A note on estimation in Hilbertian linear models *

Siegfried Hörmann‡, Łukasz Kidziński

Départment de Mathématique, Université Libre de Bruxelles, CP 215, Boulevard du Triomphe, B-1050 Bruxelles, Belgium

Abstract

We study estimation of the operator Ψ in the linear model $Y = \Psi(X) + \varepsilon$, when X and Y take values in Hilbert spaces H_1 and H_2, respectively. Our main objective is to obtain consistency without imposing some rather inconvenient technical assumptions that have been used in the literature. We develop our theory in a time dependent setup which comprises as important special case the autoregressive Hilbertian model.

Keywords: Adaptive estimation, consistency, functional regression, Hilbert spaces, infinite-dimensional data.
MSC2010: Primary 62G05, 62G20; Secundary 62J05, 62M10.

1 Introduction

In this paper we are concerned with a regression problem of the form

$$Y_k = \Psi(X_k) + \varepsilon_k, \quad k \geq 1, \tag{1.1}$$

where Ψ is a linear operator mapping from space H_1 to H_2. This model is fairly general and many special cases have been intensively studied in the literature. Our main objective is the study of this model when the regressor space H_1 is infinite dimensional. In the latter case, model (1.1) can be seen as a general formulation of a *functional linear model*, which is one of the most popular topics in functional data literature. Its various forms are introduced in Chapters 12–17 of Ramsay and Silverman [23]. To name a few recent references we mention Cuevas et al. [9], Malfait and Ramsay [21], Cardot et al. [5], Chiou et al. [6], Müller and Stadtmüller [22], Yao et al. [25], Cai and Hall [3], Li and Hsing [20], Reiss and Ogden [24], Febrero-Bande et al. [12], Crambes et al. [8], Ferraty et al. [13].

From an inferential point of view a natural problem is the estimation of the 'regression operator' Ψ. This topic has been discussed from several angles. For example, Cardot et al. [4] provide consistency results for the case of the 'functional regressor and scalar response model', while Cueavas et al. [9] consider a 'functional regressors and responses' setup assuming a non-random design. Yao et al. [25] also considered the 'functional regressors and responses model' but deal with the case where the observations are not fully observed but are obtained from sparce, irregular data measured with error.

*This research was supported by the Communauté française de Belgique—Actions de Recherche Concertées (2010–2015).

‡Corresponding author. Email: shormann@ulb.ac.be

The two main methods of estimation are based on principal component analysis (e.g. Bosq [1] and Cardot et al. [4]) or spline smoothing estimators (e.g. Hastie and Mallows [14], Marx and Eiler [11], Crambes et al. [8]).

In this paper we address the estimation problem for Ψ when the data are fully observed using the principal component approach. Let us explain what distinguishes our paper from previous work.

(i) A crucial difficulty is that we are working with an infinite dimensional operator Ψ, which needs to be approximated by a sample version $\hat{\Psi}_K$ of finite dimension K, say. A delicate issue is the choice of K. In existing papers determination of K requires very specific assumptions on the spectrum of the covariance operator of the regressor variables. In the next section we will explain why such assumptions pop up throughout the literature and show that consistency can be established without any assumption on the spectrum by proposing a *purely data-driven procedure* for the choice of K.

(ii) We allow the regressors X_k to be dependent. This is important for two reasons. First, many examples in FDA literature exhibit dependencies as the data stem from a continuous time process which is then segmented into a sequence of curves, e.g. by considering daily data. Examples of this kind include intraday patterns of pollution records, meteorogical data or financial transaction data or sequential fMRI recordings. See e.g. Horváth and Kokoszka [18].

Second, our framework detailed below, will include the important special case of a *functional autoregressive model*, which has been intensively investigated in the functional literature and is often used to model auto-regressive dynamics of a functional time series. This model is analyzed in detail in Bosq [2]. We can not only greatly simplify the assumptions needed for consistent estimation but also allow for a more general setup. E.g. in our Theorem 2.2 we show that it is not necessary to assume that Ψ is a Hilbert-Schmidt operator. This quite restrictive assumption is very often imposed though it e.g. even excludes the identity operator.

(iii) As we already mentioned before, the literature considers different forms of functional linear models. Arguably the most common are the *scalar response and functional regressor* and the *functional response and functional regressor case*. We will not distinguish between these cases, but work with a linear model between two general Hilbert spaces.

In the next section we will introduce notation, assumptions, the estimator and our main results. In Section 3 we compare the performance of the proposed estimators in a small simulation study and finally, in Section 4, we give the proofs.

2 Estimation of Ψ

2.1 Notation

Let H_1, H_2 be two (not necessarily distinct) separable Hilbert spaces. We denote by $\mathcal{L}(H_i, H_j)$, $(i, j \in \{1, 2\})$, the space of bounded linear operators from H_i to H_j. Further we write $\langle \cdot, \cdot \rangle_H$ for the inner product on Hilbert space H and $\|x\|_H = \sqrt{\langle x, x \rangle_H}$ for the corresponding norm. For $\Phi \in \mathcal{L}(H_i, H_j)$ we denote by $\|\Phi\|_{\mathcal{L}(H_i,H_j)} = \sup_{\|x\|_{H_i} \leq 1} \|\Phi(x)\|_{H_j}$ the operator norm and by $\|\Phi\|_{\mathcal{S}(H_i,H_j)} =$

$\left(\sum_{k=1}^{\infty} \|\Phi(e_k)\|_{H_j}^2\right)^{1/2}$, where $e_1, e_2, ... \in H_i$ is any orthonormal basis (ONB) of H_i, the Hilbert-Schmidt norm of Φ. It is well known that this norm is independent of the choice of the basis. Furthermore, with the inner product $\langle \Phi, \Theta \rangle_{\mathcal{S}(H_1,H_2)} = \sum_{k\geq 1} \langle \Phi(e_k), \Theta(e_k) \rangle_{H_2}$ the space $\mathcal{S}(H_1, H_2)$ is again a separable Hilbert space. For simplifying the notation we use $\mathcal{L}_{ij}$ instead of $\mathcal{L}(H_i, H_j)$ and in the same spirit $\mathcal{S}_{ij}$, $\|\cdot\|_{\mathcal{L}_{ij}}$, $\|\cdot\|_{\mathcal{S}_{ij}}$ and $\langle \cdot, \cdot \rangle_{\mathcal{S}_{ij}}$.

All random variables appearing in this paper will be assumed to defined on some common probability space $(\Omega, \mathcal{A}, P)$. A random element X with values in H is said to be in L_H^p if $\nu_{p,H}(X) = (E\|X\|_H^p)^{1/p} < \infty$. More conveniently we shall say that X has p moments. If X possesses a first moment, then X possesses a mean μ, determined as the unique element for which $E\langle X, x \rangle_H = \langle \mu, x \rangle_H$, $\forall x \in H$. For $X \in H_i$ and $Y \in H_j$ let $X \otimes Y : H_i \to H_j$ be an operator defined as $X \otimes Y(v) = \langle X, v \rangle Y$. If X and Y have 2 moments, then we say that X and Y are orthogonal ($X \perp Y$) if $EX \otimes Y = 0$. A sequence of orthogonal elements in H with a constant mean is called *H–white noise*. If $X \in L_H^2$ then it possesses a covariance operator C given by $C = E[(X - \mu) \otimes (X - \mu)]$. It can be easily seen that C is a Hilbert-Schmidt operator.

2.2 Setup

We consider the general regression problem (1.1) for *fully observed data*. Let us collect our main assumptions.

(A): *We have $\Psi \in \mathcal{L}_{12}$. Further $\{\varepsilon_k\}$ and $\{X_k\}$ are zero mean variables and are assumed to be* L^4–m–approximable *in the sense of Hörmann and Kokoszka [17] (see below). In addition $\{\varepsilon_k\}$ is H_2–white noise. For any $k \geq 1$ we have $X_k \perp \varepsilon_k$.*

Here is the weak dependence concept that we impose.

Definition 2.1 (Hörmann and Kokoszka [17])**.** *A random sequence $\{X_n\}_{n\geq 1}$ with values in H is called L^p–m–approximable if it can be represented as*

$$X_n = f(\delta_n, \delta_{n-1}, \delta_{n-2}, ...)$$

where the δ_i are iid elements taking values in a measurable space S and f is a measurable function $f : S^\infty \to H$. Moreover if δ_i' are independent copies of δ_i defined on the same probability space, then for

$$X_n^{(m)} = f(\delta_n, \delta_{n-1}, \delta_{n-2}, ..., \delta_{n-m+1}, \delta_{n-m}', \delta_{n-m-1}', ...)$$

we have

$$\sum_{m=1}^{\infty} \nu_{p,H}(X_m - X_m^{(m)}) < \infty.$$

The notion of L^p–m–approximability implicitely assumes that the process is stationary. Evidently an i.i.d. sequence with finite fourth moment is L^4–m–approximable. This leads to the classical functional linear model. However, our setup is much more general and allows e.g. to cover the

autoregressive Hilbertian model of order 1 (ARH(1)), given by the recursion $X_{k+1} = \Psi(X_k) + \varepsilon_{k+1}$. (See Section 2.4.) Other examples of functional time series covered in this framework can be found in [17]. This form of weak dependence also implies that a possible non-zero mean of X_k can be estimated consistently by the sample mean. Moreover we have (see [16])

$$\sqrt{n}\|\bar{X} - \mu\|_{H_1} = O_P(1).$$

We conclude that the mean can be accurately removed in a preprocessing step and that $EX_k = 0$ is not a stringent assumption. Since by Lemma 2.1 in [17] $\{Y_k\}$ will also be L^4–m–approximable, the same argument justifies that we study a linear model without intercept.

Our moment assumptions are mild. We are not aware of any article that works with less than 4 moments, while for several consistency results bounded variables or finite moments of all order are assumed.

2.3 The estimator

We will explain now the idea behind estimation of Ψ. Similar as in Bosq [1] it is based on a finite basis approxmation of the operator. To achieve optimal approximation in finite dimension, we choose eigenfunctions of the covariance operator $C = E[X_1 \otimes X_1]$ as our basis. Another expansion based on *predictive factors* has been proposed by Kargin and Onatski [19]. Here the intention is to use directions that minimize the prediction error in the autoregressive context.

Throughout this paper we use next to the covariance operator C the cross-covariance operator $\Delta = E[X_1 \otimes Y_1]$. By Assumption (A) both of them are Hilbert-Schmidt operators.

Now let $(\lambda_i, v_i)_{i\geq 1}$ be the eigenvalues and corresponding eigenfunctions of the operator C, such that $\lambda_1 \geq \lambda_2 \geq ....$ The eigenfunctions are orthonormal and those belonging to a non-zero eigenvalue form an orthonormal basis of $C(H_1) = \mathrm{Im}(C)$. Note that with probability one we have $X \in \mathrm{Im}(C)$. Since $\mathrm{Im}(C)$ is again a Hilbert-space, it is no real restriction to assume that $H_1 = \mathrm{Im}(C)$, i.e. that the operator is of full rank. In this case all eigenvalues are strictly positive. Using linearity of Ψ and the requirement $X_k \perp \varepsilon_k$ we obtain

$$\begin{aligned}
\Delta(v_j) &= E\langle X_1, v_j\rangle_{H_1} Y_1 \\
&= E\langle X_1, v_j\rangle_{H_1} \Psi(X_1) + E\langle X_1, v_j\rangle_{H_1} \varepsilon_1 \\
&= \Psi(E\langle X_1, v_j\rangle_{H_1} X_1) \\
&= \Psi(C(v_j)) \\
&= \lambda_j \Psi(v_j).
\end{aligned}$$

Then for any $x \in H_1$ the derived equation leads us to the representation

$$\Psi(x) = \Psi\left(\sum_{j=1}^{\infty} \langle v_j, x\rangle v_j\right) = \sum_{j=1}^{\infty} \frac{\Delta(v_j)}{\lambda_j} \langle v_j, x\rangle. \tag{2.1}$$

We assume here implicitely that $\dim(H_1) = \infty$. If $\dim(H_1) = M < \infty$, then (2.1) still holds with ∞ replaced by M. In fact, here the theory would become much simpler. To avoid distinguishing between

the different cases we will exlusively work in the infinite dimensional setup. Equation (2.1) gives a core idea for estimation of Ψ. We will estimate Δ, v_j and λ_j from our sample $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ and substitute the estimators into formula (2.1). The estimated eigenelements $(\hat{\lambda}_{j,n}, \hat{v}_{j,n};\, 1 \leq j \leq n)$ will be obtained from the empirical covariance operator

$$\hat{C}_n = \frac{1}{n} \sum_{k=1}^{n} X_k \otimes X_k.$$

In a similar straightforward manner we set

$$\hat{\Delta}_n = \frac{1}{n} \sum_{k=1}^{n} X_k \otimes Y_k.$$

For ease of notation we will suppress in the sequel the dependence on the sample size n of these estimators.

Apparently, from the finite sample we cannot estimate the entire sequence (λ_j, v_j), rather we have to work with a truncated version. This leads to

$$\hat{\Psi}_K(x) = \sum_{j=1}^{K} \frac{\hat{\Delta}(\hat{v}_j)}{\hat{\lambda}_j} \langle \hat{v}_j, x \rangle, \tag{2.2}$$

where the choice of $K = K_n$ is crucial. Notice that since we want our estimator to be consistent, K_n has to grow with the sample size to infinity in order to cover all the summands. On the other hand we know that $\lambda_j \to 0$ and hence it will be a delicate issue to control the behavior of $\frac{1}{\hat{\lambda}_j}$. A small error in the estimation of λ_j can have an enormous impact on (2.2). The usual approach is to relate K_n on the decay-rate of $\{\lambda_j\}$. For example Cardot et al. [4] assume that $n\lambda_{K_n}^4 \to \infty$ and $\frac{n\lambda_{K_n}^2}{(\sum_{j=1}^{K_n} \frac{1}{\alpha_j})^2} \to \infty$, when

$$\alpha_1 = \lambda_1 - \lambda_2 \quad \text{and} \quad \alpha_j = \min\{\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1}\}, \quad j > 1. \tag{2.3}$$

Similar requirements are used in other papers (see e.g. Theorem 8.7 in [2] or Assumption (B.5) in Yao et al. [25]). We will avoid any such assumptions by suggesting K_n that is purlely *data-driven*.

2.4 Consistency results

For our first result, Theorem 2.1, the sole assumptions on the spectrum and on Ψ are:

(B): *The eigenvalues $\{\lambda_j\}$ are mutually distinct and Ψ is a Hilbert-Schmidt operator.*

Assuming distinct eigenvalues is standard in functional data analysis and is commonly used for results involving functional principal components. Without this assumption the eigenfunctions in representation (2.1) are no longer identifiable. In Theorem 2.2 we will show that in practice we can completely avoid Assumption **(B)**. The K_n we use in Theorem 2.1, is given as follows:

(K): *Let $K_n = \min(B_n, E_n, m_n)$ where $B_n = \arg\max\{j \geq 1 | \frac{1}{\hat{\lambda}_j} \leq m_n\}$ and $E_n = \arg\max\{k \geq 1 | \max_{1 \leq j \leq k} \frac{1}{\hat{\alpha}_j} \leq m_n\}$ for some sequence $\{m_n\}$ such $m_n^6 = o(n)$. Here $\hat{\lambda}_j$ and $\hat{\alpha}_j$ are the estimates for λ_j and α_j (given in (2.3)), respectively, obtained from $\hat{C}$.*

The choice of K_n is motivated by a 'bias variance trade-off' argument. If an eigenvalue is very small (in our case $\ll 1/m_n$) it means that the direction it explains is not very important in the representation of X_k. Therefore excluding it from the representation of Ψ will not cause a big bias whereas it will considerably reduce the variance. It will be only included if the sample size is big enough, in which case we can hope for a reasonable accuracy of $\hat{\lambda}_j$. Since all the quantities involved can be computed from the sample our procedure can be a fast alternative to cross-validation or AIC type criteria as suggested in [25] for the choice of K in practical applications. In practice it is recommended to replace $\frac{1}{\hat{\lambda}_j}$ in the definition of B_n by $\frac{\hat{\lambda}_1}{\hat{\lambda}_j}$ and $\frac{1}{\hat{\alpha}_j}$ in the definition of E_n by $\frac{\hat{\lambda}_1}{\hat{\alpha}_j}$ to adapt for scaling. For the asymptotics such a modification has no influence.

Theorem 2.1. *Consider the linear Hilbertian model* (1.1) *and assume that Assumptions* **(A)** *and* **(B)** *hold and* $\{K_n\}$ *is defined as in* **(K)**. *Then the estimator described in Section 2.3 is weakly consistent, i.e.* $\|\hat{\Psi}_{K_n} - \Psi\|_{\mathcal{L}_{12}} \xrightarrow{P} 0$, *if* $n \to \infty$.

The technical Assumption **(B)** appears still unsatisfactory. Unfortunately it cannot be completely avoided. To see this, assume for example that Ψ is the identity operator, which is not Hilbert-Schmidt anymore. Then for any ONB $\{v_i\}$ we have $\Psi = \sum_{i\geq 1} v_i \otimes v_i$. Even if from the finite sample our estimators for $v_1, \ldots, v_K$ would be perfect ($v_i = \hat{v}_i$) we have $\|\Psi - \hat{\Psi}_K\|_{\mathcal{L}_{12}} = 1$ for any $K \geq 1$. This is easily seen by evaluating Ψ and $\hat{\Psi}_K$ at v_{K+1}.

A way to overcome such difficulties it to argue that in practice we will be satisfied if the estimator $\hat{\Psi}$ is such that $\|\Psi(X) - \hat{\Psi}(X)\|$ is small if $X \stackrel{d}{=} X_1$. E.g. if $\langle X, v\rangle = 0$ with probability one, then the direction v plays no role for describing X and a larger value of $\|\Psi(v) - \hat{\Psi}(v)\|$ doesn't pose a problem if for example we are interested in prediction. The next theorem shows that we can further simplify the assumptions if we are only interested in showing $\|\Psi(X) - \hat{\Psi}(X)\|_{H_2} \xrightarrow{P} 0$. In particular, **(B)** can be dropped.

Theorem 2.2. *Let Assumption* **(A)** *hold and define the estimator* $\hat{\Psi}_{K_n}$ *as in Section 2.3 with* $K_n = \arg\max\{j \geq 1 | \frac{\hat{\lambda}_1}{\hat{\lambda}_j} \leq m_n\}$, *where* $m_n = o(\sqrt{n})$. *Then* $\|\Psi(X) - \hat{\Psi}_{K_n}(X)\|_{H_2} \xrightarrow{P} 0$.

This result should be compared to Theorem 3 in Crambes and Mas [7] where an asymptotic expansion of $E\|\Psi(X) - \hat{\Psi}_k(X)\|^2_{H_2}$ is obtained (for fixed k). Their result implies consistency, but requires again assumptions on the decay rate of $\{\lambda_i\}$, that Ψ is Hilbert-Schmidt, and the existence of moments of all order of the X_k.

It should be noted that we are studying in this paper only convergence in probability, whereas for example [2] or [4] have also obtained results on almost sure convergence, but for the price of further technical assumptions (e.g. boundedness of the X_k's).

An obvious question is which rate of growth for m_n is optimal (in some sense to be specified). Though desirable, we belief that optimality results will be extremely difficult under the very general conditions of this paper. Rates of convergence seem to require more information on the spectrum or the operator Ψ and this is exactly what we wanted to avoid here. In Section 3 we perform a simulation study, which suggests that Theorem 2.2 will remain true if we use $m_n = O(\sqrt{n})$. In fact, for large n we see that $m_n = \sqrt{n}$ performs better than $m_n = 0.1\sqrt{n}$ throughout all constellations that we tested.

2.5 Applications to functional time series

We show that our framework covers the ARH(1) model of Bosq [2]. With i.i.d. innovations $\delta_k \in L_H^4$ the process $\{X_k\}$ defined via $X_{k+1} = \Psi(X_k) + \delta_{k+1}$ is L_H^4–approximable if $\Psi \in \mathcal{L}(H,H)$ such that $\|\Psi\|_{\mathcal{L}(H,H)} < 1$, see [17]. The stationary solution for X_k has the form

$$X_k = \sum_{j\geq 0} \Psi^j(\delta_{k-j}).$$

Setting $\varepsilon_k = \delta_{k+1}$ and $Y_k = X_{k+1}$ we obtain the linear model (1.1). Independence of $\{\delta_k\}$ implies that $X_k \perp \varepsilon_k$ and hence Assumption *(A)* holds. Bosq [2] has obtained a (strongly) consistent estimator of Ψ, if Ψ is Hilbert-Schmidt and again by imposing assumptions on the spectrum of C.

In our approach we don't even need that the innovations $\{\delta_k\}$ are i.i.d. As long as we can assure that $\{\delta_k\}$ and $\{X_k\}$ are L^4–m–approximable we only need that $\{\delta_k\}$ is H-white noise. Indeed, denoting A^* the conjugate of operator A, we have for any $x \in H_1$ and $y \in H_2$ that

$$\begin{aligned} E\langle X_k, x\rangle_{H_1}\langle \varepsilon_k, y\rangle_{H_2} &= \sum_{j\geq 0} E\langle \Psi^j(\delta_{k-j}), x\rangle_{H_1}\langle \delta_{k+1}, y\rangle_{H_2} \\ &= \sum_{j\geq 0} E\langle \delta_{k-j}, (\Psi^j)^*(x)\rangle_{H_1}\langle \delta_{k+1}, y\rangle_{H_2} = 0. \end{aligned}$$

This shows $X_k \perp \varepsilon_k$ and Assumption **(A)** follows.

We obtain the following

Corollary 2.1. *Let $\{X_n\}_{n\geq 1}$ be an ARH(1) process given by the recurrence equation $X_{n+1} = \Psi(X_n) + \varepsilon_{n+1}$. Assume $\|\Psi\|_{\mathcal{L}_{12}} < 1$. If $\{\varepsilon_i\}$ is H-white noise and Assumption* **(A)** *holds, then for the estimator $\hat{\Psi}_K$ given in Theorem 2.2 we have $\|\Psi(X) - \hat{\Psi}_K(X)\|_{H_2} \xrightarrow{P} 0$. In particluar if $\{\varepsilon_i\}$ is i.i.d. in L_H^4, Assumption* **(A)** *will hold.*

Corollary 2.2. *Let $\{X_n\}_{n\geq 1}$ be an ARH(1) process given by the recurrence equation $X_{n+1} = \Psi(X_n) + \varepsilon_{n+1}$. Assume $\|\Psi\|_{\mathcal{S}_{12}} < 1$. If $\{\varepsilon_i\}$ is H-white noise and Assumptions* **(A)** *and* **(B)** *hold, then the estimator $\hat{\Psi}_K$ given in Theorem 2.1 is consistent.*

Another possible application of our result refers to a recently introduced functional version of the celibrated ARCH model (Hörmann et al. [15]), which plays a fundamental role in financial econometrics. It is given by the two equations

$$y_k(t) = \varepsilon_k(t)\sigma_k(t), \quad t \in [0,1],\, k \in \mathbb{Z}$$

and

$$\sigma_k^2(t) = \delta(t) + \int_0^1 \beta(t,s) y_{k-1}^2(s) ds, \quad t \in [0,1],\, k \in \mathbb{Z}.$$

Without going into details, let us just mention that using a trick one can write the squared observations of a functional ARCH model as an autoregressive process with innovations $\nu_k(t) = y_k^2(t) - \sigma_k^2(t)$. The new noise $\{\nu_k\}$ is no longer independent and hence the results of [2] are not applicable to prove consistency of the involved estimator for the operator β. But it is shown in [15] that the innovations of this new process form Hilbertian white noise and that the new process is L^4–m–approximable. This allows us to obtain a consistent estimator for β.

3 Simulation study

Didericksen et al. [10] have investigated in an empirical study the performance of different estimators for the regression operator Ψ in the FAR(1) setup $X_{n+1} = \Psi(X_n) + \varepsilon_n$. For their study they compare different kernel operators Ψ. By the smoothness of the chosen kernels, curves X_n are mapped to quite smooth and flat curves $\Psi(X_n)$, even if X_n is irregularly shaped. In such a setup Didericksen et al. [10] conclude that choosing $K = 3$ or $K = 4$ gives broadly speaking the best results among all chosen setups. We have choosen operators Ψ that produce more distinctive curves $\Psi(X_n)$. Our conclusion below is that if the spectrum of $EX \otimes X$ is not decaying very fast, we need to choose K much bigger than 3 or 4 to get good estimates. This is true even for moderate sample sizes. We show that our procedure proposes $K = K_n$ that are close to the optimal $K = K_n^{OPT}$.

3.1 Setup

For the simulation study we obviously have to work with finite dimensional spaces H_1 and H_2. However, because of the asymptotical nature of our results, we set the dimension relatively high and define $H_1 = H_2 = \text{span}\{v_j : 0 \le j \le 34\}$, where $v_0(t) = 1$, $v_{2k-1}(t) = \sin(2\pi kt)$ and $v_{2k}(t) = \cos(2\pi kt)$ are the first 35 elements of a Fourier basis on $[0, 1]$. We work with Gaussian curves $X_i(t)$ by setting

$$X_i(t) = \sum_{j=1}^{35} A_i^{(j)} v_{j-1}(t), \tag{3.1}$$

where $(A_1^{(j)}, A_2^{(j)}, \ldots, A_{35}^{(j)})'$ are independent Gaussian random vectors with mean zero and covariance Σ. This setup allows us to easily manipulate the eigenvalues $\{\lambda_i\}$ of a covariance operator $C_X = EX \otimes X$. Indeed, if we define $\Sigma = \text{diag}(a_1, \ldots, a_{35})$, then $\lambda_i = \sqrt{a_i}$ and v_i is the corresponding eigenfunction. We test three sets of eigenvalues:

- $\Lambda_1 : (1, e^{-1/5}, e^{-2/5}, \ldots, e^{-35/5})$ [fast decay],
- $\Lambda_2 : (1, \frac{34}{35}, \ldots, \frac{1}{35})$ [slow decay],
- $\Lambda_3 : (1, 1, \ldots, 1)$ [no decay].

The noise $\{\varepsilon_k\}$ is also assumed to be of the form (3.1) with coefficients $\{A_i^{(j)},\ i, j \ge 0\}$ i.i.d. $\mathcal{N}(0, \sigma^2)$ and $\sigma^2 \in \{0.25, 1, 2.25, 4\}$. Finally we used the following 3 operators:

- Ψ_1 identity,
- $\Psi_2 = \Gamma_1 + \Gamma_2$, such that $\Gamma_1 : v_i \mapsto \frac{2}{3} v_{\pi_i}$ and $\Gamma_2 : v_i \mapsto \frac{1}{3} v_{\pi_i'}$, where $\pi_i = 1 + (i + 4 \bmod 35)$ and $\pi_i' = 1 + (i \bmod 35)$,
- $\Psi_3(x) = \sum_{i=1}^{35} \sum_{j=1}^{35} \psi_{ij} \langle x, v_i \rangle v_j$, where the coefficients ψ_{ij} have been generated as i.i.d. standard normal random variables (once generated, they were fixed for the entire simulation), normalized such that $\|\Psi_3\|_{\mathcal{L}_{12}} = 1$.

Figure 1 shows application of Ψ_i $(i = 1, 2, 3)$ on four realizations of the process X and the corresponding curves $\Psi_i(X)$, $\hat{\Psi}_i(X)$ and Y.

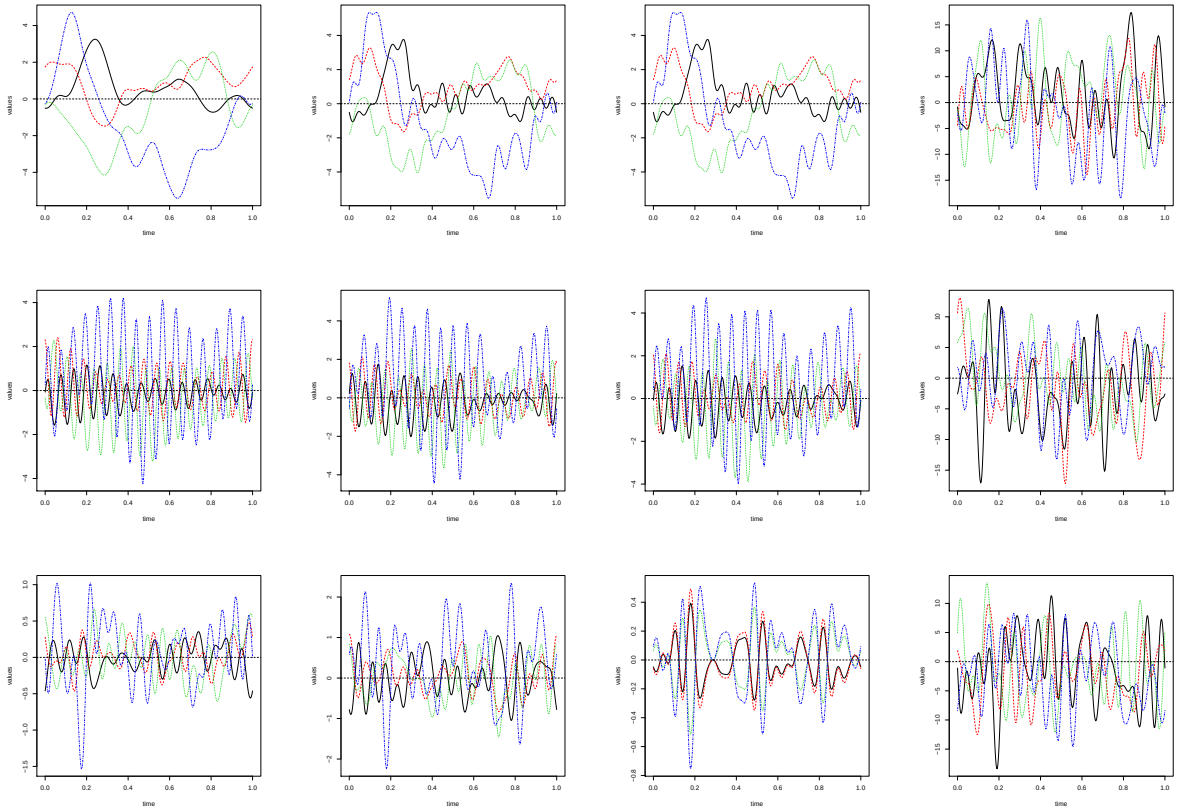

Figure 1: Columns 1 to 4 correspond to $\Psi_i(X), \hat{\Psi}_i(X), \hat{\Psi}_i^{OPT}(X)$ and $\Psi_i(X) + \varepsilon$ $(i = 1, 2, 3)$, respectively. Here $\hat{\Psi}_i$ and $\hat{\Psi}_i^{OPT}$ are operators given by the formula (2.2) with K obtained by our procedure and with the optimal one in terms of `NMSE` (see (3.2)). Estimators were computed for $n = 1280$. The same 4 curves were used with each operator. They were drawn from a distribution indicated by Λ_1 and they are presented at the top-left chart ($\Psi_1 = \mathrm{Id}$).

3.2 Results

As a performance measure for our procedure we used a normalized mean square error defined as

$$\texttt{NMSE} = \frac{\sum_{k=1}^{n} \|\Psi(X_k) - \hat{\Psi}(X_k)\|_{H_2}^2}{\sum_{k=1}^{n} \|\Psi(X_k)\|_{H_2}^2}. \tag{3.2}$$

Following Theorem 2.2 we chose $m_n = c\sqrt{n}$ with the 3 different constants $c = 0.1, 0.5, 1$ and sample sizes $n = 10 \times 2^\ell$, $\ell = 0, \ldots, 11$. The `NMSE` and the size of $K = K_n^c$ is shown for different constellations in the Appendix. We display the results only for $\sigma = 1$. Not surprisingly, the bigger the variance of the noise, the bigger `NMSE` but otherwise our findings were the same across all constellations of σ. The column K_n^{OPT} shows the value of K that gave the smallest `NMSE` among all possible values $K = 1, \ldots, 35$.

The tables in the Appendix show that the choice of K proposed by our method is very satisfactory and close to K_n^{OPT} (which gives the smallest `NMSE`) or at least that the corresponding `NMSE`'s were comparable. We can see that for small sample sizes it is preferable to use $c = 0.1$ while for large n it turns out that $c = 1$ performs best.

4 Proofs

Throughout this entire section we assume the setup and notation of Section 2.2.

4.1 Proof of Theorem 2.1

We work under Assumptions **(A)** and **(B)** and with K_n given in **(K)**. The first important lemma which we use in the proof of Theorem 2.1 is an error bound for the estimators of the operators Δ and C. Below we extend results in [17].

Lemma 4.1. *There is a constant U depending only on the law of $\{(X_k, Y_k)\}$ such that*

$$n \max\{E\|\Delta - \hat{\Delta}_n\|_{\mathcal{S}_{12}}, E\|C - \hat{C}_n\|_{\mathcal{S}_{11}}\} < U.$$

Proof of Lemma 4.1. We only prove the bound for Δ, the one for C is similar. First note that by Lemma 2.1 in [17] and Assumption **(A)** $\{Y_k\}$ is also L^4–m–approximable. Next we observe that

$$nE\big\|\Delta - \hat{\Delta}_n\big\|_{\mathcal{S}_{12}}^2 = nE\left\|\frac{1}{n}\sum_{k=1}^{n} Z_k\right\|_{\mathcal{S}_{12}}^2,$$

where $Z_k = X_k \otimes Y_k - \Delta$. Set $Z_k^{(r)} = X_k^{(r)} \otimes Y_k^{(r)} - \Delta$. Using the stationarity of the sequence $\{Z_k\}$ we obtain

$$\begin{aligned} nE\left\|\frac{1}{n}\sum_{k=1}^{n} Z_k\right\|_{\mathcal{S}_{12}}^2 &= \sum_{|r|<n}\left(1 - \frac{|r|}{n}\right) E\langle Z_0, Z_r\rangle_{\mathcal{S}_{12}} \\ &\leq E\|Z_0\|_{\mathcal{S}_{12}}^2 + 2\sum_{r=1}^{\infty} |E\langle Z_0, Z_r\rangle_{\mathcal{S}_{12}}|. \end{aligned} \tag{4.1}$$

By the Cauchy-Schwarz inequality and the independence of $Z_r^{(r-1)}$ and Z_0 we derive:

$$|E\langle Z_0, Z_r\rangle_{\mathcal{S}_{12}}| = |E\langle Z_0, Z_r - Z_r^{(r-1)}\rangle_{\mathcal{S}_{12}}| \leq (E\|Z_0\|_{\mathcal{L}_{12}}^2)^{\frac{1}{2}} (E\|Z_r - Z_r^{(r-1)}\|_{\mathcal{S}_{12}}^2)^{\frac{1}{2}}.$$

Using $\|X_0 \otimes Y_0\|_{\mathcal{S}_{12}} = \|X_0\|_{H_1}\|Y_0\|_{H_2}$ and again the Cauchy-Schwarz inequality we get

$$E\|Z_0\|_{\mathcal{L}_{12}}^2 = E\|X_0\|_{H_1}^2\|Y_0\|_{H_2}^2 \leq \nu_{4,H_1}^2(X_0)\nu_{4,H_2}^2(Y_0) < \infty.$$

To finish the proof we show that $\sum_{r=1}^{\infty}(E\|Z_r - Z_r^{(r-1)}\|_{\mathcal{L}_{12}}^2)^{\frac{1}{2}} < \infty$. By using an inequality of the type $|ab - cd|^2 \leq 2|a|^2|b - d|^2 + 2|d|^2|a - c|^2$ we obtain

$$E\|Z_r - Z_r^{(r-1)}\|_{\mathcal{S}_{12}}^2 = \|X_r \otimes Y_r - X_r^{(r-1)} \otimes Y_r^{(r-1)}\|_{\mathcal{S}_{12}}^2$$

$$\leq 2E\|X_r\|_{H_1}^2\|Y_r - Y_r^{(r-1)}\|_{H_2}^2 + 2E\|Y_r^{(r-1)}\|_{H_2}^2\|X_r - X_r^{(r-1)}\|_{H_1}^2$$
$$\leq 2\nu_{4,H_1}^2(X_r)\nu_{4,H_2}^2(Y_r - Y_r^{(r-1)}) + 2\nu_{4,H_2}^2(Y_r^{(r-1)})\nu_{4,H_1}^2(X_r - X_r^{(r-1)}).$$

Convergence of (4.1) follows now directly from L^4-m–approximability. □

Application of this lemma leads also to bounds for estimators of eigenvalues and eigenfunctions of C via the following two lemmas (see [17]).

Lemma 4.2. *Suppose $\lambda_i, \hat{\lambda}_i$ are the eigenvalues of C and $\hat{C}$, respectively, listed in decreasing order. Let $v_i, \hat{v}_i$ be the corresponding eigenvectors and let $\hat{c}_i = \langle v_i, \hat{v}_i \rangle$. Then for each $j \geq 1$,*

$$\hat{\alpha}_j\|v_j - \hat{c}_j\hat{v}_j\|_{H_1} \leq 2\sqrt{2}\|\hat{C} - C\|_{\mathcal{L}_{11}},$$

where $\hat{\alpha}_j = \min\{\hat{\lambda}_{j-1} - \hat{\lambda}_j, \hat{\lambda}_j - \hat{\lambda}_{j+1}\}$ and $\hat{\alpha}_1 = \hat{\lambda}_2 - \hat{\lambda}_1$.

Lemma 4.3. *Let $\lambda_j, \hat{\lambda}_j$ be defined as in Lemma 4.2. Then for each $j \geq 1$,*

$$|\lambda_j - \hat{\lambda}_j| \leq \|C - \hat{C}\|_{\mathcal{L}_{11}}.$$

In the following calculations we work with finite sums of the representation in (2.1):

$$\Psi_K(x) = \sum_{j=1}^{K} \frac{\Delta(v_j)}{\lambda_j}\langle v_j, x\rangle. \tag{4.2}$$

In order to prove the main result we consider the term $\|\Psi - \hat{\Psi}_K\|_{\mathcal{L}_{12}}$ and decompose it using the triangle inequality into four terms

$$\|\Psi - \hat{\Psi}_K\|_{\mathcal{L}_{12}} \leq \sum_{i=1}^{4} \|S_i(K)\|_{\mathcal{L}_{12}},$$

where

$$S_1(K) = \sum_{j=1}^{K}\left(\hat{c}_j\hat{v}_j \otimes \frac{\hat{\Delta}(\hat{c}_j\hat{v}_j)}{\hat{\lambda}_j} - \hat{c}_j\hat{v}_j \otimes \frac{\Delta(\hat{c}_j\hat{v}_j)}{\hat{\lambda}_j}\right), \tag{4.3}$$

$$S_2(K) = \sum_{j=1}^{K}\left(\hat{c}_j\hat{v}_j \otimes \frac{\Delta(\hat{c}_j\hat{v}_j)}{\hat{\lambda}_j} - \hat{c}_j\hat{v}_j \otimes \frac{\Delta(\hat{c}_j\hat{v}_j)}{\lambda_j}\right), \tag{4.4}$$

$$S_3(K) = \sum_{j=1}^{K}\left(\hat{c}_j\hat{v}_j \otimes \frac{\Delta(\hat{c}_j\hat{v}_j)}{\lambda_j} - v_j \otimes \frac{\Delta(v_j)}{\lambda_j}\right), \tag{4.5}$$

$$S_4(K) = \Psi - \Psi_K. \tag{4.6}$$

The following simple lemma gives convergence of $S_4(K_n)$, provided $K_n \xrightarrow{P} \infty$.

Lemma 4.4. *Let $\{K_n, n \geq 1\}$ be a random sequence taking values in $\mathbb{N}$, such that $K_n \xrightarrow{P} \infty$ as $n \to \infty$. Then Ψ_{K_n} defined by the equation (4.2) converges to Ψ in probability.*

Proof. Notice that since $\|\Psi\|^2_{\mathcal{S}_{12}} = \sum_{j=1}^{\infty} \|\Psi(v_j)\|^2_{H_2} < \infty$ for some orthonormal base $\{v_j\}$, we can find $m_\varepsilon \in \mathbb{N}$ such that $\|\Psi - \Psi_m\|^2_{\mathcal{S}_{12}} = \sum_{j>m} \|\Psi(v_j)\|^2_{H_2} \leq \varepsilon$, whenever $m > m_\varepsilon$. Hence

$$\begin{aligned} P(\|\Psi - \Psi_{K_n}\|^2_{\mathcal{S}_{12}} > \varepsilon) &= \sum_{m=1}^{\infty} P(\|\Psi - \Psi_m\|^2_{\mathcal{S}_{12}} > \varepsilon \cap K_n = m) \\ &= P(K_n \leq m_\varepsilon). \end{aligned}$$

□

The next three lemmas deal with terms (4.3)–(4.5).

Lemma 4.5. *Let $S_1(K)$ be defined by the equation* (4.3) *and U the constant derived in Lemma 4.1. Then*

$$P(\|S_1(K_n)\|_{\mathcal{L}_{12}} > \varepsilon) \leq \frac{U m_n^2}{\varepsilon^2 n}.$$

Proof. Note that for an orthonormal system $\{e_i \in H_1 \mid i \geq 1\}$ and any sequence $\{x_i \in H_2 \mid i \geq 1\}$ the following identity holds:

$$\left\| \sum_{i=1}^{K} e_i \otimes x_i \right\|^2_{\mathcal{S}_{12}} = \sum_{j=1}^{\infty} \left\| \sum_{i=1}^{K} \langle e_i, e_j \rangle x_i \right\|^2_{H_2} = \sum_{j=1}^{K} \|x_j\|^2_{H_2}. \tag{4.7}$$

Using this and the fact that the Hilbert-Schmidt norm bounds the operator norm we derive

$$\begin{aligned} P(\|S_1(K_n)\|^2_{\mathcal{L}_{12}} > \varepsilon) &\leq P\left(\left\| \sum_{j=1}^{K_n} \hat{c}_j \hat{v}_j \otimes \frac{1}{\hat{\lambda}_j} (\hat{\Delta} - \Delta)(\hat{c}_j \hat{v}_j) \right\|^2_{\mathcal{S}_{12}} > \varepsilon \right) \\ &\leq P\left(\frac{1}{\hat{\lambda}^2_{K_n}} \sum_{j=1}^{K_n} \|(\hat{\Delta} - \Delta)(\hat{c}_j \hat{v}_j)\|^2_{H_2} > \varepsilon \right) \\ &\leq P(m_n^2 \|\hat{\Delta} - \Delta\|^2_{\mathcal{S}_{12}} > \varepsilon). \end{aligned}$$

By the Markov inequality

$$P(\|S_1(K_n)\|^2_{\mathcal{L}_{12}} > \varepsilon) \leq E\|\hat{\Delta} - \Delta\|^2_{\mathcal{S}_{12}} \frac{m_n^2}{\varepsilon} \leq U \frac{m_n^2}{\varepsilon n},$$

where the last inequality is obtained from Lemma 4.1. □

Lemma 4.6. *Let $S_2(K)$ be defined by the equation* (4.4) *and U the constant from Lemma 4.5. Then*

$$P(\|S_2(K_n)\|_{\mathcal{L}_{12}} > \varepsilon) \leq 4U \|\Delta\|^2_{\mathcal{S}_{12}} \frac{m_n^4}{\varepsilon^2 n}.$$

Proof. Assumption $K_n \leq B_n$ and identity (4.7) imply

$$P(\|S_2(K_n)\|^2_{\mathcal{L}_{12}} > \varepsilon) = P\left(\left\| \sum_{j=1}^{K_n} \left(\frac{1}{\lambda_j} - \frac{1}{\hat{\lambda}_j} \right) \hat{c}_j \hat{v}_j \otimes \Delta(\hat{c}_j \hat{v}_j) \right\|^2_{\mathcal{L}_{12}} > \varepsilon \right)$$

$$\leq P\left(\max_{j=1}^{K_n} \left(\frac{\hat{\lambda}_j - \lambda_j}{\hat{\lambda}_j \lambda_j} \right)^2 \sum_{j=1}^{K_n} \|\Delta(\hat{c}_j \hat{v}_j)\|_{H_2}^2 > \varepsilon \right)$$
$$\leq P\left(\max_{j=1}^{K_n} \left(\frac{\hat{\lambda}_j - \lambda_j}{\lambda_j} \right)^2 > \frac{\varepsilon}{m_n^2 \|\Delta\|_{\mathcal{S}_{12}}^2} \right).$$

For simplifying the notation let $b^2 = \frac{\varepsilon}{m_n^2 \|\Delta\|_{\mathcal{S}_{12}}^2}$, then

$$P(\|S_2(K_n)\|_{\mathcal{L}_{12}}^2 > \varepsilon) \leq P\left(\max_{j=1}^{K_n} \left| \frac{\hat{\lambda}_j - \lambda_j}{\lambda_j} \right| > b \right)$$
$$\leq P\left(\frac{1}{\lambda_{K_n}} \max_{j=1}^{K_n} |\hat{\lambda}_j - \lambda_j| > b \ \cap \ \max_{j=1}^{K_n} |\hat{\lambda}_j - \lambda_j| \leq \frac{b}{2m_n} \right) + P\Big(\max_{j=1}^{K_n} |\hat{\lambda}_j - \lambda_j| > \frac{b}{2m_n} \Big).$$

The first summand vanishes because

$$P\left(\frac{1}{\lambda_{K_n}} \max_{j=1}^{K_n} |\hat{\lambda}_j - \lambda_j| > b \ \cap \ \max_{j=1}^{K_n} |\hat{\lambda}_j - \lambda_j| \leq \frac{b}{2m_n} \right)$$
$$\leq P\left(\frac{b}{2\lambda_{K_n} m_n} > b \ \cap \ |\hat{\lambda}_{K_n} - \lambda_{K_n}| \leq \frac{b}{2m_n} \right)$$
$$\leq P\left(\frac{1}{2m_n} > \lambda_{K_n} \ \cap \ |\hat{\lambda}_{K_n} - \lambda_{K_n}| \leq \frac{\sqrt{\varepsilon}}{m_n^2 2 \|\Delta\|_{\mathcal{S}_{12}^2}} \right),$$

which is equal to 0 for n large enough, since $\hat{\lambda}_{K_n} \geq \frac{1}{m_n}$ and the distance between λ_{K_n} and $\hat{\lambda}_{K_n}$ shrinks faster than $\frac{1}{2m_n}$. For the second term we use Lemma 4.3 and the Markov inequality:

$$P(\|S_2(K_n)\|_{\mathcal{L}_{12}}^2 > \varepsilon) \leq P\Big(\max_{j=1}^{K_n} |\hat{\lambda}_j - \lambda_j| > \frac{b}{2m_n} \Big)$$
$$\leq P\Big(\|\hat{C} - C\|_{\mathcal{L}_{11}} > \frac{b}{2m_n} \Big)$$
$$\leq \frac{4m_n^2}{b^2} E\|\hat{C} - C\|_{\mathcal{L}_{11}}^2$$
$$\leq 4U \|\Delta\|_{\mathcal{S}_{12}}^2 \frac{m_n^4}{\varepsilon n}.$$

□

Lemma 4.7. *Let $S_3(K)$ be defined by (4.5) and U be the constant defined in Lemma 4.5, then*

$$P(\|S_3(K_n)\|_{\mathcal{L}_{12}} < \varepsilon) \leq U(128\|\Delta\|_{\mathcal{L}_{12}}^2 + 4\varepsilon^2) \frac{m_n^6}{\varepsilon^2 n}.$$

Proof. By adding and subtracting the term $\hat{c}_j \hat{v}_j \Delta(v_j)$ and using the triangle inequality we derive

$$P(\|S_3(K_n)\|_{\mathcal{L}_{12}} > \varepsilon) = P\left(\left\| \sum_{j=1}^{K_n} \frac{1}{\lambda_j} (\hat{c}_j \hat{v}_j \otimes \Delta(\hat{c}_j \hat{v}_j) - v_j \otimes \Delta(v_j)) \right\|_{\mathcal{L}_{12}} > \varepsilon \right)$$

$$\leq P\left(\sum_{j=1}^{K_n}\frac{1}{\lambda_j}\|\hat{c}_j\hat{v}_j\otimes\Delta(\hat{c}_j\hat{v}_j-v_j)+(\hat{c}_j\hat{v}_j-v_j)\otimes\Delta(v_j)\|_{\mathcal{L}_{12}}>\varepsilon\right)$$
$$\leq P\left(\sum_{j=1}^{K_n}\frac{1}{\lambda_j}(\|\Delta\|_{\mathcal{L}_{12}}\|\hat{c}_j\hat{v}_j-v_j\|_{H_1}+\|\hat{c}_j\hat{v}_j-v_j\|_{H_1}\|\Delta\|_{\mathcal{L}_{12}})>\varepsilon\right).$$

Now we split $\Omega = A\cup A^c$ where $A=\{\frac{1}{\lambda_{K_n}}>2m_n\}$ and get

$$P(\|S_3(K_n)\|_{\mathcal{L}_{12}}>\varepsilon)\leq P\left(\frac{1}{\lambda_{K_n}}\sum_{j=1}^{K_n}\|\hat{c}_j\hat{v}_j-v_j\|_{H_1}>\frac{\varepsilon}{2\|\Delta\|_{\mathcal{L}_{12}}}\right)$$
$$\leq P\left(\sum_{j=1}^{K_n}\|\hat{c}_j\hat{v}_j-v_j\|_{H_1}>\frac{\varepsilon}{4m_n\|\Delta\|_{\mathcal{L}_{12}}}\right)+P\left(\frac{1}{\lambda_{K_n}}>2m_n\right). \tag{4.8}$$

For the first term in the inequality (4.8), by Lemma 4.2, definition of E_n and the Markov inequality we get

$$P\left(\sum_{j=1}^{K_n}\|\hat{c}_j\hat{v}_j-v_j\|_{H_1}>\frac{\varepsilon}{4m_n\|\Delta\|_{\mathcal{L}_{12}}}\right)\leq P\left(m_n\max_{j=1}^{E_n}\|\hat{c}_j\hat{v}_j-v_j\|_{H_1}>\frac{\varepsilon}{4m_n\|\Delta\|_{\mathcal{L}_{12}}}\right)$$
$$\leq P\left(\max_{j=1}^{E_n}\frac{2\sqrt{2}}{\hat{\alpha}_j}\|\hat{C}-C\|_{\mathcal{L}_{12}}>\frac{\varepsilon}{4m_n^2\|\Delta\|_{\mathcal{L}_{12}}}\right)$$
$$\leq P\left(\|\hat{C}-C\|_{\mathcal{L}_{12}}>\frac{\varepsilon}{8\sqrt{2}m_n^3\|\Delta\|_{\mathcal{L}_{12}}}\right)$$
$$\leq 128\|\Delta\|_{\mathcal{L}_{12}}^2m_n^6\frac{E\|\hat{C}-C\|_{\mathcal{L}_{12}}^2}{\varepsilon^2}$$
$$\leq 128U\|\Delta\|_{\mathcal{L}_{12}}^2\frac{m_n^6}{\varepsilon^2 n}.$$

Since $\hat{\lambda}_{K_n}\geq\frac{1}{m_n}$, the second term in the inequality (4.8) is bounded by

$$P\left(\lambda_{K_n}<\frac{1}{2m_n}\right)\leq P\left(\lambda_{K_n}<\frac{1}{2m_n}\ \cap\ |\hat{\lambda}_{K_n}-\lambda_{K_n}|\leq\frac{1}{2m_n}\right)+P\left(|\hat{\lambda}_{K_n}-\lambda_{K_n}|>\frac{1}{2m_n}\right)$$
$$\leq P\left(\|\hat{C}-C\|_{\mathcal{L}_{12}}>\frac{1}{2m_n}\right)$$
$$\leq 4m_n^2E\|\hat{C}-C\|_{\mathcal{L}_{12}}^2\leq 4U\frac{m_n^2}{n}.$$

Thus we derive

$$P(\|S_3(K_n)\|_{\mathcal{L}_{12}}>\varepsilon)\leq 128U\|\Delta\|_{\mathcal{L}_{12}}^2\frac{m_n^6}{\varepsilon^2 n}+4U\frac{m_n^2}{n}\leq U(128\|\Delta\|_{\mathcal{L}_{12}}^2+4\varepsilon^2)\frac{m_n^6}{\varepsilon^2 n}.$$

□

Finally we need a lemma which assures that K_n tends to infinity.

Lemma 4.8. *Let K_n be defined as in* **(K)**, *then* $K_n \xrightarrow{P} \infty$.

Proof. We have to show that $P(\min\{B_n, E_n\} < p) \to 0$ for any $p \in \mathbb{N}$. Since $\frac{1}{m_n} \searrow 0$, for n large enough we have, by combining Lemma 4.1 and 4.3, that

$$P(B_n < p) = P\left(\hat{\lambda}_p < \frac{1}{m_n}\right) = P\left(\lambda_p - \hat{\lambda}_p > \lambda_p - \frac{1}{m_n}\right) \leq P\left(|\hat{\lambda}_p - \lambda_p| > \lambda_p - \frac{1}{m_n}\right) \to 0.$$

□

Now we are ready to prove the main result

Proof of Theorem 2.1. First, by the triangle inequality we get

$$\begin{aligned}\|\Psi - \hat{\Psi}_{K_n}\|_{\mathcal{L}_{12}} &\leq \|\Psi - \hat{\Psi}_{K_n}\|_{\mathcal{L}_{12}} + \|\Psi - \Psi_{K_n}\|_{\mathcal{L}_{12}} \\ &\leq \|S_1(K_n)\|_{\mathcal{L}_{12}} + \|S_2(K_n)\|_{\mathcal{L}_{12}} + \|S_3(K_n)\|_{\mathcal{L}_{12}} + \|\Psi - \Psi_{K_n}\|_{\mathcal{L}_{12}}.\end{aligned}$$

By Lemmas 4.4, 4.5, 4.6, 4.7 and assumption $m_n^6 = o(n)$ we finally obtain for large enough n that

$$\begin{aligned}&P(\|\Psi - \hat{\Psi}_{K_n}\|_{\mathcal{L}_{12}} > \varepsilon) \\ &\leq U4^4\frac{m_n^2}{\varepsilon^4 n} + 4^3U\|\Delta\|_{\mathcal{S}_{12}}^2\frac{m_n^4}{\varepsilon^2 n} + 4^2U(128\|\Delta\|_{\mathcal{L}_{12}}^2 + \varepsilon^2/4)\frac{m_n^6}{\varepsilon^2 n} + P(\|\Psi - \Psi_{K_n}\|_{\mathcal{L}_{12}} > \varepsilon/4) \xrightarrow{n\to\infty} 0.\end{aligned}$$

□

4.2 Proof of Theorem 2.2

In order to simplify the notation we will denote $K = K_n$. This time as a starting point we take a representation of Ψ in the basis $\{\hat{v}_1, \hat{v}_2, ...\}$. Let $M_m = \overline{\text{sp}}\{v_1, v_2, ..., v_m\}$, $\hat{M}_m = \overline{\text{sp}}\{\hat{v}_1, \hat{v}_2, ..., \hat{v}_m\}$ where $\overline{\text{sp}}\{x_i,\, i \in I\}$ denotes the closed span of the elements $\{x_i,\, i \in I\}$. If $\text{rank}(\hat{C}) = \ell$, then $\{\hat{v}_i,\, i > \ell\}$ can be any ONB of $\hat{M}_\ell^\perp$. We write P_A for the projection operator which maps on a closed linear space A. As usual $A^\perp$ denotes the orthogonal complement of A. Since for any $m \geq 1$ we can write $x = P_{\hat{M}_m}(x) + P_{\hat{M}_m^\perp}(x)$, the linearity of Ψ and the projection operator gives

$$\begin{aligned}\Psi(x) &= \Psi(P_{\hat{M}_m}(x)) + \Psi(P_{\hat{M}_m^\perp}(x)) \\ &= \sum_{j=1}^{m}\langle \hat{v}_j, x\rangle_{H_1}\Psi(\hat{v}_j) + \Psi(P_{\hat{M}_m^\perp}(x)).\end{aligned}$$

Now we evaluate Ψ in some $\hat{v}_j$ which is not in the kernel of $\hat{C}$. By definitions of Ψ, $\hat{C}$ and again by linearity of the involved operators

$$\begin{aligned}\Psi(\hat{v}_j) &= \frac{1}{\hat{\lambda}_j}\Psi(\hat{C}(\hat{v}_j)) \\ &= \frac{1}{\hat{\lambda}_j}\frac{1}{n}\sum_{i=1}^{n}\langle X_i, \hat{v}_j\rangle_{H_1}\Psi(X_i)\end{aligned}$$

$$= \frac{1}{\hat{\lambda}_j}\frac{1}{n}\sum_{i=1}^{n}\langle X_i, \hat{v}_j\rangle_{H_1}(Y_i - \varepsilon_i)$$
$$= \frac{1}{\hat{\lambda}_j}(\hat{\Delta}(\hat{v}_j) + \hat{\Lambda}(\hat{v}_j)),$$

where $\hat{\Lambda} = -\frac{1}{n}\sum_{i=1}^{n} X_i \otimes \varepsilon_i$. Hence if m is such that $\hat{\lambda}_m > 0$ (which will now be implicitely assumed in the sequel), Ψ can be expressed as

$$\Psi(x) = \sum_{j=1}^{m}\langle \hat{v}_j, x\rangle_{H_1}\frac{1}{\hat{\lambda}_j}\hat{\Delta}(\hat{v}_j) + \sum_{j=1}^{m}\langle \hat{v}_j, x\rangle_{H_1}\frac{1}{\hat{\lambda}_j}\hat{\Lambda}(\hat{v}_j) + \Psi(P_{\hat{M}_m^{\perp}}(x)).$$

Note that the first term on the right-hand side is just $\hat{\Psi}_m(x)$. Therefore for any x, the distance between $\Psi(x)$ and $\hat{\Psi}_m(x)$ takes the following form

$$\|\Psi(x) - \hat{\Psi}_m(x)\|_{H_2} = \left\| \sum_{j=1}^{m}\langle \hat{v}_j, x\rangle_{H_1}\frac{1}{\hat{\lambda}_j}\hat{\Lambda}(\hat{v}_j) + \Psi(P_{\hat{M}_m^{\perp}}(x)) \right\|_{H_2}. \tag{4.9}$$

To assess (4.9) we need the following four lemmas.

Lemma 4.9. *Let $(\lambda_i, v_i)_{i\geq 1}$ and $(\hat{\lambda}_i, \hat{v}_i)_{i\geq 1}$ be eigenvalues and eigenfunctions of C and $\hat{C}$ respectively. Set $j, m \in \mathbb{N}$ such that $j \leq m \leq n$, then*

$$\|v_j - P_{\hat{M}_m}(v_j)\|_{H_1}^2 \leq 4\frac{\|C - \hat{C}\|_{\mathcal{L}_{11}}^2}{(\hat{\lambda}_{m+1} - \hat{\lambda}_j)^2}.$$

Proof. Note that by using Parseval's identity we get

$$\|v_j - P_{\hat{M}_m}(v_j)\|_{H_1}^2 = \sum_{k=1}^{\infty}\langle v_j - P_{\hat{M}_m}(v_j), \hat{v}_k\rangle_{H_1}^2 = \sum_{k>m}\langle v_j, \hat{v}_k\rangle_{H_1}^2.$$

Now

$$(\hat{\lambda}_{m+1} - \hat{\lambda}_j)^2\sum_{k>m}\langle v_j, \hat{v}_k\rangle_{H_1}^2 \leq \sum_{k>m}(\hat{\lambda}_k\langle v_j, \hat{v}_k\rangle_{H_1} - \hat{\lambda}_j\langle v_j, \hat{v}_k\rangle_{H_1})^2$$
$$= \sum_{k>m}(\langle v_j, \hat{C}(\hat{v}_k)\rangle_{H_1} - \hat{\lambda}_j\langle v_j, \hat{v}_k\rangle_{H_1})^2.$$

Since $\hat{C}$ is a self-adjoint operator, simple algebraic transformations yield

$$(\hat{\lambda}_{m+1} - \hat{\lambda}_j)^2\sum_{k>m}\langle v_j, \hat{v}_k\rangle_{H_1}^2 \leq \sum_{k>m}(\langle \hat{C}(v_j), \hat{v}_k\rangle_{H_1} - \hat{\lambda}_j\langle v_j, \hat{v}_k\rangle_{H_1})^2$$
$$= \sum_{k>m}(\langle (\hat{C} - C)(v_j), \hat{v}_k\rangle_{H_1} - (\hat{\lambda}_j - \lambda_j)\langle v_j, \hat{v}_k\rangle_{H_1})^2$$
$$\leq 2\sum_{k>m}|\langle (\hat{C} - C)(v_j), \hat{v}_k\rangle_{H_1}|^2 + 2\sum_{k>m}((\hat{\lambda}_j - \lambda_j)\langle v_j, \hat{v}_k\rangle_{H_1})^2.$$

By Parseval's inequality and Lemma 4.3

$$(\hat{\lambda}_{m+1} - \hat{\lambda}_j)^2\sum_{k>m}\langle v_j, \hat{v}_k\rangle_{H_1}^2 \leq 2\|(\hat{C} - C)(v_j)\|_{H_1}^2 + 2|\hat{\lambda}_j - \lambda_j|^2 \leq 4\|\hat{C} - C\|_{\mathcal{L}_{11}}^2.$$

□

Lemma 4.10. *Let Ψ, X be defined as in Lemma 2.2 and $K = K_n \xrightarrow{P} \infty$. Then $\|P_{M_K^\perp}(X)\|_{H_2} \xrightarrow{P} 0$.*

Proof. We first remark that for any $\varepsilon > 0$

$$P(\|P_{M_K^\perp}(X)\|_{H_2}^2 > \varepsilon) = P\left(\sum_{i=K+1}^{\infty} |\langle v_i, X\rangle_{H_1}|^2 > \varepsilon \right).$$

Since $\sum_{i=1}^{\infty} |\langle v_i, X\rangle_{H_1}|^2 = \|X\|_{H_1}^2$, there exists a random variable $J_\varepsilon \in \mathbb{R}$ such that $\sum_{i=J_\varepsilon}^{\infty} |\langle v_i, X\rangle_{H_1}|^2 < \varepsilon$. Since by assumption $E\|X\|_{H_1}^2 < \infty$, we conclude that J_ε is bounded in probability. Hence we obtain

$$\begin{aligned} P(\|P_{M_K^\perp}(X)\|_{H_2}^2 > \varepsilon) &\leq P\left(\sum_{i=K+1}^{\infty} |\langle v_i, X\rangle_{H_1}|^2 > \varepsilon \ \cap \ K > J_\varepsilon \right) + P(K \leq J_\varepsilon) \\ &= P(K \leq J_\varepsilon), \end{aligned}$$

where the last term converges to zero as $n \to \infty$. □

Lemma 4.11. *Let $L_n = \arg\max\{r \leq K : \sum_{i=1}^{r} (\hat{\lambda}_{K+1} - \hat{\lambda}_i)^{-2} \leq v_n\}$, where $K = K_n$ is given as in Theorem 2.2 and $v_n \to \infty$. Then $L_n \xrightarrow{P} \infty$.*

Proof. Let $r \in \mathbb{N}$ such that for all $1 \leq i \leq r$ we have $\lambda_{r+1} \neq \lambda_i$. Note that $E\|X\|_{H_1}^2 < \infty$ implies $\lambda_i \to 0$ and since $\lambda_i > 0$ we can find infinitely many r satisfying this condition. We choose such r and obtain

$$P(L_n < r) \leq P\left(\sum_{i=1}^{r} \frac{1}{(\hat{\lambda}_{K+1} - \hat{\lambda}_i)^2} > v_n \ \cap \ K \geq r \right) + P(K < r).$$

Lemma 4.8 implies that $P(K < r) \to 0$. The first term is bounded by $P\left(\sum_{i=1}^{r} \frac{1}{(\hat{\lambda}_{r+1} - \hat{\lambda}_i)^2} > v_n \right)$. Since $\hat{\lambda}_i \xrightarrow{P} \lambda_i$ and r is fixed while $v_n \to \infty$, it follows that $P(L_n < r) \to 0$ if $n \to \infty$. Since r can be chosen arbitrarily large, the proof is finished. □

Lemma 4.12. *Let Ψ and X be defined as in Lemma 2.2, then $\|P_{M_K}(X) - P_{\hat{M}_K}(X)\|_{H_1} \xrightarrow{P} 0$.*

Proof. Let us define two variables $X^{(1)} = \sum_{i=1}^{L} \langle X, v_i\rangle_{H_1} v_i$, $X^{(2)} = \sum_{i=L+1}^{\infty} \langle X, v_i\rangle_{H_1} v_i$ and L as in Lemma 4.11. Again for simplifying the notation we will write L instead of L_n. Since $X = X^{(1)} + X^{(2)}$ we derive

$$\|P_{M_K}(X) - P_{\hat{M}_K}(X)\|_{H_1} \leq \|P_{M_K}(X^{(1)}) - P_{\hat{M}_K}(X^{(1)})\|_{H_1} + \|P_{\hat{M}_K}(X^{(2)})\|_{H_1} + \|P_{M_K}(X^{(2)})\|_{H_1}. \tag{4.10}$$

The last two terms are bounded by $2\|X^{(2)}\|_{H_1}$. For the first summand in (4.10) we get

$$\|P_{M_K}(X^{(1)}) - P_{\hat{M}_K}(X^{(1)})\|_{H_1} = \left\| \sum_{i=1}^{L} \langle X, v_i\rangle_{H_1} (v_i - P_{\hat{M}_K}(v_i)) \right\|_{H_1}.$$

Let us choose $v_n = o(n)$ in Lemma 4.11. The triangle inequality, the Cauchy-Schwarz inequality, Lemma 4.9 and the definition of L entail

$$\begin{aligned}
\|P_{M_K}(X^{(1)}) - P_{\hat{M}_K}(X^{(1)})\|_{H_1} &\leq \sum_{i=1}^{L} |\langle X, v_i\rangle_{H_1}| \|v_i - P_{\hat{M}_K}(v_i)\|_{H_1} \\
&\leq \left(\sum_{i=1}^{L} |\langle X, v_i\rangle_{H_1}|^2\right)^{1/2} \left(\sum_{i=1}^{L} \|v_i - P_{\hat{M}_K}(v_i)\|_{H_1}^2\right)^{1/2} \\
&\leq \|X\|_{H_1} \left(\sum_{i=1}^{L} \|v_i - P_{\hat{M}_K}(v_i)\|_{H_1}^2\right)^{1/2} \\
&\leq 2\|X\|_{H_1} \|C - \hat{C}\|_{\mathcal{L}_{11}} \left(\sum_{i=1}^{L} \frac{1}{(\hat{\lambda}_{K+1} - \hat{\lambda}_i)^2}\right)^{1/2} \\
&\leq 2\|X\|_{H_1} \|C - \hat{C}\|_{\mathcal{L}_{11}} \sqrt{v_n}.
\end{aligned}$$

This implies the inequality

$$\|P_{M_K}(X) - P_{\hat{M}_K}(X)\|_{H_1} \leq 2\|X\|_{H_1} \|C - \hat{C}\|_{\mathcal{L}_{11}} \sqrt{v_n} + 2\|X^{(2)}\|_{H_1}. \tag{4.11}$$

Hence by Lemma 4.1 we have $2\|X\|_{H_1} \|C - \hat{C}\|_{\mathcal{L}_{11}} \sqrt{v_n} = o_P(1)$. Furthermore we have that $\|X^{(2)}\| = \left(\sum_{j>L} |\langle X, v_j\rangle|^2\right)^{1/2} \xrightarrow{P} 0$. This follows from the proof of Lemma 4.10. □

Lemma 4.13. *Let Ψ and X be defined as in Lemma 2.2, then $\|\Psi(P_{\hat{M}_K^{\perp}}(X))\|_{H_2} \xrightarrow{P} 0$.*

Proof. Some simple manipulations show

$$\begin{aligned}
\|\Psi(P_{\hat{M}_K^{\perp}}(X))\|_{H_2} &= \|\Psi(X - P_{\hat{M}_K}(X))\|_{H_2} \\
&= \|\Psi(P_{M_K}(X) + P_{M_K^{\perp}}(X) - P_{\hat{M}_K}(X))\|_{H_2} \\
&\leq \|\Psi(P_{M_K}(X)) - \Psi(P_{\hat{M}_K}(X))\|_{H_2} + \|\Psi(P_{M_K^{\perp}}(X))\|_{H_2} \\
&\leq \|\Psi\|_{\mathcal{L}_{12}} \left(\|P_{M_K}(X) - P_{\hat{M}_K}(X)\|_{H_1} + \|P_{M_K^{\perp}}(X)\|_{H_1}\right).
\end{aligned}$$

Direct applications of Lemma 4.10 and Lemma 4.12 finish the proof. □

Proof of Theorem 2.2. Set

$$\Theta_n(x) = \sum_{j=1}^{K_n} \frac{\hat{\Lambda}(\hat{v}_j)}{\hat{\lambda}_j} \langle \hat{v}_j, x\rangle_{H_1}.$$

By the representation (4.9) and the triangle inequality

$$\|\Psi(X) - \hat{\Psi}(X)\|_{H_2} \leq \|\Theta_n(X)\|_{H_2} + \|\Psi(P_{\hat{M}_{K_n}^{\perp}}(X))\|_{H_2}.$$

Lemma 4.13 shows that the second term tends to zero in probability.

If in Lemma 4.1 we define $\Psi \equiv 0$, then $\hat{\Lambda} = \hat{\Delta}$ and by independence of ε_k and X_k we get $\Lambda = 0$. By the arguments of Lemma 4.5 we infer $P(\|\Theta_n\|_{\mathcal{L}_{12}} > \varepsilon) \leq U\frac{m_n^2}{\varepsilon^2 n}$, which implies that $\|\Theta_n(X)\|_{H_2} \xrightarrow{P} 0$. □

A Appendix

n	K_n^{OPT}	NMSE	K_n^1	NMSE	$K_n^{0.5}$	NMSE	$K_n^{0.1}$	NMSE
10	1	3.26	2	3.50	1	3.26	1	3.26
20	1	1.38	3	2.59	2	1.88	1	1.38
40	1	1.14	4	1.73	3	1.29	1	1.14
80	3	0.77	6	1.58	4	1.05	1	0.80
160	5	0.48	6	0.62	5	0.48	1	0.73
320	7	0.31	7	0.31	6	0.36	2	0.57
640	8	0.18	9	0.19	7	0.19	3	0.33
1280	11	0.11	9	0.11	8	0.12	4	0.25
2560	11	0.06	10	0.07	9	0.08	5	0.17
5120	15	0.03	11	0.04	9	0.05	6	0.10
10240	17	0.02	12	0.02	10	0.03	6	0.10
20480	17	0.01	13	0.01	11	0.02	7	0.06

Table 1: Ψ_1, Λ_1, $\sigma = 1$

n	K_n^{OPT}	NMSE	K_n^1	NMSE	$K_n^{0.5}$	NMSE	$K_n^{0.1}$	NMSE
20	4	0.96	12	1.85	9	1.46	1	1.01
80	12	0.68	20	0.93	16	0.79	3	0.85
320	20	0.25	27	0.30	23	0.27	9	0.50
1280	29	0.07	30	0.08	27	0.08	17	0.19
5120	35	0.02	31	0.02	30	0.02	23	0.06
20480	34	0	33	0.01	31	0.01	26	0.02

Table 2: Ψ_1, Λ_2, $\sigma = 1$

n	K_n^{OPT}	NMSE	K_n^1	NMSE	$K_n^{0.5}$	NMSE	$K_n^{0.1}$	NMSE
20	7	0.90	18	1.39	16	1.03	5	0.93
80	28	0.58	35	0.88	32	0.62	14	0.71
320	35	0.12	35	0.12	35	0.12	33	0.16
1280	35	0.03	35	0.03	35	0.03	35	0.03
5120	35	0.01	35	0.01	35	0.01	35	0.01
20480	35	0	35	0	35	0	35	0

Table 3: Ψ_1, Λ_3, $\sigma = 1$

n	K_n^{OPT}	NMSE	K_n^1	NMSE	$K_n^{0.5}$	NMSE	$K_n^{0.1}$	NMSE
20	1	2	3	3.92	2	2.68	1	2
80	1	0.90	6	1.82	4	1.18	1	0.90
320	4	0.45	8	0.70	6	0.49	2	0.57
1280	7	0.16	9	0.17	8	0.18	4	0.27
5120	11	0.05	11	0.05	9	0.07	5	0.17
20480	17	0.02	13	0.02	11	0.02	7	0.06

Table 4: Ψ_2, Λ_1, $\sigma = 1$

n	K_n^{OPT}	NMSE	K_n^1	NMSE	$K_n^{0.5}$	NMSE	$K_n^{0.1}$	NMSE
20	1	1.08	11	2.54	9	1.97	1	1.08
80	7	0.86	21	1.68	17	1.34	3	0.92
320	20	0.40	27	0.51	24	0.43	9	0.54
1280	28	0.12	30	0.14	27	0.13	17	0.23
5120	32	0.03	31	0.03	30	0.04	22	0.08
20480	34	0.01	33	0.01	31	0.01	26	0.02

Table 5: Ψ_2, Λ_2, $\sigma = 1$

n	K_n^{OPT}	NMSE	K_n^1	NMSE	$K_n^{0.5}$	NMSE	$K_n^{0.1}$	NMSE
20	6	0.97	18	2.16	16	1.65	5	1.01
80	22	0.75	35	1.49	32	1.12	14	0.83
320	35	0.22	35	0.22	35	0.22	33	0.25
1280	35	0.05	35	0.05	35	0.05	35	0.05
5120	35	0.01	35	0.01	35	0.01	35	0.01
20480	35	0	35	0	35	0	35	0

Table 6: Ψ_2, Λ_3, $\sigma = 1$

n	K_n^{OPT}	NMSE	K_n^1	NMSE	$K_n^{0.5}$	NMSE	$K_n^{0.1}$	NMSE
20	1	35.28	3	77.60	2	68.64	1	35.28
80	1	16.90	6	71.07	4	39.40	1	16.90
320	1	1.93	7	15.85	6	13.28	2	2.87
1280	1	1.09	9	4.67	8	3.46	4	2.06
5120	3	0.64	11	1.22	9	1.18	6	0.83
20480	6	0.24	13	0.42	11	0.33	7	0.26

Table 7: Ψ_3, Λ_1, $\sigma = 1$

n	K_n^{OPT}	NMSE	K_n^1	NMSE	$K_n^{0.5}$	NMSE	$K_n^{0.1}$	NMSE
20	1	2.39	12	61.32	9	40.39	1	2.39
80	1	1.50	20	34.95	16	27.85	3	3.87
320	1	1.15	27	10.96	23	9.81	9	4.35
1280	3	1.04	30	3.20	27	2.66	17	2.01
5120	17	0.53	31	0.74	30	0.73	23	0.64
20480	29	0.18	33	0.21	31	0.20	26	0.19

Table 8: Ψ_3, Λ_2, $\sigma = 1$

n	K_n^{OPT}	NMSE	K_n^1	NMSE	$K_n^{0.5}$	NMSE	$K_n^{0.1}$	NMSE
20	1	1.33	18	47.52	16	27.09	5	4.04
80	1	1.20	35	36.28	32	26.53	14	5.84
320	1	1.05	35	5.24	35	5.24	33	4.50
1280	8	0.96	35	1.22	35	1.22	35	1.22
5120	34	0.28	35	0.28	35	0.28	35	0.28
20480	35	0.08	35	0.08	35	0.08	35	0.08

Table 9: Ψ_3, Λ_3, $\sigma = 1$

References

[1] D. Bosq, *Modelization, nonparametric estimation and prediction for continuous time processes*, Nonparametric functional estimation and related topics (Spetses, 1990), NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci., vol. 335, Kluwer Acad. Publ., 1991, pp. 509–529.

[2] ———, *Linear Processes in Function Spaces.*, Springer, New York, 2000.

[3] T. T. Cai and P. Hall, *Prediction in functional linear regression*, Ann. Statist. **34** (2006), 2159–2179.

[4] H. Cardot, F. Ferraty, and P. Sarda, *Functional linear model*, Statist. Probab. Lett. **45** (1999), 11–22.

[5] ———, *Spline estimators for the functional linear model*, Statist. Sinica **13** (2003), 571–591.

[6] J.-M. Chiou, H.-G. Müller, and J.-L. Wang, *Functional response models*, Statist. Sinica **14** (2004), 675–693.

[7] C. Crambes and A. Mas, *Optimal prediction and dimension selection in linear models with functional output*, Technical Report.

[8] C. Crambes, A. Kneip, and P. Sarda, *Smoothing splines estimators for functional linear regression*, Ann. Statist. **37** (2009), 35–72.

[9] A. Cuevas, M. Febrero, and R. Fraiman, *Linear functional regression: the case of fixed design and functional response*, Canad. J. Statist. **30** (2002), 285–300.

[10] K. P. Didericksen, D. and X. Zhang, *Empirical properties of forecasts with the functional autoregressive model*, Computational Statistics, forthcoming.

[11] P. Eilers and B. Marx, *Flexible Smoothing with B-splines and Penalties*, Statistical Sciences **11** (1996), 89–121.

[12] M. Febrero-Bande, P. Galeano, and G.-M. W., *Measures of influence for the functional linear model with scalar response*, Journal of Multivariate Analysis **101** (2010), 327–339.

[13] F. Ferraty, A. Laksaci, A. Tadj, and P. Vieu, *Kernel regression with functional response*, Electronic Journal of Statistics **5** (2011), 159–171.

[14] T. Hastie and C. Mallows, *A discussion of "a statistical view of some chemometrics regression tools" by I.E. Frank and J.H. Friedman*, Technometrics **35** (1993), 140–143.

[15] S. Hörmann, L. Horváth, and R. Reeder, *A Functional Version of the ARCH Model*, Econometric Theory, forthcoming.

[16] S. Hörmann and P. Kokoszka, *Functional Time Series*, Handbook of Statistics, forthcoming.

[17] ———, *Weakly dependent functional data*, The Annals of Statistics **38** (2010), 1845–1884.

[18] L. Horváth and P. Kokoszka, *Inference for Functional Data with Applications*, Springer, 2012.

[19] V. Kargin and A. Onatski, *Curve forecasting by functional autoregression*, J. Multivariate Anal. **99** (2008), 2508–2526.

[20] Y. Li and T. Hsing, *On rates of convergence in functional linear regression*, J. Multivariate Anal. **98** (2007), 1782–1804.

[21] N. Malfait and J. O. Ramsay, *The historical functional linear model*, Canad. J. Statist. **31** (2003), 115–128.

[22] H.-G. Müller and U. Stadtmüller, *Generalized functional linear models*, Ann. Statist. **33** (2005), 774–805.

[23] J. Ramsay and B. Silverman, *Functional Data Analysis (2nd ed.)*, Springer, New York, 2005.

[24] P. T. Reiss and R. T. Ogden, *Functional principal component regression and functional partial least squares*, J. Amer. Statist. Assoc. **102** (2007), no. 479, 984–996.

[25] F. Yao, H.-G. Müller, and J.-L. Wang, *Functional linear regression analysis for longitudinal data*, Ann. Statist. **33** (2005), 2873–2903.