section, we assume centered random functions, i.e., $\mu(u) = 0$ for all $u \in [a, b]$. This is, of course, without loss of generality.

Let us first assume the ideal situation that we know $\gamma$. But if $\gamma(u, v)$ is

known over $[a,b]^2$, we also know $\gamma^O(u,v) := \gamma(u,v)$ over $(u,v) \in O^2$ and therefore $\gamma$ defines a unique covariance operator $\Gamma^O$ of $X_i^O$ as

$$\Gamma^O(x)(u) := \int_O \gamma^O(u,v)x(v)dv, \quad x \in \mathbb{L}^2(O),$$

where $\lambda_1^O \geq \lambda_2^O \geq \dots$ and $\phi_1^O, \phi_2^O, \dots$ denote eigenvalues and eigenfunctions of $\Gamma^O$, and where $(\phi_k^O)_{k\geq 1}$ forms an orthonormal basis system that spans the space $\mathbb{L}^2(O)$. The correspondingly defined covariance operator of $X_i$ is denoted by $\Gamma$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots$, and eigenfunctions $\phi_1, \phi_2, \dots$.

Any (centered) random function $X_t^O$ adopts then the well known Karhunen-Loéve (KL) representation

$$X_i^O(u) = \sum_{k=1}^{\infty} \xi_{ik}^O \phi_k^O(u), \quad u \in O, \tag{5}$$

with the pc-scores $\xi_{ik}^O := \langle X_i^O, \phi_k^O \rangle_2$, where $\mathbb{E}(\xi_{ik}^O) = 0$ and $\mathbb{E}(\xi_{ik}^O \, \xi_{il}^O) = \lambda_k^O$ for all $k = l$ and zero else. Equivalently, the KL representation of $X_i$ is given by $X_i(u) = \sum_{k=1}^{\infty} \xi_{ik}\phi_k(u)$, $u \in [a,b]$, with pc-scores $\xi_{ik} := \langle X_i, \phi_k \rangle_2$, where $\mathbb{E}(\xi_{ik}) = 0$ and $\mathbb{E}(\xi_{ik} \, \xi_{il}) = \lambda_k$ for all $k = l$ and zero else.

The properties of the KL representation imply that if $\lambda_k^O > 0$, the value $\phi_k^O(u)$ can be obtained as the slope-coefficient of a simple linear regression of $X_i^O(u)$ on the pc-score $\xi_{ik}^O$:

$$\phi_k^O(u) = \frac{\mathbb{E}(\xi_{ik}^O \, X_i^O(u))}{\lambda_k^O} = \frac{1}{\lambda_k^O} \int_O \phi_k^O(v)\gamma^O(u,v)dv, \quad u \in O, \tag{6}$$

where the second equality in Eq. (6) follows directly form the definition of the $k$th pc-score $\xi_{ik}^O$.

Eq. (6) can obviously be generalized for all $u \in [a,b]$ by regressing $X_i(u)$ on $\xi_{ik}^O$, which leads to the following extrapolated $k$th basis function:

$$\tilde{\phi}_k(u) := \frac{\mathbb{E}(\xi_{ik}^O \, X_i(u))}{\lambda_k^O} = \frac{1}{\lambda_k^O} \int_O \phi_k^O(v)\gamma(u,v)dv, \quad u \in [a,b], \tag{7}$$

if $\lambda_k^O > 0$, while we set $\tilde{\phi}_k(x) := 0$ if $\lambda_k^O = 0$. The second equality in Eq. (7) follows directly from the definition of the $k$th pc-score $\xi_{ik}^O$ and from the KL representation of $X_i$. Note that by construction $\tilde{\phi}_k(u) = \phi_k^O(u)$ for all $u \in O$. By contrast, one generally has that $\tilde{\phi}_k(u) \neq \phi_k(u)$ for $u \in M_i$. Furthermore, continuity of $\gamma$ implies continuity of $\tilde{\phi}_k$.

This leads to the definition of our prediction functional $\mathcal{L}_u(X_i^O)$ which is based on the basis functions $\tilde{\phi}_k$ and the pc-scores $\xi_{ik}^O$:

$$\tilde{X}_i(u) = \mathcal{L}_u(X_i^O) := \sum_{k=1}^{\infty} \xi_{ik}^O \, \tilde{\phi}_k(u), \quad u \in [a,b]. \tag{8}$$

*Remark.* Note that the predictive character of Eq. (8) takes only effect for values $u \in M$. For any value $u \in O$, however, Eq. (8) is merely the KL representation of $X_i^O(u)$, since $\tilde{\phi}_k(u) = \phi_k^O(u)$ for all $u \in O$. The latter property implies that the "first predicted value", $\mathcal{L}_\vartheta(X_i^O)$, connects with the "last observed value", $X_i^O(\vartheta)$, i.e., that $X_i^O(\vartheta) = \mathcal{L}_\vartheta(X_i^O)$ for all $\vartheta \in \partial M$, where $\partial M = \partial O$ by the definition boundary points.

2.1. *A unifying framework for prediction and regression functionals.* Before we consider the optimality properties of our proposed prediction functional $\mathcal{L}_u$ in Eq. (8), we need to define a sensible class of functionals against which we can compare our candidate. We cannot simply choose the usual class of regression functionals, since our prediction functional $\mathcal{L}_u$ does not belong to this class. This is most easily seen when focusing on some $u \in O$, where our prediction functional becomes an evaluation functional, i.e., $\mathcal{L}_u(X_i) = X_i(u)$ for any $u \in O$. Though, as argued in Section 1, such evaluation functionals cannot be represented using a $\mathbb{L}^2$ inner-product and therefore are not regression functionals.

In order to introduce a large enough class of linear functionals that includes our functional $\mathcal{L}_u$, we introduce the notion of "prediction functionals".

Definition 2.1 (Prediction functionals). *Let the (centered) random predictor function $X_i^O$ have a KL representation as in Eq. (5). We call every linear functional $L_u : \mathbb{L}^2(O) \to \mathbb{R}$ a "prediction functional with respect to $X_i^O$" if $\mathbb{V}(L_u(X_i^O)) < \infty$ at any given $u \in [a,b]$.*

It is important that a prediction functional, according to Def. 2.1, is specific to the relevant random predictor function $X_i^O$. This leads to a very large class of linear functionals, since the requirement that $\mathbb{V}(L_u(X_i^O)) < \infty$ for any $u \in [a,b]$ does not have to hold for every possible random variable in $\mathbb{L}^2(O)$, but only for the relevant predictor variable $X_i^O$. We demonstrate below that any regression functional is also a prediction functional according to Def. 2.1.

Though, in order to actually work with this class of prediction functionals, we need a useful representation. Such a representation is given in the following theorem.

THEOREM 2.1 (Representation of prediction functionals). *Let the (centered) random predictor function $X_i^O$ have a KL representation as in Eq. (5) and let $L_u : \mathbb{L}^2(O) \to \mathbb{R}$ be a "prediction functional with respect to $X_i^O$" according to Def. 2.1. Then there exits a unique (deterministic) function $\alpha_u \in H$ such that*

$$L_u(X_i^O) = \langle \alpha_u, X_i^O \rangle_H \quad \textit{almost surely,}$$

*where $H := \left\{ f \in \mathbb{L}^2(O) : \ ||f||_H^2 < \infty \right\}$ is a Hilbert space with inner product $\langle f, g \rangle_H := \sum_{k=1}^{\infty} \langle f, \phi_k^O \rangle_2 \, \langle g, \phi_k^O \rangle_2 / \lambda_k^O$ for all $f, g \in \mathbb{L}^2(O)$ and induced norm $||f||_H = \sqrt{\langle f, f \rangle_H}$.*
*The proof can be found in Appendix A of the supplemental paper.*

*Remark.* $H$ is the so-called reproducing kernel Hilbert space that takes the covariance kernel $\gamma^O(u,v) = \sum_{k=1}^{\infty} \lambda_k^O \phi_k^O(u) \phi_k^O(v)$ as its reproducing kernel, therefore, it holds that $H \subset \mathbb{L}^2(O)$. Though, observe that we do *not* restrict the functions $X_i^O$ to be elements of $H$ – only the "parameter functions" $\alpha_u$ have to be an element of $H$.

Theorem 2.1 provides us with a framework to analyze linear regression functionals and linear prediction functionals in a unified manner by using the inner-product $\langle ., . \rangle_H$ representation. In the following we compare both cases with each other and finally consider our prediction functional $\mathcal{L}_u$ within this framework.

***Linear Prediction (P).*** By Theorem 2.1 there exits for every linear prediction functional $L_u^P$ a unique $\alpha_u^P \in H$ such that $L_u^P(X_i^O) = \langle \alpha_u^P, X_i^O \rangle_H$. That is, we can write the functional linear prediction problem in Eq. (1) as

$$X_i(u) = \langle \alpha_u^P, X_i^O \rangle_H + Z_i(u),$$

where the only requirement on $\alpha_u^P$ is that $||\alpha_u^P||_H^2 = \sum_{k=1}^{\infty} \langle \alpha_u^P, \phi_k^O \rangle_2^2 / \lambda_k^O < \infty$; otherwise $L_u^P$ would not be "a prediction functional with respect to $X_i^O$", since $\mathbb{V}(L_u^P(X_i^O)) = ||\alpha_u||_H$ and $\mathbb{V}(L_u^P(X_i^O))$ is required to be finite.

***Linear Regression (R).*** For every linear regression functional $L_u^R$ there exists a unique $\beta_u^R \in \mathbb{L}^2(O)$ such that $L_u^R(X_i^O) = \langle \beta_u^R, X_i^O \rangle_2$. Since $\beta_u^R$ and $X_i^O$ are both elements of $\mathbb{L}^2(O)$, $\langle \beta_u^R, X_i^O \rangle_2 < \infty$ almost surely for any $u \in [a,b]$. Therefore, $\mathbb{V}(L_u^R(X_i^O)) = \mathbb{V}(\langle \beta_u^R, X_i^O \rangle_2) < \infty$ for any $u \in [a,b]$, which shows that every regression functional is also a prediction functional according to Def. 2.1. By Theorem 2.1 there exits then also a unique $\alpha_u^R \in H$ such that $L_u^R(X_i) = \langle \alpha_u^R, X_i^O \rangle_H$. The equation $L_u^R(X_i) = \langle \alpha_u^R, X_i^O \rangle_H = \langle \beta_u^R, X_i^O \rangle_2$ establishes a one-to-one relation between the parameter functions $\beta_u^R = \sum_{k=1}^{\infty} \langle \beta_u^R, \phi_k^O \rangle_2 \phi_k^O$ and $\alpha_u^R = \sum_{k=1}^{\infty} \langle \alpha_u^R, \phi_k^O \rangle_2 \phi_k^O$, which can be

expressed through their $\mathbb{L}^2$ coefficients by $\langle \beta_u^R, \phi_k^O \rangle_2 = \langle \alpha_u^R, \phi_k^O \rangle_2 / \lambda_k^O$.
To conclude, if one considers a regression context, we can write the functional linear prediction problem in Eq. (1) as

$$X_i(u) = \langle \alpha_u^R, X_i^O \rangle_H + Z_i(u),$$

but with the additional requirement (besides $||\alpha_u^R||_H < \infty$) that

$$\sum_{k=1}^{\infty} \left( \langle \alpha_u^R, \phi_k^O \rangle_2 / \lambda_k^O \right)^2 < \infty, \tag{9}$$

which follows from the one-to-one relation between the parameter functions $\beta_u^R$ and $\alpha_u^R$ and the requirement that $||\beta_u^R||_2^2 < \infty$. Observe that the requirement in Eq. (9) is substantially more restrictive than requiring that $||\alpha_u^R||_H < \infty$ as in the case of linear prediction functionals.

***The prediction functional*** $\mathcal{L}_u$. Considering that $\tilde{\phi}_k^O = \langle \gamma_u, \phi_k^O \rangle_2 / \lambda_k^O$, we can write our prediction functional $\mathcal{L}_u$ in Eq. (8) as

$$\mathcal{L}_u(X_i^O) = \langle \gamma_u, X_i^O \rangle_H, \quad \text{for any} \quad u \in [a, b], \tag{10}$$

where $\gamma_u(v) := \gamma(u, v)$ for any $v \in O$ and any $u \in [a, b]$. Theorem 2.2 below shows that $\mathbb{V}(\mathcal{L}_u(X_i^O)) = ||\gamma_u||_H < \infty$ for any $u \in [a, b]$, i.e., $\mathcal{L}_u$ is indeed a prediction functional according to Def. 2.1.

$\mathcal{L}_u(X_i^O)$ cannot be a regression functional which is easily seen by checking the requirement in Eq. (9): For instance, for any $u \in O$ we have that $\gamma_u = \gamma_u^O = \sum_{l=1}^{\infty} \lambda_l^O \phi_l^O(u) \phi_l^O$, which yields that $\sum_{k=1}^{\infty} (\langle \gamma_u^O, \phi_k^O \rangle_2 / \lambda_k^O)^2 = \sum_{k=1}^{\infty} (\phi_k^O(u))^2$ diverges.

***The difference functional*** $\mathcal{D}_u$. By contrast to $\mathcal{L}_u(X_i^O)$, its difference from the "last observed value", $X_i^O(\vartheta)$ can be a regression functional. In order to show this, let us focus on the relevant case of $u \in M$ and observe that we can always write our prediction functional $\mathcal{L}_u(X_i^O)$ as

$$\mathcal{L}_u(X_i^O) = X_i^O(\vartheta_u) + \mathcal{D}_u(X_i^O), \quad u \in M, \quad \text{with}$$

$$\mathcal{D}_u(X_i^O) := (\mathcal{L}_u(X_i^O) - X_i^O(\vartheta_u)) = \langle (\gamma_u - \gamma_{\vartheta_u}), X_i^O \rangle_H, \tag{11}$$

where $\vartheta_u$ denotes the boundary point $\vartheta \in \partial M$ that is closest to the considered $u \in M$, i.e., $\vartheta_u = A_i$ if $|A_i - u| < |B_i - u|$ and $\vartheta_u = B_i$ else.

The difference functional functional $\mathcal{D}_u(X_i^O)$ can be a regression functional, since its parameter function $(\gamma_u - \gamma_{\vartheta_u})(v)$, with $u \in M$ and $v \in O$, can be simple enough, such that the requirement in Eq. (9) is fulfilled.

A trivial example where $\mathcal{D}_u(X_i^O)$ is indeed a regression functional is when $X_i(u)$, with $u \in [0, 1]$, is a Brownian Motion. Let's say we observe the initial

trajectory on $O = [0, \vartheta]$ with $0 < \vartheta < 1$ and want to predict the value $X_i(u)$ at some point $u \in M = (\vartheta, 1]$. (Here we write $\vartheta$ instead of $\vartheta_u$ as the same $\vartheta$ applies to all $u \in M$.) By the definition of the covariance function of the Brownian Motion we have that $\gamma_u(v) = \min(u, v) = v = \min(\vartheta, v) = \gamma_\vartheta(v)$ for all $v \in O$, all $u \in M$, and all $0 < \vartheta < 1$. Therefore, $(\gamma_u - \gamma_\vartheta)(v) = 0$ for all $v, u$, and $\vartheta$ and the requirement in Eq. (9) is trivially fulfilled, since $\sum_{k=1}^{\infty}(\langle(\gamma_u - \gamma_\vartheta), \phi_k^O\rangle_2/\lambda_k^O)^2 = 0$.

This insight motivates us to estimate $\mathcal{L}_u(X_i^O)$ through estimating $\mathcal{D}_u(X_i^O)$; see Sections 3 and 4.

2.2. *Theoretical properties.* First of all, we address two very basic, but important properties of our prediction functional $\tilde{X}_i(u) = \mathcal{L}_u(X_i^O)$ in Eq. (8). Result (a) of the following theorem assures that $\tilde{X}_i(u)$ is a well-defined random function and result (b) assures the unbiasedness of $\tilde{X}_i(u)$.

THEOREM 2.2. *Let the (centered) random predictor function $X_i^O$ have a KL representation as in Eq. (5).*

*(a) $\tilde{X}_i = \mathcal{L}_u(X_i^O)$ in Eq. (8) has a continuous and finite variance function, i.e., $\mathbb{V}(\tilde{X}_i(u)) < \infty$ for all $u \in [a, b]$.*

*(b) $\tilde{X}_i(u)$ is unbiased in the sense that $\mathbb{E}(\tilde{X}_i(u)) = \mathbb{E}(X_i(u) - \mu(u)) = 0$ for all $u \in [a, b]$.*

*The proofs can be found in Appendix A of the supplemental paper.*

In the following we analyze the theoretical properties of the prediction error $Z_i(u) = X_i(u) - \tilde{X}_i(u)$ for all $u \in [a, b]$ with $\tilde{X}_i(u) = \mathcal{L}_u(X_i^O)$. A particular focus lies on the unobserved parts $u \in M_i$, since for the observed parts $v \in O$ we have by construction that $Z_i(v) = 0$; see also below of Eq. (8).

Result (a) in the following theorem shows that the prediction error $Z_i(v)$ is orthogonal to the predictor $X_i^O(u)$ for all $v \in [a, b]$ and all $u \in O$. This result serves as an auxiliary result for result (b) which shows that $\tilde{X}_i(u)$ is the L2-optimal linear prediction of the true $X_i(u)$ for all $u \in [a, b]$. Finally, result (c) allows us to identify cases where the random functions $X_i$ can be predicted without any prediction error.

THEOREM 2.3 (Optimal Linear Prediction). *The following results hold for any (centered) random function $X_i \in \mathbb{L}^2([a, b])$ with existing covariance function $\gamma$.*

*(a) For every $v \in O$ and $u \in M_i$ we have that*

$$\mathbb{E}\left(X_i^O(v)Z_i(u)\right) = 0 \quad \textit{and} \tag{12}$$

$$\mathbb{V}(Z_i(u)) = \mathbb{E}\left(Z_i(u)^2\right) = \gamma(u,u) - \sum_{k=1}^{\infty} \lambda_k^O \tilde{\phi}_k(u)^2. \tag{13}$$

*(b) For any linear functional $\ell_u : \mathbb{L}^2(O) \to \mathbb{R}$ that is a prediction functional with respect to $X_i^O$, according to Def. 2.1, we have that*

$$\mathbb{E}\left(\left(X_i(u) - \ell_u(X_i^O)\right)^2\right) \geq \mathbb{V}(Z_i(u)),\ u \in [a,b].$$

*(c) Let $X_i$ and $X_j$, with $i \neq j$, be independent and identically distributed Gaussian processes. Then the variance of the prediction error can be written as*

$$\mathbb{V}(Z_i(u)) = \frac{1}{2}\mathbb{E}\left(\mathbb{E}\left(\left(X_i(u) - X_j(u)\right)^2 \middle| X_i^O = X_j^O\right)\right),\ i \neq j, \tag{14}$$

*for all $u \in [a,b]$, where $X_i^O = X_j^O$ denotes that $X_i^O(v) = X_j^O(v)$ for all $v \in O$.*

*The proofs can be found in Appendix A of the supplemental paper.*

Result (b) of Theorem 2.3 does not make any specific smoothness assumptions on the process $X_i$. That is, even very rough and unstructured processes such as, for instance, the Brownian Motion can be optimally predicted.

*Example: Brownian Motion.* Let's say that we observe the initial part of a Brownian Motion trajectory $X_i(u) \in \mathbb{L}^2([0,1])$ for $u \in [0,\vartheta]$ with $0 < \vartheta < 1$ and aim to predict $X_i(v)$ for $v \in [\vartheta, 1]$. Trivially, the optimal prediction of the observed part is just the observed trajectory, i.e., $\tilde{X}_i^{opt}(u) = X_i^O(u)$ for all $u \in [0,\vartheta]$. Furthermore, from the basic properties of the Brownian Motion, we know that the optimal prediction $\tilde{X}_i^{opt}(v)$ for $v \in [\vartheta, 1]$ is given by the last observed value, i.e., by $X_i^O(\vartheta)$. Summing up, the optimal prediction is given by

$$\tilde{X}_i^{opt}(u) = \begin{cases} X_i^O(u), & \text{for all} \quad u \in [0,\vartheta] \\ X_i^O(\vartheta), & \text{for all} \quad u \in [\vartheta, 1]. \end{cases} \tag{15}$$

This is exactly the prediction result of our prediction functional $\mathcal{L}_u(X_i^O) = \langle \gamma_u, X_i^O \rangle_H$. Let us consider the case $u \in [0,\vartheta]$. As already stated above, $\mathcal{L}_u$ is an evaluation functional for any $u \in O$, i.e., here $\mathcal{L}_u(X_i^O) = X_i^O(u) = \tilde{X}_i^{opt}(u)$ for any $u \in [0,\vartheta]$. This follows directly from $\mathcal{L}_u(X_i^O) = \langle \gamma_u, X_i^O \rangle_H$

by considering that $\gamma_u(v) = \gamma_u^O(v) = \sum_{k=1}^{\infty} \lambda_k^O \phi_k^O(u)\phi_k^O(v)$ if both $u, v \in [0, \vartheta]$. Next we consider the case $u \in [\vartheta, 1]$. By the definition of the covariance function of the Brownian Motion we have that $\gamma_u(v) = \min(u, v) = \min(\vartheta, v) = \gamma_\vartheta(v)$ if $u \in [\vartheta, 1]$ and $v \in [0, \vartheta]$. But this implies that $\mathcal{S}_u(X_i^O) = \langle \gamma_u, X_i^O \rangle_H = \langle \gamma_\vartheta, X_i^O \rangle_H = X_i^O(\vartheta) = \tilde{X}_i^{opt}(u)$ for all $u \in [\vartheta, 1]$, since $\langle \gamma_\vartheta, X_i^O \rangle_H$ is again an evaluation functional.

Result (c) of Theorem 2.3 can be used to identify cases that allow for a perfect prediction. By Eq. (14) there is no prediction error, i.e., $\mathbb{V}(Z_i(u)) = 0$ for $u \in M_i$ if the event $X_i^O = X_j^O$ implies that also $X_i(u) = X_j(u)$ for all $u \in M_i$. Only the unobserved $u \in M_i$ are considered here as, by construction, $\mathbb{V}(Z_i(v)) = 0$ for all observed $v \in O$.

To give an example, assume that for some integer $K < \infty$ we have $X_i(u) = \sum_{k=1}^{K} \theta_{ik} f_k(u)$ for all $u \in [a, b]$, where $f_1, \ldots, f_K$ are continuous orthonormal functions on $[a, b]$, and where $\theta_{ik} \in \mathbb{R}$ are Gaussian random variables with mean zero and (co-)variances $\mathbb{E}(\theta_{ik}\, \theta_{il}) = v_k < \infty$ for all $k = l$ and zero else. Now consider the observed interval $O \subseteq [a, b]$ and suppose that also the $K$ censored functions $f_1^O, \ldots, f_K^O$, with $f_k^O(u) := f_k(u)$ for all $u \in O$ and all $k \in \{1, \ldots, K\}$, are linearly independent. This is equivalent to require that

$$\int_O \left( \sum_{k=1}^{K} (\theta_{ik} - \alpha_k) f_k(u) \right)^2 du = 0, \text{ if and only if, } \theta_{i1} = \alpha_1, \ldots, \theta_{iK} = \alpha_K.$$

In other words, in this situation the knowledge of the structure of $X_i^O$ is sufficient to determine the corresponding coefficients $\theta_{i1}, \ldots, \theta_{iK}$ which uniquely determine the total function $X_i$ on $[a, b]$. By Eq. (14) we can then conclude that there is no prediction error, i.e., $\mathbb{V}(Z_i(u)) = 0$ for all $u \in [a, b]$. This can be made more precisely using change of basis arguments. If the orthonormal functions $f_1, \ldots, f_K$ are linear independent on $[a, b]$, we have that $\sum_{k=1}^{K} \theta_{ik} f_k(u) = \sum_{k=1}^{K} \xi_{ik}^O \phi_k^O(u)$ for all $u \in O$, where $\lambda_K^O > 0$ and $\lambda_{K+1}^O = 0$. Furthermore, there exist unique coefficients $\beta_1, \ldots, \beta_K$ such that $\phi_k^O(u) = \sum_{l=1}^{K} \beta_l f_l(u)$ for all $u \in O$. This then implies that $\tilde{\phi}_k(u) = \sum_{l=1}^{K} \beta_l f_l(u)$ as well as $X_i(u) = \sum_{k=1}^{K} \theta_{ik} f_k(u) = \sum_{k=1}^{K} \xi_{ik}^O \tilde{\phi}_k(u) = \tilde{X}_i(u)$ for all $u \in [a, b]$, i.e., $\mathbb{V}(Z_i(u)) = 0$ for all $u \in [a, b]$.

**3. Estimation.** In practice we typically do not observe the functional trajectories directly, but only the discretized functions – with or without measurement errors. For instance, Figure 1 shows the *pre-smoothed* functions, however, the actual raw data is shown in Figure 2. For this case, let $\mathbb{X}_i^O := ((Y_{i1}, U_{i1}), \ldots, (Y_{im}, U_{im}))$ denote the observable data pairs of a function $X_i^O$, where

$$Y_{ij} = X_i^O(U_{ij}) + \varepsilon_{ij}, \quad i \in \{1, \ldots, n\}, \quad j \in \{1, \ldots, m\}. \tag{16}$$

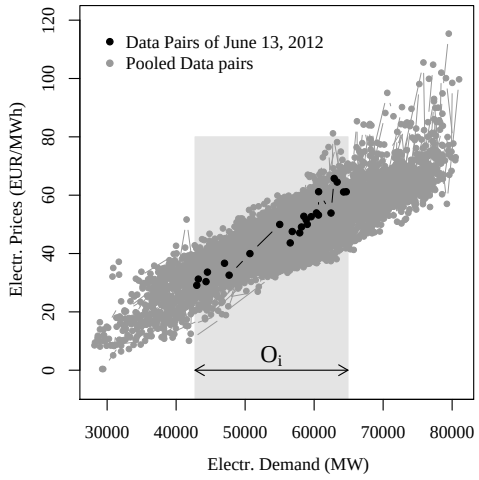


FIG 2. *Scatter plot of the observed data pairs* $(Y_{ij}, U_{ij})$.

The error term $\varepsilon_{ij}$ is a real iid random variable with mean zero and finite variance $\mathbb{V}(\varepsilon_{ij}) = \sigma_\varepsilon^2$, $0 \leq \varepsilon_{ij} < \infty$. The conditional discretization points $U_{ij}|O_i$ are iid random variables with pdf $f_{U|O_i}$ having $\text{supp}(f_{U|O_i}) = O_i \subseteq [a,b]$. The marginal pdf $f_U$ of $U_{ij}$ is assumed to have $\text{supp}(f_U) = [a,b]$. The error term $\varepsilon_{ij}$ is assumed to be independent from all other random variables. The random variables $U_{ij}$ and $O_i$ are assumed to be independent from $(X_i)_{i=1,\ldots,n}$, which leads again to the "missing completely at random" assumption already stated above in Section 1. We consider the case where $O_i = [A_i, B_i]$ with $A_i = \min_j(U_{ij})$, $B_i = \max_j(U_{ij})$, and $B_i - A_i > c_{AB}$ for some constant $0 < c_{AB} < 1$.

We propose to estimate $\mathcal{L}_u(X_i^O)$ through estimating the difference functional $\mathcal{D}_u(X_i^O)$, where $\mathcal{D}_u(X_i^O)$ is defined in Eq. (11). This allows us to derive sensitive rates of convergence depending on whether $\mathcal{D}_u(X_i^O)$ is a regression or prediction functional. Given an estimate of $\mathcal{D}_u(X_i^O)$, one directly has an estimate for the prediction functional by $\mathcal{L}_u(X_i^O) = X_i^O(\vartheta_u) + \mathcal{D}_u(X_i^O)$.

Our estimator of $\mathcal{D}_u(X_i^O)$ is simply the empirical and truncated version of $\mathcal{D}_u(X_i^O)$, i.e., for the case of non-centered data we define

$$\hat{\mathcal{D}}_{u,K}(\mathbb{X}_i^O) := \hat{\mu}(u; h_\mu) - \hat{\mu}(\vartheta_u; h_\mu) + \\ + \sum_{k=1}^{K} \frac{\hat{\xi}_{i,k}^O}{\hat{\lambda}_k^O} \int_O \hat{\phi}_k^O(v)\big(\hat{\gamma}(u, v; h_\gamma) - \hat{\gamma}(\vartheta_u, v; h_\gamma)\big)dv, \quad u \in M. \tag{17}$$

In practice, the $K$ can be chosen, e.g., by the Fraction of Variance Explained

(FVE) criterion or the AIC as suggested in Yao, Müller and Wang (2005a). In our theoretical analysis we consider $K \equiv K_{nm} \to \infty$ as the sample size $nm \to \infty$.

Then for any $u \in M$ the estimator of $\mathcal{L}_u(X_i^O)$ is

$$\hat{\mathcal{L}}_{u,K}(\mathbb{X}_i^O) := \begin{cases} X_i^O(\vartheta_u) & + \hat{\mathcal{D}}_{u,K}(\mathbb{X}_i^O), \quad \text{if } \mathbb{V}(\varepsilon_{ij}) = 0, \\ \hat{X}_i^O(\vartheta_u; h_X) & + \hat{\mathcal{D}}_{u,K}(\mathbb{X}_i^O), \quad \text{if } \mathbb{V}(\varepsilon_{ij}) > 0, \end{cases} \tag{18}$$

with $\vartheta_u = A_i$ if $|A_i - u| < |B_i - u|$ and $\vartheta_u = B_i$ else. In the case where $\mathbb{V}(\varepsilon_{ij}) > 0$ we do not directly observe $X_i^O(\vartheta_u)$, but have to estimate this value using, e.g., the LLK estimator $\hat{X}_i^O(\vartheta_u; h_X) = \hat{\beta}_0$, where

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0,\beta_1}{\arg\min} \sum_{j=1}^{m} [Y_{ij} - \beta_0 - \beta_1(U_{ij} - u)]^2 K_{h_X}(U_{ij} - u) \tag{19}$$

for $K_h(.) = \kappa(./h)/h$. The kernel function $\kappa$ is assumed to be a univariate symmetric pdf with compact support $\text{supp}(\kappa) = [-1, 1]$ such as, e.g., the Epanechnikov kernel (see Assumption A5). The usual kernel constants are given by $\nu_2(\kappa) := \int v^2 \kappa(v) dv$, and $R(\kappa) := \int \kappa(v)^2 dv$.

The empirical pc-score $\hat{\xi}_{ik}^O$ in Eq. (17) is defined by the following integral approximation of $\xi_{ik}^O$:

$$\hat{\xi}_{ik}^O = \sum_{j=2}^{m} \hat{\phi}_k^O(U_{ij})(Y_{ij} - \hat{\mu}(U_{ij}; h_\mu))(U_{ij} - U_{i,j-1}). \tag{20}$$

The mean estimator is $\hat{\mu}(u; h_\mu) = \hat{\beta}_0$, where

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0,\beta_1}{\arg\min} \sum_{i=1}^{n} \sum_{j=1}^{m} [Y_{ij} - \beta_0 - \beta_1(U_{ij} - u)]^2 K_{h_\mu}(U_{ij} - u). \tag{21}$$

As the optimality result in part (b) of Theorem 2.3 does not make any specific assumptions on the process $X_i$, we do not want to unnecessarily restrict this generality when it comes to estimation of the covariance function. For instance, when estimating the covariance function nonparametrically it is common to assume that all second order derivatives of $\gamma(u, v)$ are continuous. However, the covariance functions of many important stochastic process, such as the Brownian Motion, are not differentiable at the diagonal. Therefore, our local linear estimator of $\gamma$ slightly deviates from that proposed in Yao, Müller and Wang (2005a) and Yao, Müller and Wang (2005b) as we do not assume that $\gamma(u, v)$ is differentiable at the diagonal (see Assumption A3 below).

By the symmetry of the covariance function, it suffices to estimate either the upper diagonal part or the lower diagonal part of the covariance function. Therefore, we suggest to estimate only the upper diagonal points of the covariance function $\gamma(u,v)$ for $u \leq v$, and thus to avoid smoothing over the diagonal. To do so, we define upper-diagonal raw covariance points

$$\hat{C}_{ijl} := (Y_{ij} - \hat{\mu}(U_{ij}))(Y_{il} - \hat{\mu}(U_{il})), \quad \text{for} \quad U_{ij} < U_{il} \tag{22}$$

for $i \in \{1,\dots,n\}$ and $j,l \in \{1,\dots,m\}$. The covariance estimator is As Yao, Müller and Wang (2005a), we do not include the raw covariances $\hat{C}_{ijj}$ on the diagonal for which $U_{ij} = U_{ij}$, as these would introduce an estimation bias through taking squares of the error term $\varepsilon_{ij}$ contained in $Y_{ij}$. This leads to $nM$ many upper-diagonal raw covariance points where $M = (m^2 - m)/2$. For given data $(\hat{C}_{ijl}, U_{ij}, U_{il})$ with $U_{ij} < U_{il}$, we estimate the covariance function $\gamma(u,v)$ by $\hat{\gamma}(u,v;h_\gamma)$ if $u \leq v$ and by $\hat{\gamma}(v,u;h_\gamma)$ if $u > v$ with $\hat{\gamma}(u,v;h_\gamma) = \hat{\beta}_0$, where

$$\begin{aligned}(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2) =& \underset{\beta_0,\beta_1,\beta_2}{\arg\min} \sum_{i=1}^{n} \sum_{\substack{1\leq j,l\leq m \\ s.t.\ U_{ij}<U_{jl}}} [C_{ijl} - \beta_0 - \beta_1(U_{ij}-u) - \beta_2(U_{il}-v)]^2 \\ & \times K_{h_\gamma}(U_{ij}-u) K_{h_\gamma}(U_{il}-v).\end{aligned} \tag{23}$$

*Remark.* Using only the upper-diagonal raw covariance points means not that we are using less information. The upper- and lower-diagonal raw covariance points are anyways only the mirrored points to each other. Yao, Müller and Wang (2005a) and Yao, Müller and Wang (2005b) assume a more restrictive situation where all second order derivatives at the diagonal $\gamma(u,u)$ are continuous. Our approach is consistent there too – without any loss in terms of convergence rates.

Estimates of the eigenvalues $\lambda_k^O$ and the eigenfunctions $\phi_k^O$ are defined by the corresponding solutions of the empirical eigen-equations:

$$\int_O \hat{\gamma}(u,v;h_\gamma)\hat{\phi}_k^O(v)\,dv = \hat{\lambda}_k^O\,\hat{\phi}_k^O(u), \quad u \in O. \tag{24}$$

**4. Asymptotic results.** Our asymptotic results are developed under the following assumptions which are generally close to those in Yao, Müller and Wang (2005b) and Hall, Müller and Wang (2006). Though, we additionally allow for weakly dependent time series of random function $(X_i)_i$ (Assumption A1), since our real data application consists of a time series of

random functions. Furthermore, we consider more general asymptotic setup (Assumption A2).

**A1** (Stochastic assumptions) Our stochastic assumptions on $Y_{ij}, U_{ij}, \varepsilon_{ij}, X_i$ and $O_i$ are as defined below of Eq. (16). Furthermore, in order to assure that the mean function $\mu$ and the covariance function $\gamma$ are estimable, we assume that the random set $O_i \times O_i$ covers $[a,b]^2$, with a strictly positive probability. Additionally, we assume that $(X_i)_{i=1,\ldots,n}$ is a strictly stationary ergodic functional time series with finite fourth moments, i.e., $\mathbb{E}(||X_i||_2^4) < \infty$ and autocovariance functions with geometric decay. That is, there exist constants $c_1, c_2, r_1, r_2$ with $0 < c_1, c_2 < \infty$ and $0 < \iota_1, \iota_2 < 1$, such that $\sup_{(u,v)\in[a,b]^2} |\gamma_h(u,v)| \leq c_1 \iota_1^h$ and $\sup_{(u_1,v_1,u_2,v_2)\in[a,b]^4} |\dot{\gamma}_h((u_1,v_1),(u_2,v_2))| \leq c_2 \iota_2^h$, where $\gamma_h(u,v) := \mathrm{Cov}(X_{i+h}(u), X_i(v))$ and $\dot{\gamma}_h((u_1,v_1),(u_2,v_2)) := \mathrm{Cov}(X_{i+h}(u_1)X_{i+h}(v_1), X_i(u_2)X_i(v_2))$ for all $h \geq 0$.

**A2** (Asymptotic scenario) $nm \to \infty$, where $n \to \infty$ and $m = m_n \sim n^\theta$ with $0 \leq \theta < \infty$. Here, $a_n \sim b_n$ is used to denote that $\lim_{n\to\infty}(a_n/b_n) = c$ for some constant $0 < c < \infty$. It is assumed that $m_n \geq 2$ for all $n$.

**A3** (Smoothness) For $\hat{\mu}$: All second order derivatives of $\mu(u)$ on $[a,b]$, $f_U(u)$ on $[a,b]$, $\gamma(u,v)$ on $[a,b]^2$, and of $f_{YU}(y,u)$ on $\mathbb{R} \times [a,b]$ are uniformly continuous and bounded, where $f_{YU}$ is the joint pdf of $(Y_{ij}, U_{ij})$, and where $\mathrm{supp}(f_U) = [a,b]$. For boundary points the appropriate one sided continuity applies. For $\hat{\gamma}$: All second order derivatives of $\gamma(u,v)$ on $\{(u,v) \in [a,b]^2 : u \neq v\}$, $f_{UU}(u,v)$ on $[a,b]^2$, $\dot{\gamma}((u_1,v_1),(u_2,v_2))$ on $[a,b]^4$, and of $f_{CUU}(c,u,v)$ on $\mathbb{R} \times [a,b]^2$ are uniformly continuous and bounded, where $f_{CUU}$ is the joint pdf of $(C_{ijl}, U_{ij}, U_{il})$, and where $\mathrm{supp}(f_{UU}) = [a,b]^2$. For boundary points the appropriate one sided continuity applies.

**A4** (Bandwidths) For estimating $X_i$: $h_X \equiv h_{m,X} \to 0$ and $(m\,h_X)^{-1/2} \to \infty$ as $m \to \infty$. For estimating $\mu$: $h_\mu \equiv h_{nm,\mu} \to 0$ and $(nm\,h_\mu)^{-1/2} \to \infty$ as $nm \to \infty$. For estimating $\gamma$: $h_\gamma \equiv h_{nM,\gamma} \to 0$ and $(nM\,h_\gamma)^{-1/2} \to \infty$ as $nM \to \infty$, where $M = (m^2 - m)/2$.

**A5** (Kernel function) The kernel function $\kappa$ is a univariate, symmetric, pdf with compact support $\mathrm{supp}(\kappa) = [-1,1]$.

**A6** (Eigenvalues) For all $k \geq 1$ and all $O$ we assume that $\lambda_k^O > 0$ with multiplicity one, furthermore that $\lambda_k^O = \mathcal{O}(k^{-a(O)})$ with $a(O) > 1$, and that $\lambda_k^O - \lambda_{k+1}^O \geq c(O)k^{-a(O)-1}$ with $0 < c(O) < \infty$.

**A7** (Additional regularity assumptions on $X_i$)

$\sup_{u\in[a,b]}\{\mathbb{E}(|X_i(u)|^C)\} < \infty \quad \text{for all} \quad C > 0,$
$\sup_{u,v\in[a,b]}(\mathbb{E}[\{|u-v|^{-\epsilon}|X_i(u)-X_i(v)|\}^C]) < \infty \ \text{ for some } \ \epsilon > 0 \ \text{ and}$
$\sup_{k\geq 1}\lambda_k^{-r}\{\mathbb{E}[(\int_a(X_i(u)-\mathbb{E}(X_i(u)))\phi_k(u)\,du)^{2r}]\} < \infty \ \forall r = 1, 2, \ldots$

*Remarks on the assumptions.* The double asymptotic in Assumption A2 is inspired by the work of Zhang and Wang (2016), who investigate all data scenarios from sparse to dense functional data. For the estimation of the mean and covariance functions, the eigenvalues, and eigenfunctions we allow for the case of sparsely sampled prediction points $U_{it}$, i.e., with $\theta = 0$ and $m_n = \text{const}$ for all $n \to \infty$. However, for estimating the pc- scores, we only focus on the case of $\theta > 0$ which implies that $m_n \to \infty$ as $n \to \infty$; otherwise the pc-scores cannot be estimated consistently. Under truly sparse sampling schemes one should use the PACE procedure of Yao, Müller and Wang (2005a) and Yao, Müller and Wang (2005b) to get approximations for the pc-scores. We follow Hall, Müller and Wang (2006) and consider $m$ as deterministic, though, for a random $m$ our results are conditional on $m$. Assumptions A3-A5 contain typical smoothness, bandwidth and kernel assumptions from nonparametric statistics. In order to make use of the eigenvalue and eigenfunction expansions in Hall and Hosseini-Nasab (2006), we impose the additional regularity Assumptions A6 and A7.

All of the theoretical results in the preceding Section 2 are valid for the case of weakly dependent stationary functional data (Assumption A1), except for result (c) in Theorem 2.3 which has to be adjusted as following:

THEOREM 4.1 (Variance of the Prediction Error - Time Series Version). *Let $X_i$ and $X_j$, with $i \neq j$, be identically distributed autocorrelated Gaussian processes with autocovariance function as in Assumption A1. The variance of the prediction error can then be written as*

$$\mathbb{V}(Z_i(u)) = \tfrac{1}{2}\,\mathbb{E}\left(\mathbb{E}\left(\left(X_i(u)-X_j(u)\right)^2 \middle| X_i^O = X_j^O\right)\right) + \mathcal{O}(\iota_1^{|i-j|})$$

*for all $|i-j| > 1$.*
*The proof can be found in Appendix A of the supplemental paper.*

*Remark.* Theorem 4.1 can be interpreted in the same fashion as part (c) of Theorem 2.3, but with the additional requirement that the distance between the time points $i$ and $j$ needs to be large enough such that the $\mathcal{O}(\iota_1^{|i-j|})$ term is negligible. This additional requirement assures that the implication "$X_i^O = X_j^O \Rightarrow X_i = X_j$", is not just a mere consequence of a positive autocorrelation.

The following theorem contains our preliminary consistency results which lead to our main consistency result in Theorem 4.3.

THEOREM 4.2 (Preliminary consistency results (discrete data)).
*Under Assumptions A1-A5 we have that:*

*(a)* $\sup_{u\in[a,b]} |\hat{\mu}(u;h_\mu) - \mu(u)| = \mathcal{O}_p\left(r_\mu\right)$
*(ã)* *Conditional on* $X_i^O$*:* $\sup_{u\in O} |\hat{X}_i^O(u;h_\mu,h_X) - X_i^O(u)| = \mathcal{O}_p\left(r_X\right)$
*(b)* $\sup_{(u,v)\in[a,b]^2} |\hat{\gamma}(u,v;h_\gamma) - \gamma(u,v)| = \mathcal{O}_p\left(r_\mu + r_\gamma\right)$, *where*

$r_\mu \equiv r_\mu(h_\mu,n,m) := h_\mu^2 + 1/\sqrt{nm\,h_\mu} + 1/\sqrt{n}$
$r_X \equiv r_X(h_X,m) := h_X^2 + 1/\sqrt{m\,h_X}$
$r_\gamma \equiv r_\gamma(h_\gamma,n,M) := h_\gamma^2 + 1/\sqrt{nM\,h_\gamma^2} + 1/\sqrt{n}.$

*If additionally Assumptions A6 and A7 hold, we have that:*

*(c)* $\sup_{k\geq 1} |\hat{\lambda}_k^O - \lambda_k^O| = \mathcal{O}_p\left(r_\mu + r_\gamma\right)$ *for all* $k \geq 1$
*(d)* $\sup_{u\in O} |\hat{c}_k\hat{\phi}_k^O(u) - \phi_k^O(u)| = \mathcal{O}_p\left((\delta_k^O)^{-1}\left(r_\mu + r_\gamma\right)\right)$,
*for all* $1 \leq k \leq \bar{K}_{nM}^O - 1$, *where:*
$\hat{c}_k := \text{sign}(\langle\hat{\phi}_k^O, \phi_k^O\rangle_2)$, $\delta_k^O := \min_{1\leq i\leq k}\{\lambda_i^O - \lambda_{i+1}^O\}$,
$\bar{K}_{nM}^O := \inf\{k \geq 1 : \lambda_k^O - \lambda_{k+1}^O \leq 2\hat{\Delta}_{nM}^O\}$, *and*
$\hat{\Delta}_{nM}^O := (\int_{(u,v)\in O^2}(\hat{\gamma}^O(u,v) - \gamma(u,v))^2 d(u,v))^{1/2}$, *i.e.,*
$\hat{\Delta}_{nM}^O = \mathcal{O}_p(r_\mu + r_\gamma)$ *by result (b) above.*

*The proofs can be found in Appendix A of the supplemental paper.*

*Remark.* Similar results can be found in the Theorems 3.1, 3.3, and 3.6 of Li and Hsing (2010), who consider uniform almost sure convergence. Results (a) and (b) have also counterparts in Theorems 5.1 and 5.2 of Zhang and Wang (2016), who also consider uniform almost sure convergence. The uniform almost sure convergence rates in the latter references are slower than our weak uniform convergence rates which attain the well-known optimal rates of convergence of LLK estimators.

THEOREM 4.3 (Main consistency results).
*Under Assumptions A1-A7, the following results hold for* $1 \leq K \leq \bar{K}_{TN} - 1$, *and for optimal bandwidth* $h_\mu \sim (nm)^{-1/5}$ *and* $h_\gamma \sim (nM)^{-1/6}$*:*

*(a) For* $n \to \infty$, $m \to \infty$ *with* $m \sim n^\theta$ *and* $0 < \theta < 1/4$*:*

$$\sup_{u\in M}\left|\mathcal{D}_u(X_i^O) - \hat{\mathcal{D}}_{u,K}(\mathbb{X}_i^O)\right| = \mathcal{O}_p\left(\frac{K}{m^{1/2}} + \frac{\sum_{k=1}^K k^{a+1}}{(nM)^{1/3}}\right) + o_p(R_K)$$

*(b) For* $n \to \infty$, $m \to \infty$ *with* $m \sim n^\theta$ *and* $1/4 \leq \theta < 1$*:*

$$\sup_{u\in M}\left|\mathcal{D}_u(X_i^O) - \hat{\mathcal{D}}_{u,K}(\mathbb{X}_i^O)\right| \mathcal{O}_p\left(\frac{K}{m^{1/2}} + \frac{\sum_{k=1}^K k^{a+1}}{n^{1/2}}\right) + o_p(R_K)$$

*(c) For $n \to \infty$, $m \to \infty$ with $m \sim n^\theta$ and $1 \leq \theta < \infty$:*

$$\sup_{u \in M} \left| \mathcal{D}_u(X_i^O) - \hat{\mathcal{D}}_{u,K}(\mathbb{X}_i^O) \right| = \mathcal{O}_p\left( \frac{\sum_{k=1}^K k^{a+1}}{n^{1/2}} \right) + o_p(R_K),$$

*where the $o_p(R_K)$ terms quantify the cut-off regularization error with rates*

$$R_K = \begin{cases} 1, & \text{if } \mathcal{D}_u(X_i^O) \text{ is a prediction functional} \\ \sqrt{\sum_{k=K+1}^\infty k^{-a}}, & \text{if } \mathcal{D}_u(X_i^O) \text{ is a regression functional.} \end{cases}$$

*The proofs can be found in Appendix A of the supplemental paper.*

**5. Simulation study.** For our simulation study we generate $n$ many iid normal and and $n$ many exponential random functions $X_i(u) = \mu(u) + \xi_{i1}\phi_1(u) + \xi_{i2}\phi_2(u)$, where $\mu(u) = u + \sin(u)$, $\phi_1(u) = -\cos((\pi u)/(b-a))/\sqrt{5}$, $\phi_2(u) = -\cos(2(\pi u)/(b-a))/\sqrt{5}$, $a = 1$, and $b = 10$. For the normal case we let $\xi_{ik} \sim N(0, \lambda_k)$ with $\lambda_1 = 4$ and $\lambda_2 = 3$. For the exponential case we let $\xi_{ik}$ be a centered exponential random variable with rate $\lambda_k$ and centered by $\lambda_k^{-1}$. Furthermore, $U_{i1}, \ldots U_{im} \overset{\text{iid}}{\sim} \text{Unif}[A_i, B_i]$, where $A_i \overset{\text{iid}}{\sim} \text{Unif}[a, a + (b-a) \cdot 0.25]$ and $B_i \overset{\text{iid}}{\sim} \text{Unif}[b - (b-a) \cdot 0.25, b]$. Finally, the observations $Y_{ij}$ are generated according to $Y_{ij} = X_t(U_{ij}) + \varepsilon_{ij}$ with $\varepsilon_{ij} \sim N(0, 0.2)$. The following data dimensions are investigated: $n \in \{100, 200\}$ and $m \in \{15, 25, 50, 75\}$. The parameter $K$ is selected by the Fraction of Variance Explained (FVE) criterion with FVE $= 0.95$, i.e., we choose the highest $K$ for which $(\sum_{k=1}^K \hat{\lambda}_k)/(\sum_{l \geq 1} \hat{\lambda}_l) \leq \text{FVE}$. In the $j$th ($j = 1, \ldots, 100$) simulation run we compute the Mean Absolute Prediction Error (MAPE) as

$$\text{MAPE}_j = T^{-1} \sum_{t=1}^T \max_u |\hat{\hat{X}}_{i,K}^{(j)}(u) - X_i(u)|, \quad j = \{1, \ldots, 100\}.$$

The bandwidths $h_X$, $h_\mu$ and $h_\gamma$ are determined using generalized cross validation. The implementation is done in R and the codes are available from the authors.

Figure 3 shows the averages $\text{MAPE} = B^{-1} \sum_{j=1} \text{MAPE}_j$ for the two different data generating processes and the different data dimensions. The simulation results demonstrate that the improvements in the MAPE are larger when increasing $m$ from 15 to 25 than when increasing $m$ from 50 to 75. This is in line with our asymptotic results: If $m$ is relatively small, then increasing $m$ will reduce the estimation error (Theorem 4.3 (a)). However, a relatively large $m$ is only first order relevant for estimating $\hat{X}_i(\vartheta_u; h_X)$, as mirrored by the $m^{2/5}$ rate, but not anymore for estimating $\hat{\hat{X}}_{i,K}$ (Theorem 4.3 (c)).

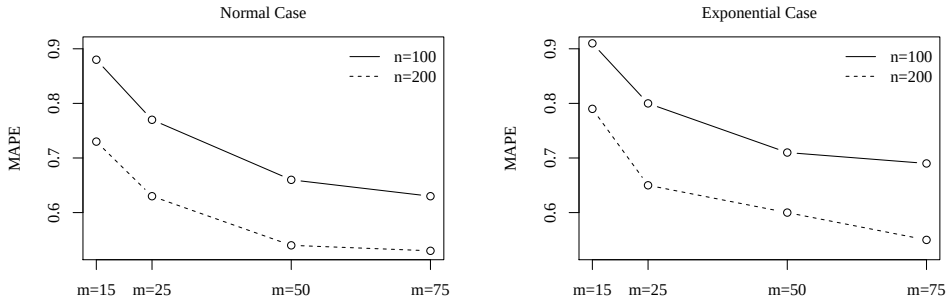


FIG 3. *Mean absolute prediction errors for the normal and exponential data generating processes.*

**6. Prediction algorithm.** In practice it can happen that the partially observed functions $X_i^O$ contain sufficient information to estimate the mean function, but not for estimating the covariance function. For instance, if the pooled sample $X_1^O, \ldots, X_n^O$ covers the total domain $[a, b]$, we can estimate the mean function, but if none of the single functions $X_i^O$ covers the total domain $[a, b]$, we cannot estimate the covariance function $\gamma$ on its total support $[a, b]^2$. There are then, e.g., no data pairs $(X_i^O(a), X_i^O(b))$ which can be used to estimate $\gamma(a, b) = \mathrm{Cov}(X_i(a), X_i(b))$. This situation is shown in Figure 7 where the mean function $\mu$ can be estimated, but not the outer off-diagonal parts of the covariance function $\gamma$.

Our prediction procedure can be used directly under such an unpleasant situation. The estimate $\hat{\hat{X}}_i$, however, will then generally not cover the total domain $[a, b]$, but a subinterval, say, $I_i \subseteq [a, b]$, where $O_i \subseteq I_i$. It can be seen from the definition of $\hat{\hat{X}}_i$, in Eq. (**??**), that the subinterval $I_i$ is determined by the interval $O_i$ and the extend to which $\gamma$ can be estimated. The determination of the subinterval $I_i$ is shown in Figure 4.

This first run of our prediction procedure may not allow to recover all functions over the total domain $[a, b]$. For instance, the predicted price function in the left panel of Figure 4 still lacks the upper fragment for values of electricity demand $u \in [77362\ (\mathrm{MW}), 82282\ (\mathrm{MW})]$. In order to predict also the further missing fragments, we suggest to repeat the prediction procedure. In the following we describe the $\ell$th, $\ell \geq 2$, run of the prediction algorithm:

**$\ell$th Run:** Compute a new prediction $\hat{\hat{X}}_{i,K}^{(\ell)}$ based on the artificial predictor $X_i^{O^*}(u) := \hat{\hat{X}}_{i,K}^{(\ell-1)}(u)$, $u \in O_i^*$, where $O_i^*$ is appropriately chosen such that a still missing fragment of $X_i$ can be predicted.

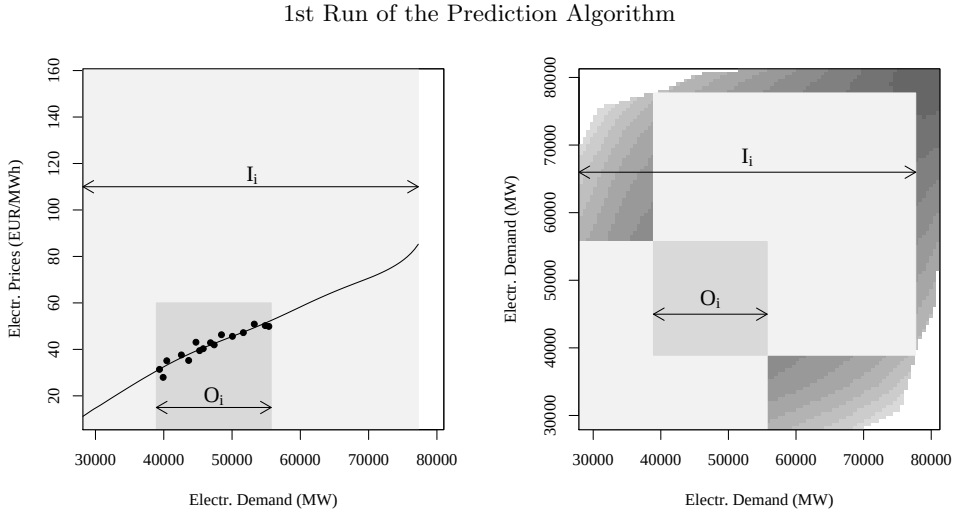


FIG 4. *Explanatory plots for the first run of the prediction algorithm.*

An exemplary second ($\ell = 2$) run is visualized in Figure 5. There the new interval $O_i^* \subseteq I_i^{(1)}$ is chosen such that the missing upper fragment can be predicted: the new large interval $I_i^{(2)}$ contains the missing upper interval $[77362\ (\text{MW}), 82282\ (\text{MW})]$. One can use this new fragment in order to extend the prediction from the first run.

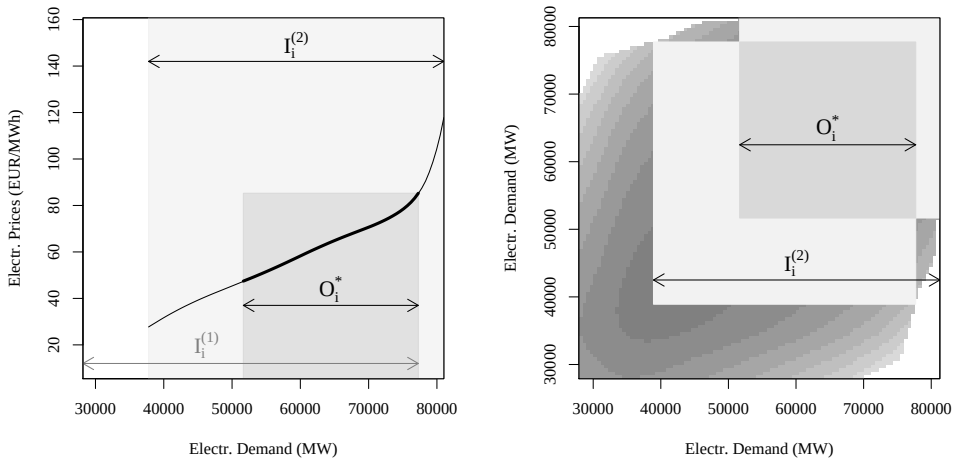


FIG 5. *Explanatory plots for the second run of the prediction algorithm.*

*Practical remark:.* In the $\ell$th, $\ell \geq 2$, run of the prediction algorithm, we can estimate the pc-scores by the classical integral method, i.e., by

$$\hat{\xi}_{ik}^{O^*} = \int_{O_i^*} (X_i^{O^*}(u) - \hat{\mu}(u))\hat{\phi}_k^{O^*}(u)du.$$

## 7. Application.

*Data sources and preprocessing.* The data for our analysis come from three different sources. Hourly spot prices of the German electricity market are provided by the European Energy Power Exchange (EPEX) (www.epexspot.com), hourly values of Germany's gross electricity demand and electricity exchanges with other countries are provided by the European Network of Transmission System Operators for Electricity (www.entsoe.eu), and German wind and solar power infeed data are provided by the transparency platform of the European energy exchange (www.eex-transparency.com). For academic, i.e., non-commercial purposes the data can be accessed free of charges from these sources. The data dimensions are given by $m = 24$ hours and $n = 241$ working days between March 15, 2012 and March 14, 2013. Very few (0.4%) of the data pairs $(Y_{ij}, U_{ij})$ with prices $Y_{ij} > 120$ EUR/MWh and $U_{ij} > 82000$ MW are considered as outliers and reset to $Y_{ij} = 120$.

The German electricity market, like many other electricity markets, provides purchase guarantees for Renewable Energy Sources (RES). Therefore, the relevant variable for pricing at the energy exchange is electricity demand minus electricity infeeds from RES (Nicolosi, 2010). Correspondingly, the hourly values of electricity demand $U_{ij}$ actually denote *residual* electricity demand, i.e., electricity demand minus infeeds from RES. That is, $U_{ij} := \texttt{Elect.Demand}_{ij} - \texttt{RES}_{ij}$, where $\texttt{RES}_{ij} = \texttt{Wind.Infeed}_{ij} + \texttt{Solar.Infeed}_{ij}$. The effect of further RES such as biomass is still negligible for the German electricity market.

### 7.1. *Checking for perfect predictability.*

According to result (c) of Theorem 2.3, the prediction error can be zero if the structure of the random function $X_i$ is simple enough. In the following, we check empirically whether the functions are perfectly predictable. The idea is to use the observed functions $X_i^O$ and to partition them artificially into an observed part and a missing part. This way we can check whether the predicted missing part equals to the artificially missing part, i.e., whether the functions are perfectly predictable. In practice, however, we have to take into account the estimation errors, which imply that the empirical prediction error $\hat{Z}_i(u) \neq 0$ even though the theoretical prediction error $Z_i(u) = 0$.

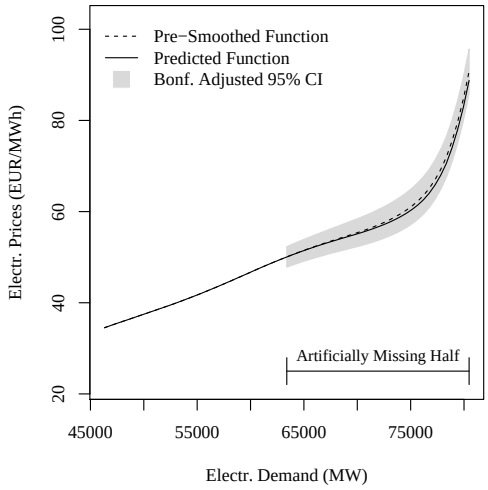


FIG 6. *Checking whether the predicted function is within the Bonferroni adjusted 95% confidence interval of the pre-smoothed function.*

We take the lower halves of the pre-smoothed functions $\hat{X}_1^O, \ldots, \hat{X}_n^O$ as the observed parts and use them to predict their artificially missing upper halves and vise-versa which leads to in total $2n$ checks. To account for the pre-smoothing estimation error we use bootstrap $(1 - \alpha_{\text{Bonf}})$ confidence intervals with Bonferroni adjusted $\alpha_{\text{Bonf}} = 0.05/(2n)$ in order to account for the $2n = 482$ checks. Only if the prediction result $\hat{\hat{X}}_{i,K}$ is within this confidence interval, the prediction is considered as perfect. The function-wise pre-smoothing is done using the local linear estimator. To archive a stable bandwidth selection, we use a common bandwidth for all pre-smoothing problems which is computed as the mean of the function-wise cross-validation bandwidths. Figure 6 demonstrates this approach for the case where we get a perfect prediction of the artificially missing upper half given the information from the observed lower half of the price function. Using this approach, 96% of the $2n$ predictions are considered as perfect which is in line with the chosen significance level of $\alpha = 5\%$ where we expect only 5% of the predictions to be falsely classified as imperfect. This supports our hypothesis of perfect predictability. Though, we emphasize that the predictions are optimal among all linear predictions – no matter whether the prefect predictability condition holds or not.

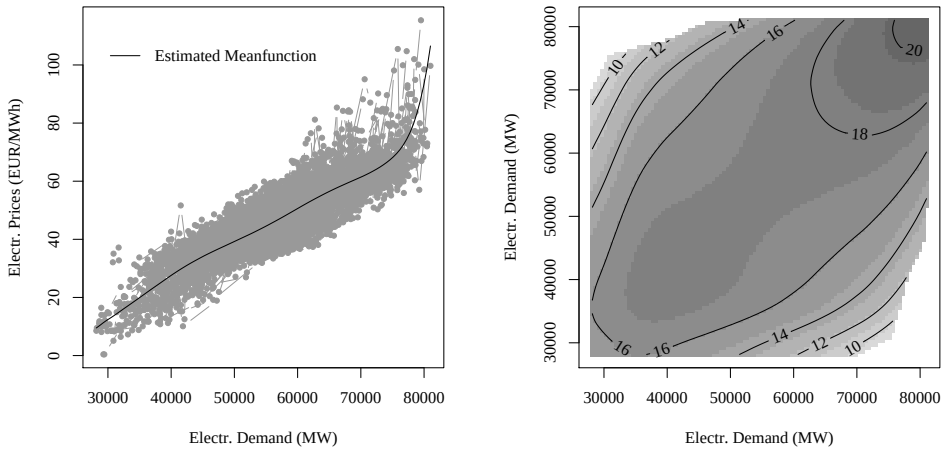


FIG 7. LEFT PANEL: *Estimated mean function plus scatter plot of the data pairs* $(Y_{ij}, U_{ij})$. RIGHT PANEL: *Contour plot of the estimated covariance function. The white regions reflect the outer off-diagonal parts which are infeasible to estimate.*

7.2. *Empirical results.* The estimated mean function and the estimated covariance function are shown in Figure 7. The outer off-diagonal parts of the covariance function $\gamma$ cannot be estimated, since these parts of the domain are not covered by data pairs $(U_{ij}, U_{il})$, $j \neq l$. For predicting the missing parts of each function $X_i^O$ we use the repetitive prediction algorithm as described in Section 6. We use at most three repetitions for each partially observed price function. In the first run, we use the information with respect to the original observations on $X_i^O$ in order to predict the missing parts as far as possible – as visualized in Figure 4. In the second run, we choose an artificial predictor $X_i^{O^*}$ where $O_i^* \subseteq I^{(1)}$ is appropriately chosen such that a possibly missing upper fragment of $X_i$ can be predicted – as visualized in Figure 5. In the third run, we choose the artificial predictor $X_i^{O^*}$ such that a possibly missing lower fragment of $X_i$ can be predicted. This approach allows us to recover 91% of the price functions over the total support. The predicted functions are shown in Figure 8.

Our preceding analysis on checking for perfect predictability, suggests that the recovered functions are reliable estimates of the true price functions. Note that the prediction of negative prices is well justified as the EPEX allows for negative prices since the price reform in 2008 (Nicolosi, 2010). Electricity producers are willing to sell electricity at negative prices (i.e., to pay for selling) if shutting-off and re-starting their power plants is more expansive than selling their electricity at negative prices.

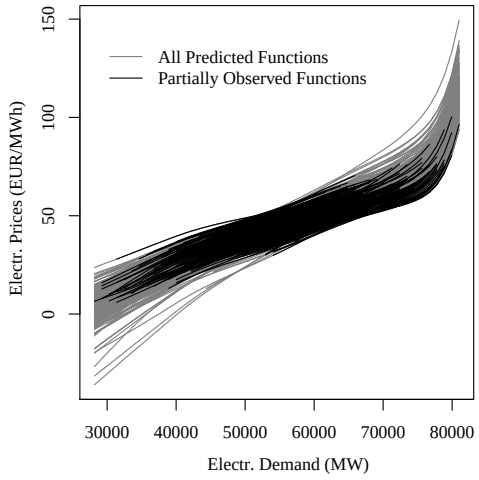


FIG 8. *All predicted/recovered functions (gray) plus a visualization of the original partially observed functions* $X_i^O$ *(black).*

**8. Conclusion.** We propose a new prediction functional that allows to recover partially observed functional data in an optimal manner. By contrast to the literature, we build upon the very particular nature of this prediction problem: a response function and a predictor function that are both coming from the same function. This allows us to derive very strong optimality results, which go beyond those found in the literature. In our estimation theory, we allow for autocorrelated functional data and focus on the relevant practical situation in which we do not directly observe the functional data. We derive the uniform rates of consistency for our nonparametric estimation procedures. The finite sample properties are investigated through simulations and a real data application.

## References.

CAI, T. T. and HALL, P. (2006). Prediction in functional linear regression. *The Annals of Statistics* **34** 2159–2179.

DELAIGLE, A. and HALL, P. (2013). Classification using censored functional data. *Journal of the American Statistical Association* **108** 1269–1283.

DELAIGLE, A. and HALL, P. (2016). Approximating fragmented functional data by segments of Markov chains. *Biometrika* **00** 1–21.

GOLDBERG, Y., RITOV, Y. and MANDELBAUM, A. (2014). Predicting the continuation of a function with applications to call center data. *Journal of Statistical Planning and Inference* **147** 53–65.

HALL, P. and HOROWITZ, J. L. (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics* **35** 70–91.

HALL, P. and HOSSEINI-NASAB, M. (2006). On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68** 109–126.

HALL, P., MÜLLER, H. G. and WANG, J. L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *The Annals of Statistics* **34** 1493–1517.

HIRTH, L. (2013). The market value of variable renewables: The effect of solar wind power variability on their relative price. *Energy economics* **38** 218–236.

KRAUS, D. (2015). Components and completion of partially observed functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77** 777–801.

LI, Y. and HSING, T. (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *The Annals of Statistics* **38** 3321–3351.

NICOLOSI, M. (2010). Wind power integration and power system flexibility – An empirical analysis of extreme events in Germany under the new negative price regime. *Energy Policy* **38** 7257–7268.

RUPPERT, D. and WAND, M. P. (1994). Multivariate locally weighted least squares regression. *The Annals of statistics* 1346–1370.

TSYBAKOV, A. (2008). *Introduction to Nonparametric Estimation*, 1. ed. *Springer Series in Statistics*. Springer.

WEIGT, H. (2009). Germany's wind energy: The potential for fossil capacity replacement and cost saving. *Applied Energy* **86** 1857–1863.

YAO, F., MÜLLER, H. G. and WANG, J. L. (2005a). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **100** 577–590.

YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005b). Functional linear regression analysis for longitudinal data. *The Annals of Statistics* **33** 2873–2903.

ZHANG, X. and WANG, J.-L. (2016). From sparse to dense functional data and beyond. *The Annals of Statistics* **44** 2281–2321.

DOMINIK LIEBL AND ALOIS KNEIP

STATISTISCHE ABTEILUNG

UNIVERSITY OF BONN

ADENAUERALLEE 24-26

53113 BONN, GERMANY

E-MAIL: dliebl@uni-bonn.de

# Supplemental paper for:

# On the Optimal Reconstruction of Partially Observed Functional Data

by Dominik Liebl and Alois Kneip

## CONTENT

In the following we give the the proofs of Theorems 2.2, 2.3, 4.2, and 4.3. The main steps in our proofs of the Theorems 4.2 and 4.3 are as in Yao, Müller and Wang (2005a). Though, by contrast to Yao, Müller and Wang (2005a), we allow for a time series context (see Assumption A1) and consider a different asymptotic setup (see Assumption A2).

## APPENDIX A: PROOFS

**Proof of Theorem 2.1:** For every linear functional $L_u : \mathbb{L}^2(O_i) \to \mathbb{R}$ that is a prediction functional with respect to $X_i^O$, according to Def. 2.1, we have that

$$\mathbb{V}(L_u(X_i^O)) = \sum_{k=1}^{\infty} \lambda_k^O \left(L_u(\phi_k^O)\right)^2 \quad \text{for every} \quad u \in [a,b]. \tag{25}$$

Existence: Writing $L_u(X_i^O)$ as $L_u(X_i^O) = \langle \alpha_u, X_i^O \rangle_H$ for some $\alpha_u \in H$ and computing the variance of $L_u(X_i^O)$ yields that

$$\mathbb{V}(L_u(X_i^O)) = \sum_{k=1}^{\infty} \lambda_k^O \left( \frac{\langle \alpha_u, \phi_k^O \rangle_2^2}{\lambda_k^O} \right)^2 . \tag{26}$$

Since (25) and (26) must be equal, we have that $L_u(\phi_k^O) = \langle \alpha_u, \phi_k^O \rangle_2^2 / \lambda_k^O$ for all $k \geq 1$, which establishes that there exits a $\alpha_u \in H$ for every prediction functional $L_u$.
Uniqueness: Assume that there is an alternative $\tilde{\alpha}_u \in H$ such that $L_u(\phi_k^O) = \langle \tilde{\alpha}_u, \phi_k^O \rangle_2^2 / \lambda_k^O$ for all $k \geq 1$. Then $\langle \alpha_u - \tilde{\alpha}_u, \phi_k^O \rangle_2^2 / \lambda_k^O = 0$ for all $k \geq 1$ or equivalently $\langle \alpha_u, \phi_k^O \rangle_2 - \langle \tilde{\alpha}_u, \phi_k^O \rangle_2 = 0$ for all $k \geq 1$ which shows that $\tilde{\alpha}_u - \alpha_u = 0$.

**Proof of Theorem 2.2, part (a):** First, note that continuity of $\gamma(u,v)$ implies continuity of $\mathbb{V}(\tilde{X}_i(u))$. Second, note that for any $K$ and every $u \in$

$[a,b]$, we have

$$0 \leq \mathbb{E}\left(\left(X_i(u) - \sum_{k=1}^{K} \xi_{ik}^O \tilde{\phi}_k(u)\right)^2\right) = \gamma(u,u) - \sum_{k=1}^{K} \lambda_k^O \tilde{\phi}_k(u)^2. \tag{27}$$

But this implies that $\mathbb{V}(\tilde{X}_i^K(u)) = \mathbb{V}(\sum_{k=1}^K \xi_{ik}^O \tilde{\phi}_k(u)) = \sum_{k=1}^K \lambda_k^O \tilde{\phi}_k(u)^2$ converges to a fixed limit $0 \leq \mathbb{V}(\tilde{X}_i(u)) < \infty$ as $K \to \infty$ for all $u \in [a,b]$.

**Part (b):** Follows directly from observing that $\mathbb{E}(\tilde{X}_i(u)) = 0$ for all $u \in [a,b]$. □

**Proof of Theorem 2.3, part (a):** For all $v \in O_i$ and $u \in M_i$ we have that

$$\begin{aligned}
&\mathbb{E}\left(X_i^O(v) Z_i(u)\right) = \mathbb{E}\left(X_i^O(v)\left(X_i(u) - \tilde{X}_i(u)\right)\right) = \\
&= \mathbb{E}\left(\sum_{k=1}^{\infty} \xi_{ik}^O \phi_k^O(v)\Big(X_i(u) - \sum_{k=1}^{\infty} \xi_{ik}^O \tilde{\phi}_k(u)\Big)\right) = \\
&= \sum_{k=1}^{\infty} \phi_k^O(v)\Big(\mathbb{E}(\xi_{ik}^O X_i(u)) - \lambda_k \tilde{\phi}_k(u)\Big).
\end{aligned}$$

From the definition $\tilde{\phi}_k(u)$ in Eq. (7) we get that $\mathbb{E}(\xi_{ik}^O X_i(u)) = \lambda_k^O \tilde{\phi}_k(u)$, which leads to $\mathbb{E}(X_i^O(v) Z_i(u)) = 0$ for all $u \in [a,b]$. This proves Eq. (12), while Eq. (13) directly follows from the definition of $Z_i(x)$.

**Part (b):** By Theorem 2.1 there exists a unique $b_u \in H$ such that

$$\ell_u(X_i^O) = \langle b, X_i^O \rangle_H.$$

By Eq. (12) and the orthogonality property of the least squares projection we thus obtain

$$\begin{aligned}
&\mathbb{E}\left(\left(X_i(u) - \ell_u(X_i^O)\right)^2\right) = \\
&= \mathbb{E}\left(\left(\tilde{X}_i(u) + Z_i(u) - \langle b, X_i^O \rangle_H\right)^2\right) = \\
&= \mathbb{E}\left(\left(\tilde{X}_i(u) - \langle b, X_i^O \rangle_H\right)^2\right) + \mathbb{E}(Z_i(u)^2) + \\
&+ 2\left(\mathbb{E}(\tilde{X}_i(u)\, Z_i(u)) - \mathbb{E}\left(\langle b, X_i^O \rangle_H Z_i(u)\right)\right) = \\
&= \mathbb{E}\left(\left(\tilde{X}_i(u) - \langle b, X_i^O \rangle_H\right)^2\right) + \mathbb{E}(Z_i^2(u)) \geq \mathbb{E}(Z_i^2(u)).
\end{aligned}$$

**Part (c):** Observe that $\mathbb{V}(Z_i(u) - Z_j(u)) = \mathbb{V}(Z_i(u)) + \mathbb{V}(Z_j(u)) - 2\,\mathrm{Cov}(Z_i(u), Z_j(u)) = 2\,\mathbb{V}(Z_i(u))$ for all $u \in [a,b]$ and $i \neq j$. Rearranging and using that $\mathbb{E}(Z_i(u)) = \mathbb{E}(Z_j(u)) = 0$ for all $u \in [a,b]$ and all $i,j \in \{1,\ldots,n\}$ yields $\mathbb{V}(Z_i(u)) = \frac{1}{2}\,\mathbb{E}((Z_i(u) - Z_j(u))^2)$. From result (a) we know that $Z_i(u)$ and $X_i^O(v)$ are orthogonal and therefore uncorrelated for all $u \in [a,b]$ and all $v \in O_i$, that is, $\mathbb{E}\left(X_i^O(v)Z_i(u)\right) = \mathrm{Cov}\left(X_i^O(v), Z_i(u)\right) = 0$. Under the assumption of an independent Gaussian process, we have then independence between $Z_i(u)$ and $X_i^O(v)$, such that

$$\mathbb{V}(Z_i(u)) = \frac{1}{2}\,\mathbb{E}\left(\mathbb{E}\left((Z_i(u) - Z_j(u))^2\right)|X_i^O = X_j^O\right),$$

where $X_i^O = X_j^O$ means that $X_i^O(u) = X_i^O(u)$ for all $u \in O_i$.

□

**Proof of Theorem 4.1**: The proof follows the same arguments as the proof of result (c) in Theorem 2.3, but with the adjustment that $\mathbb{V}(Z_i(u) - Z_j(u)) = \mathbb{V}(Z_i(u)) + \mathbb{V}(Z_j(u)) - 2\,\mathrm{Cov}(Z_i(u), Z_j(u)) = 2\,\mathbb{V}(Z_i(u)) + \mathcal{O}(\iota_1^{|i-j|})$.

□

**Proof of Theorem 4.2, part (a)**: Let us rewrite the estimator $\hat{\mu}$ using matrix notation as in Ruppert and Wand (1994), i.e.,

$$\hat{\mu}(u;h_\mu) = e_1^\top \left([\mathbf{1}, \mathbf{U}_u]^\top \mathbf{W}_{\mu,u}[\mathbf{1}, \mathbf{U}_u]\right)^{-1} [\mathbf{1}, \mathbf{U}_u]^\top \mathbf{W}_{\mu,u}\mathbf{Y}, \tag{28}$$

where $e_1 = (1,0)^\top$, $[\mathbf{1}, \mathbf{U}_u]$ is a $nm \times 2$ dimensional data matrix with typical rows $(1, U_{ij} - u)$, the $nm \times nm$ dimensional diagonal weighting matrix $\mathbf{W}_{\mu,u}$ holds the kernel weights $K_{\mu,h}(U_{ij} - u) = h_\mu^{-1}\,\kappa(h_\mu^{-1}(U_{ij} - u))$. The objects $\mathbf{U}_u$ and $\mathbf{W}_{\mu,u}$ are filled in correspondence with the $nm$ dimensional vector $\mathbf{Y} = (Y_{11}, Y_{12}, \ldots, Y_{n,m-1}, Y_{n,m})^\top$.

This way we can decompose the estimator $\hat{\mu}(u;h_\mu)$ as

$$\hat{\mu}(u;h_\mu) = e_1^\top L_{1,nm,u}^{-1} L_{2,nm,u}, \tag{29}$$

with $2 \times 2$ matrix

$$\begin{aligned} L_{1,nm,u} &= (nm)^{-1}[\mathbf{1}, \mathbf{U}_u]^\top \mathbf{W}_{\mu,u}[\mathbf{1}, \mathbf{U}_u] \\ &= \begin{pmatrix} \frac{1}{nmh_\mu}\sum_{ij}\kappa\left(\frac{U_{ij}-u}{h_\mu}\right) & \frac{1}{nmh_\mu}\sum_{ij}\kappa\left(\frac{U_{ij}-u}{h_\mu}\right)(U_{ij}-u) \\ \frac{1}{nmh_\mu}\sum_{ij}\kappa\left(\frac{U_{ij}-u}{h_\mu}\right)(U_{ij}-u) & \frac{1}{nmh_\mu}\sum_{ij}\kappa\left(\frac{U_{ij}-u}{h_\mu}\right)(U_{ij}-u)^2 \end{pmatrix}, \end{aligned}$$

and $2 \times 1$ vector

$$L_{2,nm,u} = (nm)^{-1}[\mathbf{1}, \mathbf{U}_u]^\top \mathbf{W}_{\mu,u}\mathbf{Y} = \begin{pmatrix} \frac{1}{nmh_\mu}\sum_{ij}\kappa\left(\frac{U_{ij}-u}{h_\mu}\right)Y_{ij} \\ \frac{1}{nmh_\mu}\sum_{ij}\kappa\left(\frac{U_{ij}-u}{h_\mu}\right)(U_{ij}-u)\,Y_{ij} \end{pmatrix}.$$

Using the notation and the results from Lemma A.1 we have that

$$L_{1,nm,u} = \begin{pmatrix} \Psi_{0,nm}(u;h_\mu) & \Psi_{1,nm}(u;h_\mu) \\ \Psi_{1,nm}(u;h_\mu) & \Psi_{2,nm}(u;h_\mu) \end{pmatrix}$$

(30)
$$= \begin{pmatrix} f_U(u) & 0 \\ 0 & f_U(u)\nu_2(\kappa) \end{pmatrix} + \mathcal{O}_p^{\text{Unif}}\left(h_\mu^2 + \frac{1}{\sqrt{nm\,h_\mu}} + \frac{1}{\sqrt{n}}\right) \quad \text{and}$$

(31)
$$L_{2,nm,u} = \begin{pmatrix} \Psi_{3,nm}(u;h_\mu) \\ \Psi_{4,nm}(u;h_\mu) \end{pmatrix} = \begin{pmatrix} \mu(u)f_U(u) \\ 0 \end{pmatrix} + \mathcal{O}_p^{\text{Unif}}\left(h_\mu^2 + \frac{1}{\sqrt{nm\,h_\mu}} + \frac{1}{\sqrt{n}}\right),$$

where we write $\Psi_{q,nm}(u;h_\mu) - m_q(u) = \mathcal{O}_p^{\text{Unif}}(\texttt{rate})$ in order to denote that $\sup_{u\in[a,b]} |\Psi_{q,nm}(u;h_\mu) - m_q(u)| = \mathcal{O}_p(\texttt{rate})$. Taking the inverse of (30) gives

(32)
$$L_{nm,u}^{-1} = \begin{pmatrix} 1/f_U(u) & 0 \\ 0 & 1/(f_U(u)\nu_2(\kappa)) \end{pmatrix} + \mathcal{O}_p^{\text{Unif}}\left(h_\mu^2 + \frac{1}{\sqrt{nm\,h_\mu}} + \frac{1}{\sqrt{n}}\right).$$

Plugging (32) and (31) into (29) leads to

$$\sup_{u\in[a,b]} |\hat{\mu}(u;h_\mu) - \mu(u)| = \mathcal{O}_p\left(h_\mu^2 + \frac{1}{\sqrt{nm\,h_\mu}} + \frac{1}{\sqrt{m}}\right).$$

□

**Proof of Theorem 4.2, part (ã):** Observe that

$$\sup_{u\in O} |\hat{X}_i^O(u;h_\mu,h_X) - X_i^O(u)| \leq$$
$$\sup_{u\in O} |\hat{X}_i^{c,O}(u;h_X) - (X_i^O(u) - \mu(u))| + \sup_{u\in O} |\hat{\mu}(u;h_\mu) - \mu(u)|.$$

From Theorem 4.2, part (a), we have that $\sup_{u\in O} |\hat{\mu}(u;h_\mu) - \mu(u)| = \mathcal{O}_p(r_\mu)$ with $r_\mu = h_\mu^2 + 1/\sqrt{nm\,h_\mu} + 1/\sqrt{m}$. From a simplified version of the proof of Theorem 4.2, part (a), with $n = 1$, it follows that

$$\sup_{u\in O} |\hat{X}_i^{c,O}(u;h_X) - (X_i^O(u) - \mu(u))| = \mathcal{O}_p\left(h_X^2 + \frac{1}{\sqrt{mh_X}}\right)$$

□

**Proof of Theorem 4.2, part (b):**

Let us rewrite the estimator $\hat{\gamma}$ using matrix notation as in Ruppert and Wand (1994), i.e., $\hat{\gamma}(u,v;h_\gamma) =$

$$= e_1^\top \left( [\mathbf{1}, \mathbf{U}_u, \mathbf{U}_v]^\top \mathbf{W}_{\gamma,u,v} [\mathbf{1}, \mathbf{U}_u, \mathbf{U}_v] \right)^{-1} [\mathbf{1}, \mathbf{U}_u, \mathbf{U}_v]^\top \mathbf{W}_{\gamma,u,v} \hat{\mathbf{C}}, \tag{33}$$

where $e_1 = (1,0,0)^\top$, $[\mathbf{1}, \mathbf{U}_u, \mathbf{U}_v]$ is a $nM \times 3$ dimensional data matrix with typical rows $(1, U_{ij}-u, U_{il}-v)$, the $nM \times nM$ dimensional diagonal weighting matrix $\mathbf{W}_{\gamma,u,v}$ holds the bivariate kernel weights $K_{\gamma,h}(U_{ij}-u, U_{il}-v)$. For the bivariate kernel weights $K_{\gamma,h}(z_1,z_2) = h_\gamma^{-2}\kappa_\gamma(z_1,z_2)$ we use a multiplicative kernel function $\kappa_\gamma(z_1,z_2) = \kappa(z_1)\kappa(z_2)$ with $\kappa$ as defined above. The usual kernel constants are then $\nu_2(\kappa_\gamma) := (\nu_2(\kappa))^2$ and $R(\kappa_\gamma) := R(\kappa)^2$. The rows of the matrices $[\mathbf{1}, \mathbf{U}_u, \mathbf{U}_v]$ and $\mathbf{W}_{\gamma,u,v}$ are filled in correspondence with the $nM$ elements of the vector of raw covariances $\hat{\mathbf{C}} = (\dots, \hat{C}_{ijl}, \dots)^\top$.

Let us initially consider the infeasible estimator $\hat{\gamma}_C$ that is based on the infeasible "clean" raw covariances $C_{ijl} = (Y_{ij} - \mu(U_{ij}))(Y_{il} - \mu(U_{il}))$ instead of the estimator $\hat{\gamma}$ in (33) that is based on the "dirty" raw covariances $\hat{C}_{ijl} = (Y_{ij} - \hat{\mu}(U_{ij}))(Y_{il} - \hat{\mu}(U_{il}))$. Equivalently to the estimator $\hat{\mu}$ above, we can write the estimator $\hat{\gamma}_C$ as

$$\hat{\gamma}_C(u,v;h_\gamma) = e_1^\top \tilde{S}^{-1}_{1,nM,(u,v)} \tilde{S}_{2,nM,(u,v)}, \tag{34}$$

with

$$\tilde{S}^{-1}_{1,nM,(u,v)} = \begin{pmatrix} \Theta_{0,nM}(u,v;h_\gamma) & \Theta_{1,nM}(u,v;h_\gamma) \\ \Theta_{1,nM}(u,v;h_\gamma) & \Theta_{2,nM}(u,v;h_\gamma) \end{pmatrix}^{-1}$$

$$= \begin{pmatrix} 1/f_{UU}(u,v) & 0 \\ 0 & 1/f_{UU}(u,v)(\nu_2(\kappa))^2 \end{pmatrix} + \mathcal{O}_p^{\text{Unif}} \left( h_\gamma^2 + \frac{1}{\sqrt{nM\, h_\gamma^2}} + \frac{1}{\sqrt{n}} \right) \tag{35}$$

and $\tilde{S}_{2,nM,(u,v)} =$

(36)

$$= \begin{pmatrix} \Theta_{3,nM}(u,v;h_\gamma) \\ \Theta_{4,nM}(u,v;h_\gamma) \end{pmatrix} = \begin{pmatrix} \gamma(u,v) f_{UU}(u,v) \\ 0 \end{pmatrix} + \mathcal{O}_p^{\text{Unif}} \left( h_\gamma^2 + \frac{1}{\sqrt{nM\, h_\gamma^2}} + \frac{1}{\sqrt{n}} \right),$$

where we use the notation and the results from Lemma A.2, and where we write $\Theta_{q,nM}(u,v;h_\gamma) - \eta_q(u,v) = \mathcal{O}_p^{\text{Unif}}(\texttt{rate})$ in order to denote that $\sup_{(u,v)\in[a,b]^2} |\Theta_{q,nM}(u,v;h_\gamma) - \eta_q(u,v)| = \mathcal{O}_p(\texttt{rate})$.

Plugging (35) and (36) into (34) leads to

$$\sup_{(u,v)\in[a,b]^2} |\hat{\gamma}_C(u,v;h_\mu) - \gamma(u,v)| = \mathcal{O}_p \left( h_\gamma^2 + \frac{1}{\sqrt{nM\, h_\gamma^2}} + \frac{1}{\sqrt{n}} \right). \tag{37}$$

It remains to consider the additional estimation error, which comes from using the "dirty" response variables $\hat{C}_{ijl}$ instead of "clean" dependent variables $C_{ijl}$. Observe that we can expand $\hat{C}_{ijl}$ as following:

$$\begin{aligned}\hat{C}_{ijl} = C_{ijl} &+ (Y_{ij} - \mu(U_{ij}))(\mu(U_{il}) - \hat{\mu}(U_{il})) \\ &+ (Y_{il} - \mu(U_{il}))(\mu(U_{ij}) - \hat{\mu}(U_{ij})) \\ &+ (\mu(U_{ij}) - \hat{\mu}(U_{ij}))(\mu(U_{il}) - \hat{\mu}(U_{il})).\end{aligned}$$

Using our finite moment assumptions on $Y_{ij}$ (Assumption A1) and our result in Theorem 4.2, part (a), we have that

$$\begin{aligned}\hat{C}_{ijl} &= C_{ijl} + \mathcal{O}_p(1)\mathcal{O}_p\left(h_\mu^2 + \frac{1}{\sqrt{nm\,h_\mu}} + \frac{1}{\sqrt{n}}\right) \\ &+ \mathcal{O}_p(1)\mathcal{O}_p\left(h_\mu^2 + \frac{1}{\sqrt{nm\,h_\mu}} + \frac{1}{\sqrt{n}}\right) \\ &+ \left(\mathcal{O}_p\left(h_\mu^2 + \frac{1}{\sqrt{nm\,h_\mu}} + \frac{1}{\sqrt{n}}\right)\right)^2 = C_{ijl} + \mathcal{O}_p\left(h_\mu^2 + \frac{1}{\sqrt{nm\,h_\mu}} + \frac{1}{\sqrt{n}}\right),\end{aligned}$$

uniformly for all $j \neq l \in \{1, \ldots, m\}$ and $i \in \{1, \ldots, n\}$. Therefore

$$\sup_{(u,v)\in[a,b]^2} |\hat{\gamma}(u, v; h_\mu) - \gamma(u, v)| = \mathcal{O}_p\left(h_\gamma^2 + h_\mu^2 + \frac{1}{\sqrt{nM\,h_\gamma^2}} + \frac{1}{\sqrt{nm\,h_\mu^2}} + \frac{1}{\sqrt{n}}\right).$$

□

LEMMA A.1. *Define*

$$\Psi_{q,nm}(u; h_\mu) = \frac{1}{nmh_\mu} \sum_{ij} \kappa\left(\frac{U_{ij} - u}{h_\mu}\right) \psi_q\left(U_{ij} - u, Y_{ij}\right), \tag{38}$$

*where*

$$\psi_q\left(U_{ij} - u, Y_{ij}\right) = \begin{cases} (U_{ij} - u)^q & \text{for } q \in \{0, 1, 2\} \\ Y_{ij} & \text{for } q = 3 \\ (U_{ij} - u)\, Y_{ij} & \text{for } q = 4. \end{cases}$$

*Then, under Assumptions A1-A5,*

$$\tau_{q,nm} = \sup_{u\in[a,b]} |\Psi_{q,nm}(u; h_\mu) - m_q(u)| = \mathcal{O}_p\left(h_\mu^2 + \frac{1}{\sqrt{nm\,h_\mu}} + \frac{1}{\sqrt{n}}\right),$$

*where* $m_0(u) = f_U(u)$, $m_1(u) = 0$, $m_2(u) = f_U(u)\nu_2(\kappa)$, $m_3(u) = \mu(u)f_U(u) = \mathbb{E}(Y_{ij}|U_{ij} = u)f_U(u)$, *and* $m_4(u) = 0$.

LEMMA A.2. *Define*

(39)
$$\Theta_{q,nM}(u,v;h_\gamma) = \frac{1}{nMh_\gamma}\sum_{i,j\neq l}\kappa\left(\frac{U_{ij}-u}{h_\gamma}\right)\kappa\left(\frac{U_{il}-v}{h_\gamma}\right)\vartheta_q\left(U_{ij}-u, U_{il}-u, C_{ijl}\right),$$

*where*

$$\vartheta_q\left(U_{ij}-u, U_{il}-v, C_{ijl}\right) = \begin{cases} (U_{ij}-u)^q\,(U_{il}-v)^q & \text{for } q\in\{0,1,2\} \\ C_{ijl} & \text{for } q=3 \\ (U_{ij}-u)\,(U_{il}-v)\,C_{ijl} & \text{for } q=4. \end{cases}$$

*Then, under Assumptions A1-A5,*

$$\varrho_{q,nM} = \sup_{(u,v)\in[a,b]^2}|\Theta_{q,nM}(u,v;h_\gamma)-\eta_q(u,v)| = \mathcal{O}_p\left(h_\gamma^2+\frac{1}{\sqrt{nM\,h_\gamma^2}}+\frac{1}{\sqrt{n}}\right),$$

*where* $\eta_0(u,v) = f_{UU}(u,v)$, $\eta_1(u,v) = 0$, $\eta_2(u,v) = f_{UU}(u,v)(\nu_2(\kappa))^2$, $\eta_3(u,v) = \gamma(u,v)f_{UU}(u,v) = \mathbb{E}(C_{ijl}|(U_{ij},U_{il}) = (u,v))f_{UU}(u,v)$, *and* $m_4(u,v) = 0$.

**Proof of Lemma A.1:** Remember that $\mathbb{E}(|\tau_{q,nm}|) = \mathcal{O}(\texttt{rate}_{nm})$ implies that $\tau_{q,nm} = \mathcal{O}_p(\texttt{rate}_{nm})$, therefore, we focus in the following on $\mathbb{E}(|\tau_{q,nm}|)$, where $\mathbb{E}(|\tau_{q,nm}|) = \mathbb{E}(\tau_{q,nm})$. Adding a zero and applying the triangle inequality yields that $\mathbb{E}(\tau_{q,nm}) =$

$$\begin{aligned}\mathbb{E}(\sup_{u\in[a,b]}|\Psi_{q,nm}(u;h_\mu)-m_q(u)|) \leq \sup_{u\in[a,b]}|\,\mathbb{E}(\Psi_{q,nm}(u;h_\mu))-m_p(u)|+ \\ +\,\mathbb{E}(\sup_{u\in[a,b]}|\Psi_{q,nm}(u;h_\mu)-\mathbb{E}(\Psi_{q,nm}(u;h_\mu))|). \qquad (40)\end{aligned}$$

Let us first focus on the second summand in (40). The next steps will make use of the Fourier transformation of the kernel function $\kappa$ (see, e.g., Tsybakov, 2008, Ch. 1.3):

$$\kappa^{\text{ft}}(x) := \mathcal{F}[\kappa](x) = \int_{\mathbb{R}}\kappa(z)\exp(-\mathrm{i}zx)dz = \int_{-1}^{1}\kappa(z)\exp(-\mathrm{i}zx)dz$$

with $\mathrm{i} = \sqrt{-1}$. By Assumption A5, $\kappa(.)$ has a compact support $[-1,1]$. The inverse transform gives then

$$\kappa\left(s\right) = \frac{1}{2\pi}\int_{\mathbb{R}}\kappa^{\text{ft}}(x)\exp\left(\mathrm{i}xs\right)dx = \frac{1}{2\pi}\int_{\mathbb{R}}\kappa^{\text{ft}}(x)\exp\left(\mathrm{i}xs\right)dx\;\mathbb{1}_{(|s|<1)}.$$

Furthermore, we can use that (see Tsybakov, 2008, Ch. 1.3, Eq. (1.34)) $\mathcal{F}[\kappa(./h_\mu)/h_\mu](x) = \mathcal{F}[\kappa](xh_\mu) = \kappa^{\text{ft}}(xh_\mu)$ which yields

$$\begin{aligned}\kappa\left(s/h_\mu\right)/h_\mu &= \frac{1}{2\pi}\int_{\mathbb{R}} \mathcal{F}[\kappa(./h_\mu)/h_\mu](x)\exp\left(\mathrm{i}xs\right)dx\ \mathbb{1}_{(|s|<h_\mu)} \\ &= \frac{1}{2\pi}\int_{\mathbb{R}} \kappa^{\text{ft}}(xh_\mu)\exp\left(\mathrm{i}xs\right)dx\ \mathbb{1}_{(|s|<h_\mu)}. \end{aligned} \tag{41}$$

Plugging (41) into (38) yields $\Psi_{q,nm}(u;h_\mu) =$

$$\begin{aligned}&= \frac{1}{nm}\sum_{ij}\kappa\left(\frac{U_{ij}-u}{h_\mu}\right)\frac{1}{h_\mu}\psi_q\left(U_{ij}-u, Y_{ij}\right) \\ &= \frac{1}{nm}\sum_{ij}\frac{1}{2\pi}\int_{\mathbb{R}}\kappa^{\text{ft}}(xh_\mu)\exp\big(\mathrm{i}x(U_{ij}-u)\big)dx\ \mathbb{1}_{(|U_{ij}-u|<h_\mu)}\ \psi_q\left(U_{ij}-u, Y_{ij}\right) \\ &= \frac{1}{2\pi}\int_{\mathbb{R}}\left[\frac{1}{nm}\sum_{ij}\exp\left(\mathrm{i}xU_{ij}\right)\psi_q\left(U_{ij}-u, Y_{ij}\right)\ \mathbb{1}_{(|U_{ij}-u|<h_\mu)}\right]\exp\big(\mathrm{i}xu\big)\kappa^{\text{ft}}(xh_\mu)dx.\end{aligned}$$

Using that $|\exp(ixu)| \leq 1$ leads to

$$\mathbb{E}(\sup_{u\in[a,b]}|\Psi_{q,nm}(u;h_\mu) - \mathbb{E}(\Psi_{q,nm}(u;h_\mu))|) \leq \frac{1}{2\pi}\,\mathbb{E}\left(\sup_{u\in[a,b]}\left|\int_{\mathbb{R}}\tilde{\omega}_{q,nm}(u,x)\cdot\kappa^{\text{ft}}(xh_\mu)dx\right|\right),$$

where

$$\begin{aligned}\tilde{\omega}_{q,nm}(u,x) =& \frac{1}{nm}\sum_{ij}\big[\exp\left(\mathrm{i}xU_{ij}\right)\psi_q\left(U_{ij}-u, Y_{ij}\right)\ \mathbb{1}_{(|U_{ij}-u|<h_\mu)} - \\ &\mathbb{E}\left(\exp\left(\mathrm{i}xU_{ij}\right)\psi_q\left(U_{ij}-u, Y_{ij}\right)\mathbb{1}_{(|U_{ij}-u|<h_\mu)}\right)\big].\end{aligned}$$

Using further that $\kappa^{\text{ft}}$ is symmetric, since $\kappa$ is symmetric by Assumption A5, and that $\exp\left(\mathrm{i}xU_{ij}\right) = \cos\left(xU_{ij}\right) + \mathrm{i}\sin\left(xU_{ij}\right)$ leads to

$$\frac{1}{2\pi}\,\mathbb{E}\left(\sup_{u\in[a,b]}\left|\int_{\mathbb{R}}\tilde{\omega}_{q,nm}(u,x)\cdot\kappa^{\text{ft}}(xh_\mu)dx\right|\right) = \frac{1}{2\pi}\,\mathbb{E}\left(\sup_{u\in[a,b]}\left|\int_{\mathbb{R}}\omega_{q,nm}(u,x)\cdot\kappa^{\text{ft}}(xh_\mu)dx\right|\right),$$

where

$$\begin{aligned}\omega_{q,nm}(u,x) =& \frac{1}{nm}\sum_{ij}\Big[\cos\left(xU_{ij}\right)\psi_q\left(U_{ij}-u, Y_{ij}\right)\ \mathbb{1}_{(|U_{ij}-u|<h_\mu)} - \\ &\mathbb{E}\left(\cos\left(xU_{ij}\right)\psi_q\left(U_{ij}-u, Y_{ij}\right)\mathbb{1}_{(|U_{ij}-u|<h_\mu)}\right)\Big],\end{aligned} \tag{42}$$

such that

$$\begin{aligned}
&\mathbb{E}(\sup_{u\in[a,b]}|\Psi_{q,nm}(u;h_\mu)-\mathbb{E}(\Psi_{q,nm}(u;h_\mu))|)\\
&\leq \frac{1}{2\pi}\int_{\mathbb{R}}\mathbb{E}\left(\sup_{u\in[a,b]}\left|\omega_{q,nm}(u,x)\right|\right)\cdot\left|\kappa^{\mathrm{ft}}(xh_\mu)\right|dx\\
&\leq \frac{1}{2\pi}\int_{\mathbb{R}}\sqrt{\mathbb{E}\left(\left(\sup_{u\in[a,b]}\left|\omega_{q,nm}(u,x)\right|\right)^2\right)}\cdot\left|\kappa^{\mathrm{ft}}(xh_\mu)\right|dx\\
&= \frac{1}{2\pi}\int_{\mathbb{R}}\sqrt{\mathbb{E}\left(\sup_{u\in[a,b]}\left(\omega_{q,nm}(u,x)\right)^2\right)}\cdot\left|\kappa^{\mathrm{ft}}(xh_\mu)\right|dx. \qquad (43)
\end{aligned}$$

In order to simplify the notation we will denote

$$W_{ij}^q(x,u)=\cos\left(xU_{ij}\right)\psi_q\left(U_{ij}-u,Y_{ij}\right),$$

such that $\mathbb{E}\left(\sup_{u\in[a,b]}\left(\omega_{q,nm}(u,x)\right)^2\right)=$

$$\begin{aligned}
\mathbb{E}\Big(\sup_{u\in[a,b]}\Big(&\frac{1}{(nm)^2}\sum_{ij}\left[W_{ij}^q(x,u)\mathbb{1}_{(|U_{ij}-u|<h_\mu)}-\mathbb{E}(W_{ij}^q(x,u)\mathbb{1}_{(|U_{ij}-u|<h_\mu)})\right]^2+\\
&\frac{1}{(nm)^2}\sum_{(i,j)\neq(r,l)}\left[(W_{ij}^q(x,u)\mathbb{1}_{(|U_{ij}-u|<h_\mu)}-\mathbb{E}(W_{ij}^q(x,u)\mathbb{1}_{(|U_{ij}-u|<h_\mu)}))\cdot\right.\\
&\left.\cdot(W_{rl}^q(x,u)\mathbb{1}_{(|U_{rl}-u|<h_\mu)}-\mathbb{E}(W_{rl}^q(x,u)\mathbb{1}_{(|U_{rl}-u|<h_\mu)}))\right]\Big)\Big).
\end{aligned}$$

As $u$ takes only values within the compact interval $[a,b]$, there exist constants $C_1$ and $C_2$ such that, uniformly for all $u\in[a,b]$, $\mathbb{P}(|U_{ij}-u|<h_\mu)\leq C_1h_\mu<\infty$, for all $i,j$, and $\mathbb{P}(|U_{ij}-u|<h_\mu \text{ AND } |U_{rl}-u|<h_\mu)\leq C_2h_\mu^2<\infty$, for all $(i,j)\neq(r,l)$. Together with the triangle inequality, this yields that $\mathbb{E}\left(\sup_{u\in[a,b]}\left(\omega_{q,nm}(u,x)\right)^2\right)\leq$

$$\begin{aligned}
&\frac{C_1h_\mu}{(nm)^2}\sum_{ij}\mathbb{E}\left(\sup_{u\in[a,b]}\left[W_{ij}^q(x,u)-\mathbb{E}(W_{ij}^q(x,u))\right]^2\right)+\\
&\frac{C_2h_\mu^2}{(nm)^2}\sum_{(i,j)\neq(r,l)}\mathbb{E}\left(\sup_{u\in[a,b]}\left[(W_{ij}^q(x,u)-\mathbb{E}(W_{ij}^q(x,u)))(W_{rl}^q(x,u)-\mathbb{E}(W_{rl}^q(x,u)))\right]\right)
\end{aligned}$$

From our moment assumptions (Assumption A1) and the fact that $[a, b]$ is compact, we can conclude that there must exist a constant $C_3$ such that, point-wise for every $x \in \mathbb{R}$,

$$\mathbb{E}\left(\left(\sup_{u\in[a,b]}\left|W_{ij}^q(x,u) - \mathbb{E}(W_{ij}^q(x,u))\right|\right)^2\right) \leq C_3 < \infty \tag{44}$$

for all $i$ and $j$.

"Within function" dependencies: By the same reasoning there must exist a constant $C_4$ such that, point-wise for every $x \in \mathbb{R}$,

$$\begin{aligned}&\mathbb{E}\left(\sup_{u\in[a,b]}\left|W_{ij}^q(x,u) - \mathbb{E}(W_{ij}^q(x,u))\right| \cdot \sup_{u\in[a,b]}\left|W_{il}^q(x,u) - \mathbb{E}(W_{il}^q(x,u))\right|\right)\\ &\leq C_4 < \infty\end{aligned} \tag{45}$$

for all $j \neq l$ and all $i$.

"Between function" dependencies: Our weak dependency assumption (Assumption A1) and the fact that $[a, b]$ is compact yields that point-wise for every $x \in \mathbb{R}$

$$\begin{aligned}&\mathbb{E}\left(\sup_{u\in[a,b]}\left|W_{ij}^q(x,u) - \mathbb{E}(W_{ij}^q(x,u))\right| \cdot \sup_{u\in[a,b]}\left|W_{rl}^q(x,u) - \mathbb{E}(W_{rl}^q(x,u))\right|\right)\\ &\leq c_1\iota_1^{|i-r|}\end{aligned} \tag{46}$$

for all $j, l$ and $|i - r| \geq 1$, where $0 < c_1 < \infty$ and $0 < \iota_1 < 1$.

Eq.s (44), (45), and (46) yield that $\mathbb{E}\left(\sup_{u\in[a,b]}\left(\omega_{q,nm}(u,x)\right)^2\right) \leq$

$$\begin{aligned}&\leq \frac{C_1 h_\mu}{(nm)^2}\sum_{ij} C_3 + \frac{C_2 h_\mu^2}{(nm)^2}\sum_{i,j\neq l} C_4 + \frac{C_2 h_\mu^2}{(nm)^2}\sum_{i\neq r,jl} c_1\iota_1^{|i-r|}\\ &= \mathcal{O}\left(\frac{h_\mu}{nm} + \frac{h_\mu^2(m-1)}{nm} + \frac{h_\mu^2}{n}\right) = \mathcal{O}\left(\frac{h_\mu}{nm} + \frac{h_\mu^2}{n}\right),\end{aligned}$$

such that

$$\sqrt{\mathbb{E}\left(\sup_{u\in[a,b]}\left(\omega_{q,nm}(u,x)\right)^2\right)} = \mathcal{O}\left(\sqrt{\frac{h_\mu}{nm}} + \frac{h_\mu}{\sqrt{n}}\right). \tag{47}$$

Plugging (47) into (43) and integration by substitution leads to

$$\mathbb{E}\Big(\sup_{u\in[a,b]}|\Psi_{q,nm}(u;h_\mu)-\mathbb{E}(\Psi_{q,nm}(u;h_\mu))|\Big)\leq$$

(48)

$$\frac{1}{2\pi}\int_{\mathbb{R}}\sqrt{\mathbb{E}\left(\sup_{u\in[a,b]}(\omega_{q,nm}(u,x))^2\right)}\cdot\left|\kappa^{\mathrm{ft}}(xh_\mu)\right|dx=\mathcal{O}\left(\frac{1}{\sqrt{nm\,h_\mu}}+\frac{1}{\sqrt{n}}\right).$$

Let us now focus on the first summand in (40). From standard arguments in nonparametric statistics (see, e.g., Ruppert and Wand (1994)) we know that

$$\mathbb{E}(\Psi_{q,nm}(u;h_\mu))-m_q(u)=\mathcal{O}(h_\mu^2)$$

for each $u\in[a,b]$ and for all $q\in\{0,\dots,4\}$. Under our smoothness Assumption A3, the "$\mathcal{O}(h_\mu^2)$" term becomes uniformly valid for all $u\in[a,b]$ and all $q\in\{0,1,2,4\}$, since all of the involved functions have uniformly bounded second order derivatives. We can conclude with respect to the first summand in (40) that

$$\sup_{u\in[a,b]}|\,\mathbb{E}(\Psi_{q,nm}(u;h_\mu))-m_p(u)|=\mathcal{O}(h_\mu^2)\quad\text{for all}\quad q\in\{0,\dots,4\}. \tag{49}$$

Finally, plugging our results (48) and (49) into (40) leads to

$$\tau_{q,nm}=\sup_{u\in[a,b]}|\Psi_{q,nm}(u;h_\mu)-m_q(u)|=\mathcal{O}_p\left(h_\mu^2+\frac{1}{\sqrt{nm\,h_\mu}}+\frac{1}{\sqrt{n}}\right) \tag{50}$$

for all $q\in\{0,\dots,4\}$. □

**The proof of Lemma A.2:** Analogously to that of Lemma A.1.

**The proof of Theorem 4.2, parts (c) and (d)** Parts (c) and (d) follow from the asymptotic expansions of eigenvalues and eigenfunctions of an estimated covariance function developed by Hall and Hosseini-Nasab (2006), and by applying the weak uniform convergence rate of part (b) in Theorem 4.2. The proofs follow exactly the same steps as the proofs of parts (a) and (c) in Theorem 3.6 of Li and Hsing (2010) and is omitted therefore.

**Proof of Theorem 4.3.**
In the following, we consider the asymptotic behavior of

$$\hat{\mathcal{D}}_{u,K}(\mathbb{X}_i^O)=\hat{\mu}(u;h_\mu)-\hat{\mu}(\vartheta_u;h_\mu)+\sum_{k=1}^{K}\hat{\xi}_{i,k}^O(\hat{\tilde{\phi}}_k(u)-\hat{\tilde{\phi}}_k(\vartheta_u)),\quad u\in M,$$

where $\hat{\tilde{\phi}}(u) = \frac{1}{\hat{\lambda}_k^O} \int_O \hat{\phi}_k^O(v)\hat{\gamma}(u,v;h_\gamma)dv$.

Using the results (b), (c), and (d) of Theorem 4.2 it follows directly that

$$\sup_{u\in M} |\hat{\tilde{\phi}}(u) - \tilde{\phi}(u)| = \mathcal{O}_p\left(\frac{r_\mu + r_\gamma}{\delta_k^O}\right). \tag{51}$$

Approximating the pc-scores:

$$\hat{\xi}_{ik}^O = \sum_{j=2}^{n} \hat{\phi}_k^O(U_{ij})(Y_{ij} - \hat{\mu}(U_{ij};h_\mu))(U_{ij} - U_{i,j-1}),$$

where $r_\mu = h_\mu^2 + 1/\sqrt{nm\,h_\mu} + 1/\sqrt{n}$ and $r_\gamma = h_\gamma^2 + 1/\sqrt{nM\,h_\gamma^2} + 1/\sqrt{n}$.

Using that $Y_{ij} = X_i^O(U_{ij}) + \varepsilon_{ij}$, a weak law of large numbers, and results (a) and (d) of Theorem 4.2 lead to

$$\begin{aligned}\hat{\xi}_{ik}^O &= \int_O \hat{\phi}_k^O(u)(X_i^O(u) - \hat{\mu}(u;h_\mu))\,du + \mathcal{O}_p\left(m^{-1/2}\right) \\ &= \xi_{kt}^O + \mathcal{O}_p\left(m^{-1/2} + (\delta_k^O)^{-1}(r_\mu + r_\gamma)\right).\end{aligned} \tag{52}$$

Next, let us denote

$$\mathcal{D}_{u,K}(X_i^O) = \mu(u) - \mu(\vartheta_u) + \sum_{k=1}^{K} \xi_{i,k}^O(\tilde{\phi}_k(u) - \tilde{\phi}_k(\vartheta_u)) \quad \text{and}$$

$$\mathcal{D}_{u,K^+}(X_i^O) = \sum_{k=K+1}^{\infty} \xi_{i,k}^O(\tilde{\phi}_k(u) - \tilde{\phi}_k(\vartheta_u))$$

such that $\mathcal{D}_u(X_i^O) = \mathcal{D}_{u,K}(X_i^O) + \mathcal{D}_{u,K^+}(X_i^O)$.

This allows us to decompose the prediction error into an estimation error part and an regularization error part:

$$\sup_{u\in M} \left|\mathcal{D}_u(X_i^O) - \hat{\mathcal{D}}_{u,K}(\mathbb{X}_i^O)\right| \le \sup_{u\in M} \left|\mathcal{D}_{u,K}(X_i^O) - \hat{\mathcal{D}}_{u,K}(\mathbb{X}_i^O)\right| + \sup_{u\in M} \left|\mathcal{D}_{u,K^+}(X_i^O)\right|. \tag{53}$$

Let us, first, focus on the estimation error, i.e., the first term on the right hand sides of (53):

$$\begin{aligned}&\sup_{u\in M} \left|\mathcal{D}_{u,K}(X_i^O) - \hat{\mathcal{D}}_{u,K}(\mathbb{X}_i^O)\right| \le \sup_{u\in M} |\mu(u) - \hat{\mu}(u)| + \sup_{u\in M} |\mu(\vartheta_u) - \hat{\mu}(\vartheta_u)| \\ &+ \sum_{k=1}^{K} \sup_{u\in M} |\xi_{ik}^O \tilde{\phi}_k(u) - \hat{\xi}_{ik}^O \hat{\tilde{\phi}}_k(u)| + \sum_{k=1}^{K} \sup_{u\in M} |\xi_{ik}^O \tilde{\phi}_k(\vartheta_u) - \hat{\xi}_{ik}^O \hat{\tilde{\phi}}_k(\vartheta_u)|.\end{aligned}$$

Using the approximations in (52) and (51) we get that:

$$\sum_{k=1}^{K} \sup_{u\in M} |\xi_{ik}^{O}\tilde{\phi}_k(u) - \hat{\xi}_{ik}^{O}\hat{\tilde{\phi}}_k(u)| =$$
$$= \sum_{k=1}^{K} \sup_{u\in M} |\xi_{ik}^{O}\tilde{\phi}_k(u) - (\xi_{ik}^{O} + \mathcal{O}_p(m^{-1/2} + \frac{r_\mu + r_\gamma}{\delta_k^O}))(\tilde{\phi}_k(u) + \mathcal{O}_p(\frac{r_\mu + r_\gamma}{\delta_k^O}))| =$$
$$= \mathcal{O}_p(Km^{-1/2}) + \sum_{k=1}^{K} \mathcal{O}_p(\frac{r_\mu + r_\gamma}{\delta_k^O})$$

The same result holds for

$$\sum_{k=1}^{K} \sup_{u\in M} |\xi_{ik}^{O}\tilde{\phi}_k(\vartheta_u) - \hat{\xi}_{ik}^{O}\hat{\tilde{\phi}}_k(\vartheta_u)| = \mathcal{O}_p(Km^{-1/2}) + \sum_{k=1}^{K} \mathcal{O}_p(\frac{r_\mu + r_\gamma}{\delta_k^O}).$$

Since $\sup_{u\in M} |\mu(u)-\hat{\mu}(u)|+\sup_{u\in M} |\mu(\vartheta_u)-\hat{\mu}(\vartheta_u)| = \mathcal{O}_p(r_\mu)$ is negligible in comparison to the other terms we get that

$$\sup_{u\in M} \left|\mathcal{D}_{u,K}(X_i^O) - \hat{\mathcal{D}}_{u,K}(\mathbb{X}_i^O)\right| = \mathcal{O}_p(Km^{-1/2}) + \sum_{k=1}^{K} \mathcal{O}_p(\frac{r_\mu + r_\gamma}{\delta_k^O}).$$

Using that Assumption (A6) implies that $(\delta_k^O)^{-1} = \mathcal{O}(k^{a+1})$ yields

(54)
$$\sup_{u\in M} \left|\mathcal{D}_{u,K}(X_i^O) - \hat{\mathcal{D}}_{u,K}(\mathbb{X}_i^O)\right| = \mathcal{O}_p\left(Km^{-1/2} + (r_\mu + r_\gamma)\sum_{k=1}^{K} k^{a+1}\right),$$

where we suppress the dependencies of the constants $0 < c = c(O) < \infty$ and $a = a(O) > 1$ on $O$.

Next we focus on the regularization error, i.e., the second term on the right hand side of (53).

The case of "prediction functionals": Observe that we can write

$$\mathcal{D}_{u,K^+}(X_i^O) = \sum_{k=K+1}^{\infty} \frac{\langle(\gamma_u - \gamma_{\vartheta_u}), \phi_k^O\rangle_2 \langle X_i^O, \phi_k^O\rangle_2}{\lambda_k^O},$$

where $\gamma_u(v) = \gamma(u, v)$ and $\gamma_{\vartheta_u}(v) = \gamma(\vartheta_u, v)$, with $v \in O$. Thich yields that

$$\mathbb{E}(|\mathcal{D}_{u,K^+}(X_i^O)|^2) = \sum_{k=K+1}^{\infty} \frac{\langle(\gamma_u - \gamma_{\vartheta_u}), \phi_k^O\rangle_2^2}{\lambda_k^O}.$$

Therefore

$$\sup_{u\in M} |\mathcal{D}_{u,K^+}(X_i^O)| = \mathcal{O}_p\left(r_K^P\right), \tag{55}$$

where $r_K^P = \sqrt{\sup_{u\in M}\sum_{k=K+1}^{\infty}\frac{\langle(\gamma_u-\gamma_{\vartheta_u}),\phi_k^O\rangle_2^2}{\lambda_k^O}}$

Since $\mathcal{D}_u(X_i^O)$ is a "prediction functional", we know that $\mathbb{V}(\mathcal{D}_u(X_i^O)) = ||(\gamma_u-\gamma_{\vartheta_u})||_H < \infty$ independently of the value of $u \in M$. This implies that $\mathbb{E}(|\mathcal{D}_{u,K^+}(X_i^O)|^2) = o(1)$, i.e., that $|\mathcal{D}_{u,K^+}(X_i^O)| = o_p(1)$ as $K = K_{nm} \to \infty$, independently of the value of $u \in M$. This shows that

$$r_K^P \to 0 \quad \text{as} \quad K = K_{nm} \to \infty.$$

The case of "regression functionals": Observe that

$$\begin{aligned}
\mathcal{D}_{u,K^+}(X_i^O) &= \sum_{k=K+1}^{\infty}\frac{\langle(\gamma_u-\gamma_{\vartheta_u}),\phi_k^O\rangle_2\langle X_i^O,\phi_k^O\rangle_2}{\lambda_k^O}\\
|\mathcal{D}_{u,K^+}(X_i^O)| &\le \sqrt{\sum_{k=K+1}^{\infty}\frac{\langle(\gamma_u-\gamma_{\vartheta_u}),\phi_k^O\rangle_2^2}{(\lambda_k^O)^2}}\sqrt{\sum_{k=K+1}^{\infty}\langle X_i^O,\phi_k^O\rangle_2^2}\\
\mathbb{E}(|\mathcal{D}_{u,K^+}(X_i^O)|^2) &\le \sum_{k=K+1}^{\infty}\frac{\langle(\gamma_u-\gamma_{\vartheta_u}),\phi_k^O\rangle_2^2}{(\lambda_k^O)^2}\sum_{k=K+1}^{\infty}\lambda_k^O
\end{aligned}$$

If $\mathcal{D}_u(X_i^O)$ is a "regression functional", we know (by requirement (9)) that $\sum_{k=1}^{\infty}\frac{\langle(\gamma_u-\gamma_{\vartheta_u}),\phi_k^O\rangle_2^2}{(\lambda_k^O)^2} < \infty$, which implies that

$$\sum_{k=K+1}^{\infty}\frac{\langle(\gamma_u-\gamma_{\vartheta_u}),\phi_k^O\rangle_2^2}{(\lambda_k^O)^2} = o(1) \quad \text{as} \quad K = K_{nm} \to \infty.$$

Therefore,

$$\mathbb{E}(|\mathcal{D}_{u,K^+}(X_i^O)|^2) = o\left(\sum_{k=K+1}^{\infty}\lambda_k^O\right).$$

By Assumption (A6) we have that $\lambda_k^O = \mathcal{O}(k^{-a})$, with $a > 1$, such that

$$\mathbb{E}(|\mathcal{D}_{u,K^+}(X_i^O)|^2) = o\left(\sum_{k=K+1}^{\infty}k^{-a}\right)$$

independently of $u \in M$. This leads to

$$\sup_{u\in M} |\mathcal{D}_{u,K^+}(X_i^O)| = o_p\left(\sqrt{\sum_{k=K+1}^{\infty} k^{-a}}\right) = o_p(1). \tag{56}$$

Summing up: From (54), (55), and (56) we have that

$$\sup_{u\in M}\left|\mathcal{D}_u(X_i^O) - \hat{\mathcal{D}}_{u,K}(\mathbb{X}_i^O)\right| =$$
$$= \begin{cases} \mathcal{O}_p\left(Km^{-1/2} + (r_\mu + r_\gamma)\sum_{k=1}^K k^{a+1}\right) + \mathcal{O}_p(r_K^P) & \text{for prediction functionals} \\ \mathcal{O}_p\left(Km^{-1/2} + (r_\mu + r_\gamma)\sum_{k=1}^K k^{a+1}\right) + o_p\left(\sqrt{\sum_{k=K+1}^{\infty} k^{-a}}\right) & \text{for regression functionals} \end{cases}$$

**Case-by-case considerations under optimal bandwidth choices:** Let us in the following focus on the case of optimal bandwidth choices, i.e., $h_\mu \sim (nm)^{-1/5}$ and $h_\gamma \sim (nM)^{-1/6}$. Then we have that for $0 < \theta < 1/4$: $r_\mu \sim (nm)^{-2/5}$ and $r_\gamma \sim (nM)^{-1/3}$, which implies that $r_\mu = o(r_\gamma)$. These considerations yield to

$$\sup_{u\in M}\left|\mathcal{D}_u(X_i^O) - \hat{\mathcal{D}}_{u,K}(\mathbb{X}_i^O)\right| =$$
$$= \begin{cases} \mathcal{O}_p\left(Km^{-1/2} + (nM)^{-1/3}\sum_{k=1}^K k^{a+1}\right) + \mathcal{O}_p(r_K^P) & \text{for prediction functionals} \\ \mathcal{O}_p\left(Km^{-1/2} + (nM)^{-1/3}\sum_{k=1}^K k^{a+1}\right) + o_p\left(\sqrt{\sum_{k=K+1}^{\infty} k^{-a}}\right) & \text{for regression functionals} \end{cases} \tag{57}$$

Under optimal bandwidth choices and for $1/4 \leq \theta < 1$ we have that $r_{nm}^\mu \sim r_{nM}^\gamma \sim n^{-1/2}$. These considerations yield to

$$\sup_{u\in M}\left|\mathcal{D}_u(X_i^O) - \hat{\mathcal{D}}_{u,K}(\mathbb{X}_i^O)\right| =$$
$$= \begin{cases} \mathcal{O}_p\left(Km^{-1/2} + n^{-1/2}\sum_{k=1}^K k^{a+1}\right) + \mathcal{O}_p(r_K^P) & \text{for prediction functionals} \\ \mathcal{O}_p\left(Km^{-1/2} + n^{-1/2}\sum_{k=1}^K k^{a+1}\right) + o_p\left(\sqrt{\sum_{k=K+1}^{\infty} k^{-a}}\right) & \text{for regression functionals} \end{cases} \tag{58}$$

Under optimal bandwidth choices and for $1 \leq \theta < \infty$ we have that $r_{nm}^\mu \sim r_{nM}^\gamma \sim n^{-1/2}$ and that $m^{-1/2} = \mathcal{O}\left(n^{-1/2}\right)$. But note that then $Km^{-1/2} =$

$\mathcal{O}(n^{-1/2}\sum_{k=1}^{K} k^{a+1})$. These considerations yield to

$$\sup_{u\in M}\left|\mathcal{D}_u(X_i^O) - \hat{\mathcal{D}}_{u,K}(\mathbb{X}_i^O)\right| =$$

(59)

$$= \begin{cases} \mathcal{O}_p\left(n^{-1/2}\sum_{k=1}^{K} k^{a+1}\right) + \mathcal{O}_p(r_K^P) & \text{for prediction functionals} \\ \mathcal{O}_p\left(n^{-1/2}\sum_{k=1}^{K} k^{a+1}\right) + o_p\left(\sqrt{\sum_{k=K+1}^{\infty} k^{-a}}\right) & \text{for regression functionals} \end{cases}$$

Dominik Liebl and Alois Kneip
Statistische Abteilung
University of Bonn
Adenauerallee 24-26
53113 Bonn, Germany
E-mail: dliebl@uni-bonn.de