# Estimation in functional lagged regression

Siegfried Hörmann[1*] Łukasz Kidziński[2]
Piotr Kokoszka[3]

[1] Department of Mathematics, Université libre de Bruxelles CP210, Bd. du Triomphe, B-1050 Brussels, Belgium.
[2] Computer-Human Interaction Lab for Learning & Instruction, École Polytechnique Fédérale de Lausanne, RLC D1 740, Station 20, CH-1015 Lausanne, Switzerland.
[3] Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877, USA.

**Abstract.** The paper introduces a *functional* time series (lagged) regression model. The impulse response coefficients in such a model are operators acting on a separable Hilbert space, which is the function space $L^2$ in applications. A spectral approach to the estimation of these coefficients is proposed and asymptotically justified under a general nonparametric condition on the temporal dependence of the input series. Since the data are infinite dimensional, the estimation involves a spectral domain dimension reduction technique. Consistency of the estimators is established under general data dependent assumptions on the rate of the dimension reduction parameter. Their finite sample performance is evaluated by a simulation study which compares two ad hoc approaches to dimension reduction with a new asymptotically justified method. The new method is superior when the MSE of the in sample prediction error is used as a criterion.

## 1. Introduction

This paper is concerned with the estimation of impulse response operators in functional lagged regression. Time series (or lagged) regression goes back to the origins of modern time series analysis, Kolmogorov (1941), Wiener (1949). Accounts are given in many monographs and textbooks, e.g. Brillinger (1975), Priestley (1981), Shumway and Stoffer (2011). It forms the most important input–output paradigm in modeling engineering, geophysical and economic systems. In an abstract form,

*Corresponding author. Email: shormann@ulb.ac.be
[0]This research was supported by the Communauté française de Belgique—Actions de Recherche Concertées (2010–2015) and NSA grant 14-1-0165.

the lagged regression model is

$$Y_\ell = a + \sum_{k\in\mathbb{Z}} b_k(X_{\ell-k}) + \varepsilon_\ell. \tag{1.1}$$

The regressors $X_k$ are elements of a Hilbert space $H$, the responses $Y_\ell$ and the errors $\varepsilon_\ell$ belong to a possibly different Hilbert space $H'$, and $b_k : H \to H'$ are linear operators. In the most common setting, all quantities are scalars, applications with several scalar input series are not uncommon. While model (1.1) can be formulated in abstract Hilbert spaces, the existing statistical theory relies on the assumption that all spaces are finite dimensional. This is because solutions to problems of estimation, prediction and interpolation require inverting various matrices, and these inverses do not exist (as bounded operators) in infinite dimensional spaces. A dimension reduction methodology with a requisite theory must be developed.

Such issues have been extensively investigated in the field of Functional Data Analysis, with the most relevant research relating to the functional linear model, e.g. Ramsay and Silverman (2005), Horváth and Kokoszka (2012). There are many types of functional linear models; those most relevant to this paper are known as the *scalar response model* and the *fully functional model*. The scalar response model has the form $Y_k = a + \int_{\mathcal{T}} b(u)X_k(u)du + \varepsilon_k$, where the $X_k$ are functions and the responses $Y_k$ are scalars. This model has been investigated from many angles, to give a selection of references, we cite Cardot et al. (2003), Müller and Stadtmüller (2005), Cai and Hall (2006), Li and Hsing (2007), Crambes et al. (2009), James et al. (2009), McKeague and Sen (2010) and Comte and Johannes (2012). The fully functional model is defined by

$$Y_k(t) = a(t) + \int_{\mathcal{T}} b(t,u)X_k(u)du + \varepsilon_k(t),$$

where now the responses $Y_k$ and the errors $\varepsilon_k$ are also functions. This model is more complex, and has not been so thoroughly investigated as the scalar response model, but it is safe to say that it is presently well understood: Yao et al. (2005), Chiou and Müller (2007), Hörmann and Kokoszka (2010) Gabrys et al. (2010) and Hörmann and Kidziński (2014) are just a few examples of recent work.

As in the usual linear regression, the assumption imposed on the above models is that the pairs $(X_k, Y_k)$ are independent and identically distributed, and these models do not involve lagged values of the input series. However, many problems in science, economics and engineering can be formulated in terms of statistical inference for functional time series (FTS) which are defined as $\{X_n(t), t \in \mathcal{T},\ n = 1, 2, \ldots\}$, where $n$ is the index referring to day, week, year or a similar unit of time, and plays the role of the time index in time series analysis. The random elements $X_n$ are functions defined on a common domain $\mathcal{T}$, typically an interval. This concept has been applied over the last two decades in many settings, and a fairly complete theory for estimation, prediction and testing for a single FTS exists, both in time and spectral domains: Bosq (2000), Horváth and Kokoszka (2012), Hörmann and Kokoszka (2012), to name a few accounts.

The objective of this paper is to advance the existing framework by considering the input–output paradigm for two FTS in the context of model (1.1). There are two broad approaches to inference and prediction in the lagged regression model: 1) time domain approach based on ARMA modeling of the series $(X_\ell)$ and response function modeling of the coefficients $b_k$, Box et al. (1994); 2) spectral domain approach based on coherency analysis, Brillinger (1975). While the Box–Jenkins approach has an appealing heuristic justification, the coherency approach is viewed as a more principled one, and has been extensively used in geosciences and engineering. Recent advances in the spectral theory for functional data, Panaretos and Tavakoli (2013b, 2013a), Hörmann et al. (2014), have opened up a prospect of developing a usable and asymptotically supported methodology for model (1.1). As with most functional procedures, the main challenge is a suitable dimension reduction technique and the need to deal with unbounded operators, difficulties not encountered in the scalar and vector theory; details are explained in Section 3.

The remainder of this paper is organized as follows. Section 2 introduces model (1.1) in greater detail by specifying the assumptions on its parameters and dependence structure. Estimation methodology is explained in Section 3 and asymptotically justified in Section 4. Its finite sample performance is evaluated in Section 5. All proofs are collected in Section 6. In the Appendix, we describe a new method for selecting an important dimension reduction parameter.

# 2. Model specification

We consider model (1.1) with a strictly stationary sequence $(X_k)$ and a thereof independent i.i.d. sequence $(\varepsilon_k)$, with realizations in separable Hilbert spaces $H$ and $H'$, respectively. These spaces are equipped with the norms $\|f\| = \sqrt{\langle f, f\rangle}$, where $\langle\cdot,\cdot\rangle$ is the inner product. The inner product in $H$ and $H'$ is denoted in the same way. Even though we consider only real–valued observations, it is convenient to assume that $H$ and $H'$ are Hilbert spaces over the complex field $\mathbb{C}$, so that $\langle f, g\rangle = \overline{\langle g, f\rangle}$, where $\overline{z}$ is the complex conjugate of $z$.

Throughout we suppose that $E\|X_k\|^2 < \infty$ and $E\|\varepsilon_k\|^2 < \infty$ and that $E\varepsilon_k = 0$. A sufficient condition for the convergence of (1.1) is $\sum_{k\in\mathbb{Z}} \|b_k\|_{\mathcal{L}} < \infty$, where $\|b\|_{\mathcal{L}} = \sup_{f:\,\|f\|=1} \|b(f)\|$ denotes the usual operator norm. A slightly stronger, but more convenient assumption is

$$\sum_{k\in\mathbb{Z}} \|b_k\|_{\mathcal{S}} < \infty, \tag{2.1}$$

where $\|\cdot\|_{\mathcal{S}}$ is the Hilbert–Schmidt norm. Recall that $\|\Psi\|_{\mathcal{S}}^2 = \sum_{j\geq 1} \|\Psi(e_j)\|^2 = \sum_{j\geq 1} \lambda_j^2$, where $(e_j)$ is any orthonormal basis in $H$ and the $\lambda_j$ are the eigenvalues of $\Psi$. Recall that $\|\Psi\|_{\mathcal{L}} \leq \|\Psi\|_{\mathcal{S}}$. Our assumptions imply that $(Y_\ell)$ is also strictly stationary and $E\|Y_\ell\|^2 < \infty$.

The means $\mu_X = EX_\ell$ and $\mu_Y = EY_\ell$ are estimated by sample averages which, under quite general dependence assumptions, are $\sqrt{n}$-consistent, see Section 2.4 of Bosq (2000) for general results in Banach spaces, and Theorem 16.3. of Horváth and Kokoszka (2012) for the form of dependence used in this paper. Since $\mu_Y =$

$a + \sum_{k\in\mathbb{Z}} b_k(\mu_X)$, once the $b_k$ have been estimated, an estimator for the intercept operator $a$ can be easily obtained. We therefore consider from now on the model

$$Y_\ell = \sum_{k\in\mathbb{Z}} b_k(X_{\ell-k}) + \varepsilon_\ell, \quad (EY_\ell = 0,\ EX_\ell = 0). \tag{2.2}$$

Since the process $(Y_\ell, X_\ell)$ is strictly stationary and has second order moments, the operators

$$C_h^X := \mathrm{Cov}(X_h, X_0) \quad \text{and} \quad C_h^{YX} := \mathrm{Cov}(Y_h, X_0)$$

defined by the relation

$$\mathrm{Cov}(X, Y)(f) = E[(X - EX)\langle f, Y - EY\rangle]$$

exist as elements of the space of Hilbert–Schmidt operators. The autocovariances of the input series are assumed to be summable:

$$\sum_{h\in\mathbb{Z}} \|C_h^X\|_{\mathcal{S}} < \infty. \tag{2.3}$$

For ease of reference, we collect the time domain conditions imposed so far in the following assumption.

**Assumption 2.1.** All random elements are square integrable. Model (2.2) and conditions (2.1) and (2.3) hold. The sequences $(X_k)$ and $(\varepsilon_k)$ are strictly stationary and independent of each other. The errors $\varepsilon_k$ are independent.

Setting $\mathrm{i} = \sqrt{-1}$, we introduce the *spectral density operator*

$$\mathcal{F}_\theta^X := \frac{1}{2\pi} \sum_{h\in\mathbb{Z}} C_h^X e^{-\mathrm{i}h\theta}, \quad \theta \in [-\pi, \pi],$$

and the *cross–spectral density operator*

$$\mathcal{F}_\theta^{YX} := \frac{1}{2\pi} \sum_{h\in\mathbb{Z}} C_h^{YX} e^{-\mathrm{i}h\theta}, \quad \theta \in [-\pi, \pi].$$

By (2.3) and Lemma 6.2 these two series are absolutely convergent in the Hilbert–Schmidt norm.

We will use the following assumption.

**Assumption 2.2.** For any $\theta \in [-\pi, \pi]$ the operators $\mathcal{F}_\theta^X : H \to H$ are full rank, that is $\ker\left(\mathcal{F}_\theta^X\right) = 0$.

For a scalar time series, Assumption 2.2 is equivalent to requiring that the spectral density of $(X_\ell)$ be positive over $[-\pi, \pi]$. An analogous nonsingularity condition must be imposed for vector–valued time series, Theorem 8.3.1 of Brillinger (1975).

Next, we introduce the *frequency response operator*

$$B_\theta := \sum_{h\in\mathbb{Z}} b_h e^{-\mathrm{i}h\theta}, \quad \theta \in [-\pi, \pi].$$

The mapping $\theta \mapsto \frac{1}{2\pi} B_\theta$ is the Fourier transform (FT) of the sequence $(b_k)$ from which we may recover $b_h$ by

$$b_h = \frac{1}{2\pi} \int_{-\pi}^{\pi} B_\theta e^{\mathrm{i}h\theta} d\theta.$$

In case of the general model (1.1), $b_h$ are operators. We refer to Hörmann et al. (2014) for details on how this type of FT is rigorously defined. If the $Y_k$ are scalars and the $b_k$ functions on an interval, then $B_\theta = B_\theta(u) = \sum_{h\in\mathbb{Z}} b_h(u) e^{-\mathrm{i}h\theta}$ reduces to a pointwise FT.

We conclude this section by specifying the assumptions on the dependence structure of the process $(X_k)$. We use the concept of $L^p$-$m$-approximability introduced by Hörmann and Kokoszka (2010). This moment based notion of dependence is convenient to apply and has been verified to hold for several popular functional time series models, including functional linear processes. We conjecture that our results could also be established under the cumulant type assumptions used by Panaretos and Tavakoli (2013a), but the latter framework seems to be more restrictive than ours. We write that $X \in L^p_H$ if $X$ takes values in Hilbert space $H$ and

$$\nu_p(X) := \big(E\|X\|^p\big)^{1/p} < \infty.$$

**Definition 2.1.** A sequence $(X_n) \in L^p_H$ is called *$L^p$-$m$-approximable* if each $X_n$ admits the representation

$$X_n = f(u_n, u_{n-1}, \ldots), \tag{2.4}$$

where the $u_i$ are i.i.d. elements taking values in a measurable space $S$, and $f$ is a measurable function $f : S^\infty \to H$. Moreover we assume that if $\{u_i'\}$ is an independent copy of $\{u_i\}$ defined on the same probability space, then letting

$$X_n^{(m)} = f(u_n, u_{n-1}, \ldots, u_{n-m+1}, u'_{n-m}, u'_{n-m-1}, \ldots) \tag{2.5}$$

we have

$$\sum_{m=1}^{\infty} \nu_p\big(X_0 - X_0^{(m)}\big) < \infty. \tag{2.6}$$

Notice that by construction $X_n^{(m)} \overset{\mathcal{L}}{=} X_0$ (equality in law), and that $X_n^{(m)}$ is independent of $(X_{n-k} \colon k \geq m)$. Representation (2.4) implies that the $X_k$ form a stationary and ergodic sequence in $L^2$. Similar assumptions have been used extensively in recent theoretical work, as all stationary time series models in practical use can be represented as Bernoulli shifts (2.4), see Wu (2005), Shao and Wu (2007), Aue et al. (2009), Hörmann and Kokoszka (2010), Hörmann et al. (2013) and Kokoszka and Reimherr (2013).

**Assumption 2.3.** The input sequence $(X_k)$ is $L^4$-$m$-approximable.

# 3. Estimation of the impulse response operators

In a scalar lagged regression model $y_\ell = \sum_k b_k x_{\ell-k} + \varepsilon_k$, the frequency response function is estimated by $\widehat{B}_\theta = \hat{f}_{yx}(\theta)/\hat{f}_{xx}(\theta)$, where $\hat{f}_{yx}(\theta)$ and $\hat{f}_{xx}(\theta)$ are, respectively, estimates of the cross–spectrum of $(y_k)$ and $(x_k)$ and the spectral density of the $(x_k)$. The response coefficients $b_h$ are then estimated by the inverse FT of $\widehat{B}_\theta$. To develop a similar procedure for functional data, we begin with the relation

$$\mathcal{F}_\theta^{YX} = B_\theta \circ \mathcal{F}_\theta^X \tag{3.1}$$

which follows by changing the order of summation (cf. Lemma 6.2):

$$\mathcal{F}_\theta^{YX}(f) = \sum_{k\in\mathbb{Z}} b_k \left( \frac{1}{2\pi} \sum_{h\in\mathbb{Z}} C_{h-k}^X(f) e^{-\mathrm{i}(h-k)\theta} \right) e^{-\mathrm{i}k\theta} = B_\theta \circ \mathcal{F}_\theta^X(f).$$

Heuristically, (3.1) yields the relation $B_\theta = \mathcal{F}_\theta^{YX} \circ \left(\mathcal{F}_\theta^X\right)^{-1}$. This relation is heuristic only because $\left(\mathcal{F}_\theta^X\right)^{-1}$ is not a bounded operator, see Lemma 6.3. We now explain how to overcome this problem and construct consistent estimators of the impulse response operators $b_k$.

For any $\theta \in [-\pi, \pi]$, the operator $\mathcal{F}_\theta^X$ is Hilbert-Schmidt, symmetric and non-negative definite. The verification is not difficult, see Hörmann et al. (2014). Assumption 2.2 implies that its eigenvalues $\lambda_m(\theta)$ are positive and the eigenfunctions $(\varphi_m(\theta)\colon m \geq 1)$ form an orthonormal basis of $H$. Hence, by the spectral theorem, it can be decomposed as

$$\mathcal{F}_\theta^X(f) = \sum_{m\geq 1} \lambda_m(\theta) \langle f, \varphi_m(\theta)\rangle \varphi_m(\theta), \qquad f = \sum_{m\geq 1} \langle f, \varphi_m(\theta)\rangle, \varphi_m(\theta), \tag{3.2}$$

where the $\lambda_m(\theta)$ are arranged in descending order and the corresponding eigenfunctions are normalized to unit length. Relations (3.1) and (3.2) imply

$$\mathcal{F}_\theta^{YX}(\varphi_m(\theta)) = \lambda_m(\theta) B_\theta\left(\varphi_m(\theta)\right), \quad m \geq 1.$$

Therefore

$$B_\theta(f) = \sum_{m\geq 1} \frac{\langle f, \varphi_m(\theta)\rangle}{\lambda_m(\theta)} \mathcal{F}_\theta^{YX}(\varphi_m(\theta)). \tag{3.3}$$

The latter sum is convergent for any $f \in H$, and relation (3.3) forms the starting point of our estimations approach.

Consider a sample $(Y_1, X_1), \ldots, (Y_n, X_n)$ and the sample cross–covariance operators:

$$\widehat{C}_h^{YX}(f) = \begin{cases} \frac{1}{n}\sum_{k=1}^{n-h} Y_{k+h}\langle f, X_k\rangle, & h = 0, \ldots, n-1; \\ \frac{1}{n}\sum_{k=1-h}^{n} Y_{k+h}\langle f, X_k\rangle, & h = -n+1, \ldots, -1; \\ 0, & |h| \geq n. \end{cases}$$

The estimators $\widehat{C}_h^X$ for the autocovariances of $(X_\ell)$ are defined analogously. Now we set

$$\widehat{\mathcal{F}}_{\theta|q}^{YX} = \frac{1}{2\pi} \sum_{|h|\leq q} \omega_q(h) \widehat{C}_h^{YX} e^{-\mathrm{i}h\theta}.$$

This is the functional version of the smoothed periodogram. A popular choice is to use the Bartlett weights $\omega_q(h) = 1 - |h|/(q+1)$, but any weights satisfying the following assumption can be used.

**Assumption 3.1.** The weights $\omega_q(h)$ satisfy $\omega_q(-h) = \omega_q(h)$, $|\omega_q(h)| \leq \omega^\star$, for some $\omega^\star$ independent of $q$ and $h$, and $\lim_{q\to\infty} \omega_q(h) = 1$, for every fixed $h$.

All kernels used in practice lead to weights which satisfy Assumption 3.1.

In an analogous way define $\widehat{\mathcal{F}}^X_{\theta|q}$. The latter operator is non-negative definite, symmetric and Hilbert-Schmidt for any frequency $\theta \in [-\pi, \pi]$. Thus, the spectral theorem applies and we can use its eigenfunctions $\hat{\varphi}_m(\theta) = \hat{\varphi}_{m|q}(\theta)$ and eigenvalues $\hat{\lambda}_m(\theta) = \hat{\lambda}_{m|q}(\theta)$ as estimators for the population spectrum of $\mathcal{F}^X_\theta$. Clearly, $\varphi_m(\theta)$ and $\hat{\varphi}_m(\theta)$ can only be close if they have the same direction. In other words, the best we can hope is that $\|\varphi_m(\theta) - \hat{c}_m(\theta)\hat{\varphi}_m(\theta)\|$ is small, when $\hat{c}_m(\theta) = \langle \varphi_m(\theta), \hat{\varphi}_m(\theta)\rangle / |\langle \varphi_m(\theta), \hat{\varphi}_m(\theta)\rangle|$. This implies that all formulas defining estimators and test statistics must be invariant with respect to $\hat{c}_m(\theta)$.

We are now ready to define

$$\hat{b}_h = \frac{1}{2\pi} \int_{-\pi}^{\pi} \widehat{B}_\theta e^{\mathrm{i}h\theta} d\theta,$$

where

$$\widehat{B}_\theta(f) = \widehat{B}_{\theta|p,q,K}(f) = \sum_{m=1}^{K} \frac{\langle f, \hat{\varphi}_{m|q}(\theta)\rangle}{\hat{\lambda}_{m|q}(\theta)} \widehat{\mathcal{F}}^{YX}_{\theta|p}(\hat{\varphi}_{m|q}(\theta)). \tag{3.4}$$

The estimator $\widehat{B}_\theta$ involves three tuning parameters: $p$, $q$ and $K$, which in principle may each depend on $\theta$. For the sake of readability, we shall in the sequel often suppress the dependence on these parameters in the notation. The selection of these parameters is discussed in the following sections.

Notice that (3.4) is invariant with respect to rotations of $\hat{\varphi}_m(\theta)$; if $c$ is a number on the complex unit circle, then we can replace $\hat{\varphi}_m(\theta)$ by $c\hat{\varphi}_m(\theta)$ without changing the estimator. This follows from

$$\langle f, c\hat{\varphi}_m(\theta)\rangle \widehat{\mathcal{F}}^{YX}_\theta(c\hat{\varphi}_m(\theta)) = \overline{c}c\langle f, \hat{\varphi}_m(\theta)\rangle \widehat{\mathcal{F}}^{YX}_\theta(\hat{\varphi}_m(\theta)),$$

and $\overline{c}c = 1$. Hence, in theoretical arguments, we can replace $\hat{\varphi}_m(\theta)$ in (3.4) by $\hat{c}_m(\theta)\hat{\varphi}_m(\theta)$.

# 4. Consistency of the estimators

Asymptotic assumptions commonly used in the context of functional regression models are formulated in terms of decay rates of eigenvalues and conditions on the gaps between eigenvalues of the covariance operator $C^X_0$. Under such assumptions, convergence rates for the estimators can be obtained. In our spectral context, taking this route would necessitate translating such conditions about the eigenvalues of $C^X_0$ to the eigenvalues of $\mathcal{F}^X_\theta$, $\theta \in [-\pi, \pi]$, and this does not yield clean conditions. It is

much more natural to base the rates of convergence of the $\hat{b}_h$ directly on the rates of convergence of the spectral density operator stated in the following lemma which is proven in Section 6.

**Lemma 4.1.** *Suppose Assumptions 2.1, 2.2, 2.3 and 3.1 hold. If $q = q_n \to \infty$, such that $q^2 = o(n)$, then there exist null sequences $(\psi_n^X)$ and $(\psi_n^{YX})$, such that*

$$\sup_{\theta\in[-\pi,\pi]} \|\mathcal{F}_\theta^X - \widehat{\mathcal{F}}_\theta^X\|_{\mathcal{L}} = o_P(\psi_n^X) \quad \text{and} \quad \sup_{\theta\in[-\pi,\pi]} \|\mathcal{F}_\theta^{YX} - \widehat{\mathcal{F}}_\theta^{YX}\|_{\mathcal{L}} = o_P(\psi_n^{YX}).$$

Such an approach will allow us to specify the value of $K$ (the truncation level in (3.4)) which implies the consistency of the $\hat{b}_h$ directly in terms of the sequences $(\psi_n^X)$ and $(\psi_n^{YX})$.

To establish the consistency of the $\hat{b}_h$, we need technical assumptions which ensure the identifiability of the eigenfunctions of the spectral density estimators. These assumptions do not enter into the convergence rates or the selection of $K$. Introduce the following function, which measures the size of spectral gaps:

$$\begin{aligned}\alpha_1(\theta) &:= \lambda_1(\theta) - \lambda_2(\theta);\\ \alpha_m(\theta) &:= \{\lambda_m(\theta) - \lambda_{m+1}(\theta)\} \wedge \{\lambda_{m-1}(\theta) - \lambda_m(\theta)\}, \quad m \geq 2.\end{aligned}$$

In case of $\alpha_m(\theta) \neq 0$, the eigenspace corresponding to $\lambda_m(\theta)$ is one-dimensional, and $\varphi_m(\theta)$ is unique up to multiplication with a number on the complex unit circle. If $\alpha_m(\theta) = 0$ for some $\theta$, the eigenspace corresponding to $\lambda_m(\theta)$ has dimension greater than one. Then, $\varphi_m(\theta)$ cannot be identified. We shall thus impose the following assumptions.

**Assumption 4.1.** It holds that $\inf_\theta \alpha_k(\theta) > 0$, for all $k \geq 1$.

To formulate our consistency result, we need the following random variables. We define $K = \min\{K^{(i)}, 1 \leq i \leq 4\}$ with

$$\begin{aligned}K^{(1)} &= \max\{k \geq 1 \colon \inf_\theta \hat{\lambda}_k(\theta) \geq 2\psi_n^X\},\\ K^{(2)} &= \max\{k \geq 1 \colon \psi_n^{YX} \int_{-\pi}^{\pi} W_\lambda^k(\theta) d\theta \leq 1\},\\ K^{(3)} &= \max\{k \geq 1 \colon \int_{-\pi}^{\pi} \big(W_\lambda^k(\theta)\big)^2 d\theta \leq \big(\psi_n^X\big)^{-1/2}\},\\ K^{(4)} &= \max\{k \geq 1 \colon \int_{-\pi}^{\pi} \big(W_\alpha^k(\theta)\big)^2 d\theta \leq \big(\psi_n^X\big)^{-1/2}\},\end{aligned}$$

and

$$W_\lambda^k(\theta) = \left(\sum_{m=1}^k \frac{1}{\hat{\lambda}_m^2(\theta)}\right)^{1/2} \quad \text{and} \quad W_\alpha^k(\theta) = \left(\sum_{m=1}^k \frac{1}{\hat{\alpha}_m^2(\theta)}\right)^{1/2}.$$

By convention, the maximum over the empty set is equal to zero.

**Theorem 4.1.** *Suppose that Assumptions 2.1, 2.2, 2.3, 3.1, 4.1 hold. For any null sequences* $(\psi_n^X)$ *and* $(\psi_n^{YX})$ *in Lemma 4.1 define* $K = \min\{K^{(i)} \colon 1 \le i \le 4\}$. *If* $q, p \to \infty$ *such that* $q + p = o(n^{1/2})$, *then*

$$\max_{h \in \mathbb{Z}} \|\hat{b}_h - b_h\|_{\mathcal{S}} \xrightarrow{\mathcal{P}} 0.$$

Theorem 4.1 is proven in Section 6

Theorem 4.1 provides general conditions on the dimension parameter $K$ to ensure that the estimator $\hat{b}_h$ is consistent. We now propose three specific rules whose finite sample performance will be compared in the next section.

**Cross–validation (CV).** We divide $\{1, \dots, n\}$ into a training set $S_{\text{tr}} = \{1, \dots, m\}$ and a test set $S_{\text{test}} = \{m+1, \dots, n\}$, with $m = \lfloor \alpha n \rfloor$ and $\alpha \in (0, 1)$. With the variables $\{X_j \colon j \in S_{\text{tr}}\}$ we estimate the operators $\hat{b}_{h|k}$ for $h \in \{-H, \dots, H\}$, $H \ge 0$, with a fixed dimension $K_\theta = k$ for all $\theta$. Then, we compute

$$V_k^2 := \sum_{j \in S_{\text{test}}} \Big\| Y_j - \sum_{|h| \le H} \hat{b}_{h|k} (X_{j-h} I\{j - h \in S_{\text{test}}\}) \Big\|^2, \quad k \ge 1.$$

We set $K = \text{argmin}_{k \ge 1}\{V_k\}$. In Section 5, we use $\alpha = 0.8$ and $H = 3$.

A potential disadvantage of this method is that $K$ is fixed for all frequencies $\theta$. In principle, one could vary $K$ over a partition of $[-\pi, \pi]$. However, such a method is numerically unstable and increases computational costs.

**Eigenvalue thresholding (ET).** A major source of variability of estimator (3.4) is small eigenvalues $\hat{\lambda}_{m|q}(\theta)$ in the denominator. Hence, another natural tuning approach consists in truncating the sum in (3.4) as soon as $\hat{\lambda}_{m|q}(\theta)$ is below a certain threshold $\epsilon = \epsilon_n$. Hence, we choose

$$K_\theta = \text{argmax}_{m \ge 1}\{\hat{\lambda}_{m|q}(\theta) > \epsilon_n\}.$$

In Section 5, we use $\epsilon_n = n^{-1/2}$.

**Final prediction error (FPE).** This method is more complex and is of independent interest since it is applicable to the static functional linear model as well. The new data driven approach is explained in Appendix A, which also contains its theoretical justification.

# 5. Assessment of the performance in finite samples

## 5.1. Data generating processes and numerical implementation of the estimators

We work with a scalar response model of the form

$$Y_t = \sum_{|h| \le H} \langle X_{t-h}, b_h \rangle + e_t, \qquad H \ge 0.$$

In order to generate it, we assume a finite dimensional specification $b_h = \sum_{k=1}^{d} b_{h;k} f_k$, where the functions $f_k$ form an orthonormal system in $H$. If we expand the curves $X_t$ along the basis $(f_k)$ we have $X_t = \sum_{k\geq 1} \langle X_t, f_k\rangle f_k$, and thus

$$Y_t = \sum_{|h|\leq H} \sum_{k=1}^{d} b_{h;k} \langle X_{t-h}, f_k\rangle + e_t.$$

Hence, we can rearrange this functional lagged regression model in vector form as

$$Y_t = \sum_{|h|\leq H} \mathbf{X}'_{t-h} \mathbf{b}_h + e_t,$$

where $\mathbf{b}_h = (b_{h;1}, \ldots, b_{h;d})'$ and $\mathbf{X}_t := (\langle X_t, f_1\rangle, \ldots, \langle X_t, f_d\rangle)'$.

A similar discretization can be done for the operator $B_\theta$ and the covariance operators $C_h^X$ and $C_h^{YX}$. More specifically, we can write

$$B_\theta(x) = \left( \sum_{|h|\leq H} \mathbf{b}'_h \exp(-\mathrm{i}h\theta) \right) \mathbf{x} =: \mathfrak{B}_\theta \mathbf{x},$$

and

$$C_h^X(x) = \mathbf{f}'_d \big( E\mathbf{X}_{t+h}\mathbf{X}'_t \big)\mathbf{x} := \mathbf{f}'_d \mathfrak{C}_h^X \mathbf{x} \quad \text{and} \quad C_h^{YX}(x) = \big( EY_{t+h}\mathbf{X}'_t \big)\mathbf{x} := \mathfrak{C}_h^{YX}\mathbf{x},$$

where $\mathbf{x} = (\langle x, f_1\rangle, \ldots, \langle x, f_d\rangle)'$ and $\mathbf{f}'_d = (f_1, \ldots, f_d)$. In other words, all involved operators have their corresponding matrices, which act on the coefficients of $x$ projected onto $\mathrm{span}(f_1, \ldots, f_d)$, instead of acting on $x$ itself. Following this line of argumentation it follows that

$$B_\theta(x) = \mathfrak{F}_\theta^{YX} \big( \mathfrak{F}_\theta^X \big)^{-1} \mathbf{x},$$

where $\mathfrak{F}_\theta^X$ and $\mathfrak{F}_\theta^{YX}$ are spectral densities related to the matrices $(\mathfrak{C}_h^X)$ and $(\mathfrak{C}_h^{YX})$, respectively. Furthermore, it can be readily shown that

$$\widehat{B}_{\theta|p,q,k}(x) = \widehat{\mathfrak{B}}_{\theta|p,q,k}\, \mathbf{x} := \widehat{\mathfrak{F}}_{\theta|p}^{YX} \left( \sum_{m=1}^{k} \frac{1}{\hat{\lambda}_{m|q}(\theta)} \widehat{\varphi}_{m|q}(\theta) \widehat{\varphi}^*_{m|q}(\theta) \right) \mathbf{x},$$

where $\hat{\lambda}_{m|q}(\theta)$ and $\widehat{\varphi}_{m|q}(\theta)$ are the eigenvalues and eigenvectors of

$$\widehat{\mathfrak{F}}_{\theta|q}^X = \frac{1}{2\pi} \sum_{|h|\leq q} w_q(h) \widehat{\mathfrak{C}}_h^X e^{-\mathrm{i}h\theta},$$

and where

$$\widehat{\mathfrak{F}}_{\theta|p}^{YX} = \frac{1}{2\pi} \sum_{|h|\leq p} w_p(h) \widehat{\mathfrak{C}}_h^{YX} e^{-\mathrm{i}h\theta}.$$

Here $\widehat{\mathfrak{C}}_h^X$ and $\widehat{\mathfrak{C}}_h^{YX}$ are the usual empirical covariance matrices related the sequence $((Y_t, \mathbf{X}_t)\colon 1 \leq t \leq n)$ and $w_q(h)$ are the Bartlett weights.

Finally, $\hat{b}_\ell = \frac{1}{2\pi}\int_{-\pi}^{\pi} \widehat{\mathfrak{B}}_{\theta|p,q,k} e^{\mathrm{i}h\theta} d\theta$. (Note that we do allow $p, q$ and $k$ to depend on $\theta$.) Since this term cannot be computed explicitly, we use the numerical approximation

$$\hat{b}_\ell(x) = \mathbf{f}_d' \left( \frac{1}{|\Theta|} \sum_{\theta \in \Theta} \widehat{\mathfrak{B}}_{\theta|p,q,k} e^{\mathrm{i}\ell\theta} \right) \mathbf{x} =: \mathbf{f}_d' \, \widehat{\mathbf{b}}_\ell \, \mathbf{x}, \tag{5.1}$$

where $\Theta$ is a fine mesh on $[-\pi, \pi]$.

## 5.2. Simulation settings and results

For the simulation study we have chosen the following settings.

- We set $b_h = 0$ if $h \notin \{0, \ell\}$ where $\ell \in \{1, 3\}$. Furthermore, $b_{0;k} = \alpha_0(d - k + 1)$ and $b_{\ell;k} = \alpha_1(d - k + 1)^2$, with $\alpha_0$ and $\alpha_1$ such that $\|b_0\| = \beta_0$ and $\|b_\ell\| = \beta_\ell$ and $d = 15$. We choose $\beta_0 \in \{0.5, 1\}$ and $\beta_\ell \in \{0.5, 1\}$. The curves $\mathbf{f}_d$ are the first $d$ Fourier basis functions.

- We assume that $(e_t) \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, with $\sigma^2 \in \{0.1, 0.5\}$, and suppose that $\mathbf{X}_t = \Psi \mathbf{X}_{t-1} + \boldsymbol{\varepsilon}_t$, where $(\boldsymbol{\varepsilon}_t) \overset{\text{i.i.d.}}{\sim} N_d(0, \Sigma_d)$, $\Psi = (\psi_{ij} \colon 1 \le i, j \le d)$ satisfies $\psi_{ij} = c/(ij)$, with $\|\Psi\| \in \{0, 0.7\}$. Obviously, $\|\Psi\| = 0$ amounts to the i.i.d. setting. Since
$$E\|X_t\|^2 = \sum_{k \ge 1} \mathrm{Var}(\langle X_t, f_k \rangle) < \infty,$$
we assume that the elements of $\mathrm{diag}(\Sigma_d)$ are decaying. More precisely we set $\mathrm{diag}(\Sigma_d) = (1, 1/2^i, 1/3^i, \ldots, 1/d^i)$, with $i \in \{2, 4\}$.

- The sample size is $n \in \{250, 500\}$. The parameters $p$ and $q$ are set equal to 10. Variation of these parameters did invoke much changes for the output. Under each settings specified above, we make 100 simulation runs and use the three methods described in Section 4 for tuning the dimension parameter $K$. For the ET criterion we chose $\epsilon_n = 1/\sqrt{n}$.

- We compare two measures of fit. The first is the relative absolute error of the estimators of the two non-zero lags:
$$\delta^{\mathrm{err}} = \frac{1}{2} \left( \frac{\|b_0 - \hat{b}_0\|}{\beta_0} + \frac{\|b_\ell - \hat{b}_\ell\|}{\beta_\ell} \right).$$
The second is the mean square criterion:
$$\delta^{\mathrm{MSE}} = \frac{1}{n} \sum_{t=1}^{n} \left( Y_t - \sum_{|h| \le 3} \hat{b}_h (X_{t-h} I\{1 \le t - h \le n\}) \right)^2.$$
Each simulation run gives with each setting a sample $\delta_1, \ldots, \delta_{100}$. We compute the mean and the standard deviation.

| $\|\Psi\|$ | $n$ | $\ell$ | CV | | FPE | | TH | |
|---|---|---|---|---|---|---|---|---|
| | | | mean | sd | mean | sd | mean | sd |
| 0 | 250 | 1 | 0.687 | 0.154 | **0.621** | 0.090 | 0.629 | 0.045 |
| | | 3 | 0.669 | 0.107 | 0.660 | 0.103 | **0.653** | 0.047 |
| | 500 | 1 | 0.564 | 0.117 | **0.514** | 0.080 | 0.565 | 0.034 |
| | | 3 | 0.608 | 0.113 | **0.593** | 0.070 | 0.596 | 0.038 |
| 0.7 | 250 | 1 | 0.526 | 0.242 | 0.512 | 0.144 | **0.323** | 0.052 |
| | | 3 | 0.674 | 0.219 | 0.631 | 0.091 | **0.513** | 0.045 |
| | 500 | 1 | 0.531 | 0.292 | 0.571 | 0.190 | **0.339** | 0.063 |
| | | 3 | 0.592 | 0.144 | 0.552 | 0.099 | **0.483** | 0.039 |

Table 1: Mean and standard deviation of $\delta^{\text{err}}$ for three methods under $\beta_0 = \beta_1 = 1, \sigma^2 = 0.5$ and $i = 2$.

| $\|\Psi\|$ | $n$ | $\ell$ | CV | | FPE | | TH | |
|---|---|---|---|---|---|---|---|---|
| | | | mean | sd | mean | sd | mean | sd |
| 0 | 250 | 1 | 0.479 | 0.081 | **0.474** | 0.066 | 0.481 | 0.045 |
| | | 3 | **0.507** | 0.065 | 0.514 | 0.069 | 0.512 | 0.045 |
| | 500 | 1 | **0.480** | 0.045 | 0.486 | 0.039 | 0.490 | 0.031 |
| | | 3 | **0.519** | 0.046 | 0.528 | 0.042 | 0.524 | 0.029 |
| 0.7 | 250 | 1 | 0.485 | 0.064 | **0.453** | 0.050 | 0.476 | 0.047 |
| | | 3 | 0.541 | 0.096 | **0.518** | 0.065 | 0.531 | 0.050 |
| | 500 | 1 | 0.505 | 0.042 | **0.479** | 0.036 | 0.497 | 0.032 |
| | | 3 | 0.549 | 0.044 | **0.533** | 0.039 | 0.547 | 0.032 |

Table 2: Mean and standard deviation of $\delta^{\text{MSE}}$ for three methods under $\beta_0 = \beta_1 = 1, \sigma^2 = 0.5$ and $i = 2$.

**Discussion of results.** Due to a large number of settings we do not show all our results. Rather, we display in Tables 1 and 2 a few selected and representative settings. Overall we found that the relative absolute error $\delta^{\text{err}}$ is typically smallest when we use method TH. In particular, for the dependent setting and fast decay of eigenvalues, this method clearly outperforms CV and FPE. The FPE method performs best when the $X_t$ are independent and when $i = 2$ (slower decaying eigenvalues).

The situation is quite different if we look at the model fit using $\delta^{\text{MSE}}$. Then, overall the method FPE performs best, especially under dependence. Here TH method produces the largest errors among the three approaches. The CV method can slightly outperform FPE when the $X_t$ are independent and when $i = 2$.

We also considered a setting with slowly decaying response coefficients, namely, $\|b_0\| = 1.0, \|b_1\| = 0.9, \|b_2\| = 0.7, \|b_3\| = 0.5, \|b_4\| = 0.3$ and $\|b_5\| = 0.1$. The conclusions reported above remain the same: TH method performs when $\delta^{\text{err}}$ is used as the performance criterion, methods FPE and CV perform better if $\delta^{\text{MSE}}$ is used.

In conclusion we recommend to use the method TH if the target is estimation. Possibly this method could be further improved by tuning the selection of the threshold $\epsilon_n$. If the target is to use the model for prediction, then FPE is preferable, though

it comes with larger numerical costs than TH. Method CV cannot be recommended, because it is numerically expensive and was not a clear winner in any of the settings tested.

### 5.3. Response to the last bullet point of referee jtsa-14 ...

We performed the additional simulations required by the referee. Their results made us more confident in our general conclusions and recommendations. In the discussion of results, we added the following sentence: "We also considered a setting with slowly decaying response coefficients, namely, $\|b_0\| = 1.0, \|b_1\| = 0.9, \|b_2\| = 0.7, \|b_3\| = 0.5, \|b_4\| = 0.3$ and $\|b_5\| = 0.1$. The conclusions reported above remain the same: TH method performs when $\delta^{\text{err}}$ is used as the performance criterion, methods FPE and CV perform better if $\delta^{\text{MSE}}$ is used."

We now provide more details for the referee to evaluate. The simulation setting is the same as described in the paper except that the first and last bullet points are modified to:

- We set $b_h = 0$ if $h \notin \{0, 1, ..., 5\}$. Furthermore, for $i \in \{0, 1, ..., 5\}$ we set $b_{i;k} = \alpha_i(d-k+1) + \sqrt{d-k+1}c_{i;k}$, with indepentdent $c_{i;k} \sim \mathcal{N}(0, 1)$. In each simulation we draw coefficients $c_{i;k}$ and they remain fixed within the given run. We set $(\alpha_i)_{0\le i\le 5}$ such that $\|b_0\| = 1.0, \|b_1\| = 0.9, \|b_2\| = 0.7, \|b_3\| = 0.5, \|b_4\| = 0.3$ and $\|b_5\| = 0.1$. The curves $\mathbf{f}_d$ are the first $d = 15$ Fourier basis functions.

- We compare two measures of fit. The first is the relative absolute error of the estimators of the two non-zero lags:

$$\delta^{\text{err}} = \frac{1}{6}\sum_{\ell=0}^{5} \frac{\|b_\ell - \hat{b}_\ell\|}{\|b_\ell\|}.$$

  The second is the mean square criterion:

$$\delta^{\text{MSE}} = \frac{1}{n}\sum_{t=1}^{n}\left(Y_t - \sum_{|h|\le 5}\hat{b}_h(X_{t-h}I\{1 \le t-h \le n\})\right)^2.$$

Results of the additional simulations are presented in Tables 3 and 4.

# 6. Proofs

It is assumed that all random elements in the sequel are defined on a common probability space $(\Omega, \mathcal{A}, P)$. Recall that the vector space of all Hilbert–Schmidt operators acting on a Hilbert space $H$ is itself a Hilbert space with the inner product $\langle K, L\rangle_{\mathcal{S}} = \sum_{m\ge 1}\langle K(e_m), L(e_m)\rangle$, where $(e_m)$ is any orthonormal basis of $H$. The tensor product $x\otimes y$ of $x, y \in H$ is a Hilbert–Schmidt operator defined by $x\otimes y(z) = x\langle z, y\rangle$ whose norm is $\|x \otimes y\|_{\mathcal{S}} = \|x\|\|y\|$.

| $\|\Psi\|$ | $i$ | $n$ | $\sigma^2$ | CV | | FPE | | TH | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | mean | sd | mean | sd | mean | sd |
| 0 | 2 | 250 | 0.1 | **0.971** | 0.130 | 0.992 | 0.150 | 0.981 | 0.090 |
| | | | 0.5 | 1.217 | 0.376 | 1.315 | 0.260 | **1.049** | 0.096 |
| | | 500 | 0.1 | **0.835** | 0.114 | 0.843 | 0.117 | 0.887 | 0.074 |
| | | | 0.5 | 1.138 | 0.396 | 1.175 | 0.213 | **0.965** | 0.086 |
| | 4 | 250 | 0.1 | 1.763 | 2.194 | 3.515 | 1.713 | **1.091** | 0.066 |
| | | | 0.5 | 2.082 | 3.359 | 6.225 | 3.722 | **1.117** | 0.080 |
| | | 500 | 0.1 | 1.447 | 1.685 | 2.921 | 1.372 | **1.023** | 0.036 |
| | | | 0.5 | 2.323 | 4.270 | 5.534 | 3.013 | **1.068** | 0.071 |
| 0.7 | 2 | 250 | 0.1 | 1.052 | 0.494 | 1.541 | 0.352 | **0.907** | 0.125 |
| | | | 0.5 | 1.028 | 0.477 | 1.742 | 0.443 | **0.987** | 0.135 |
| | | 500 | 0.1 | **0.950** | 0.382 | 1.637 | 0.294 | 0.964 | 0.124 |
| | | | 0.5 | **1.004** | 0.477 | 1.801 | 0.386 | 1.024 | 0.120 |
| | 4 | 250 | 0.1 | 3.944 | 3.813 | 5.024 | 1.633 | **0.868** | 0.115 |
| | | | 0.5 | 3.185 | 2.530 | 7.695 | 3.192 | **0.882** | 0.107 |
| | | 500 | 0.1 | 3.797 | 2.381 | 5.185 | 1.485 | **0.832** | 0.084 |
| | | | 0.5 | 5.143 | 5.880 | 7.041 | 2.836 | **0.871** | 0.127 |

Table 3: Mean and standard deviation of $\delta^{\mathrm{err}}$ for three methods.

| $\|\Psi\|$ | $i$ | $n$ | $\sigma^2$ | CV | | FPE | | TH | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | mean | sd | mean | sd | mean | sd |
| 0 | 2 | 250 | 0.1 | **0.1020** | 0.0198 | 0.1231 | 0.0324 | 0.1339 | 0.0131 |
| | | | 0.5 | **0.4562** | 0.0824 | 0.4644 | 0.0789 | 0.4742 | 0.0448 |
| | | 500 | 0.1 | **0.1071** | 0.0130 | 0.1168 | 0.0128 | 0.1294 | 0.0106 |
| | | | 0.5 | **0.4721** | 0.0498 | 0.4882 | 0.0435 | 0.4925 | 0.0275 |
| | 4 | 250 | 0.1 | **0.1026** | 0.0149 | 0.1037 | 0.0199 | 0.1244 | 0.0148 |
| | | | 0.5 | 0.4703 | 0.0654 | **0.4529** | 0.0582 | 0.5029 | 0.0492 |
| | | 500 | 0.1 | **0.1069** | 0.0096 | 0.1078 | 0.0142 | 0.1139 | 0.0072 |
| | | | 0.5 | 0.4834 | 0.0482 | **0.4768** | 0.0401 | 0.4965 | 0.0319 |
| 0.7 | 2 | 250 | 0.1 | 0.2286 | 0.0542 | **0.1854** | 0.0347 | 0.2240 | 0.0320 |
| | | | 0.5 | 0.5885 | 0.1019 | **0.4998** | 0.0641 | 0.5649 | 0.0646 |
| | | 500 | 0.1 | 0.2406 | 0.0374 | **0.2034** | 0.0226 | 0.2322 | 0.0244 |
| | | | 0.5 | 0.6116 | 0.0721 | **0.5399** | 0.0390 | 0.5915 | 0.0338 |
| | 4 | 250 | 0.1 | **0.1072** | 0.0236 | 0.1091 | 0.0184 | 0.2132 | 0.0256 |
| | | | 0.5 | 0.4779 | 0.0953 | **0.4552** | 0.0594 | 0.5825 | 0.0535 |
| | | 500 | 0.1 | **0.1082** | 0.0129 | 0.1092 | 0.0148 | 0.2151 | 0.0232 |
| | | | 0.5 | **0.4825** | 0.0565 | 0.4829 | 0.0428 | 0.6060 | 0.0398 |

Table 4: Mean and standard deviation of $\delta^{\mathrm{MSE}}$ for three methods.

## 6.1. Auxiliary lemmas

We collect in this section several simple lemmas referred to in Sections 2 and 3, and used in the arguments that follow.

**Lemma 6.1.** *If $X, Z \in H$ are square integrable, and $\Psi$ is a Hilbert–Schmidt operator, then*

$$\|\mathrm{Cov}(\Psi(X), Z)\|_{\mathcal{S}} \leq \|\Psi\|_{\mathcal{S}} \, \|\mathrm{Cov}(X, Z)\|_{\mathcal{S}}.$$

**Proof.** To lighten the notation, assume $EX = 0, EZ = 0$. Then, for any orthonormal basis $(e_j, j \geq 1)$,

$$\|\mathrm{Cov}(\Psi(X), Z)\|_{\mathcal{S}}^2 = \sum_{j=1}^{\infty} \|E[\Psi(X)\langle e_j, Z\rangle]\|^2 = \sum_{j=1}^{\infty} \|\Psi(E[\langle e_j, Z\rangle X])\|^2,$$

where we used the fact that expectation commutes with any bounded operator. It follows that

$$\|\mathrm{Cov}(\Psi(X), Z)\|_{\mathcal{S}}^2 \leq \sum_{j=1}^{\infty} \|\Psi\|_{\mathcal{L}}^2 \|\mathrm{Cov}(X, Z)(e_j)\|^2 \leq \|\Psi\|_{\mathcal{L}}^2 \, \|\mathrm{Cov}(X, Z)\|_{\mathcal{S}}^2.$$

The claim then follows because $\|\Psi\|_{\mathcal{L}} \leq \|\Psi\|_{\mathcal{S}}$. □

**Lemma 6.2.** *Under Assumption 2.1, $\sum_{h\in\mathbb{Z}} \|C_h^{YX}\|_{\mathcal{S}} < \infty$.*

**Proof.** Since $\mathrm{Cov}(\varepsilon_\ell, X_k) = 0$,

$$\|C_h^{YX}\|_{\mathcal{S}} = \|\sum_k \mathrm{Cov}(b_k(X_{h-k}), X_0)\|_{\mathcal{S}} \leq \sum_k \|\mathrm{Cov}(b_k(X_{h-k}), X_0)\|_{\mathcal{S}}.$$

Therefore, by Lemma 6.1,

$$\begin{aligned}\sum_h \|C_h^{YX}\|_{\mathcal{S}} &\leq \sum_h \sum_k \|b_k\|_{\mathcal{S}} \, \|\mathrm{Cov}(X_{h-k}, X_0)\|_{\mathcal{S}} \\ &= \sum_k \|b_k\|_{\mathcal{S}} \sum_h \|C_h^X\|_{\mathcal{S}},\end{aligned}$$

so the claim follows from (2.3). □

**Lemma 6.3.** *Suppose Assumptions 2.1 and 2.2 hold. Then, for any $\theta \in [-\pi, \pi]$, the operator $\mathcal{F}_\theta^X$ is unbounded. It is invertible on*

$$D_\theta = \left( f \in H \colon \sum_{m\geq 1} \langle f, \varphi_m(\theta)\rangle^2 \lambda_m^{-2}(\theta) < \infty \right).$$

**Proof.** To show that the inverse of $\mathcal{F}_\theta^X$ does not exist as a bounded operator, we must find a sequence $f_n \to 0$ such that $\liminf_{n\to\infty} \| \left(\mathcal{F}_\theta^X\right)^{-1}(f_n)\| > 0$. As noted in the discussion leading to (3.2), for any $\theta \in [-\pi, \pi]$, the operator $\mathcal{F}_\theta^X$ is Hilbert-Schmidt, symmetric and non-negative definite. Since $\mathcal{F}_\theta$ is Hilbert-Schmidt,

$\sum_{m\geq 1}\lambda_m^2(\theta)<\infty$, and thus $\lambda_m(\theta)\to 0$, as $m\to\infty$. Since all eigenvalues $\lambda_m(\theta)$ are positive and $(\varphi_m(\theta)\colon m\geq 1)$ is an orthonormal basis of $H$,

$$\left(\mathcal{F}_\theta^X\right)^{-1}(f)=\sum_{m\geq 1}\frac{1}{\lambda_m(\theta)}\langle f,\varphi_m(\theta)\rangle\varphi_m(\theta),$$

if the series converges, i.e. if $f\in D_\theta$. Setting $f_n=\lambda(\theta)\varphi_n(\theta)$, we see that $f_n\to 0$ and $\|\left(\mathcal{F}_\theta^X\right)^{-1}(f_n)\|=\|\varphi_n\|=1$. □

## 6.2. Proofs of Lemma 4.1 and Theorem 4.1

We begin with a lemma which allows us to bound the difference between sample and population auto– and cross–covariance operators. It is an extension of a fundamental result that the difference between the sample and population covariance operators ($h=0, X=Y$) is of the order $n^{-1/2}$, see Bosq (2000) and Horváth and Kokoszka (2012). It is a result likely to find applications in many asymptotic arguments in the context of functional time series.

**Lemma 6.4.** *Suppose Assumption 2.3 holds. Then there is a constant $\kappa$, independent of $n$ and $h$, such that $E\|\widehat{C}_h^X-C_h^X\|_{\mathcal{S}}\leq\kappa\, n^{-1/2}$. If, in addition, Assumption 2.1 holds, then $E\|\widehat{C}_h^{YX}-C_h^{YX}\|_{\mathcal{S}}\leq\kappa\, n^{-1/2}$.*

**Proof.** We present the argument for $C_h^X$ and $h\geq 0$, which contains the key points. The result for the cross–covariance operators is established in a similar way using the lemmas of Section 6.1. We will repeatedly use the following simple relation: $|\langle x_1\otimes y_1, x_2\otimes y_2\rangle_{\mathcal{S}}|=|\langle x_1,y_1\rangle\langle x_2,y_2\rangle|\leq\|x_1\|\|x_2\|\|y_1\|\|y_2\|$.

Using stationarity and diagonal summation, we obtain

$$nE\|\widehat{C}_h^X-C_h^X\|_{\mathcal{S}}^2=\sum_{|r|<n}\left(1-\frac{|r|}{n}\right)E\langle X_{r+h}\otimes X_r-C_h^X, X_h\otimes X_0-C_h^X\rangle_{\mathcal{S}}.$$

By Definition 2.1, if $r\in\{0,\ldots,h-1\}$, then $X_{r+h}^{(r)}$ is independent of $X_r$, $X_h$ and $X_0$. It follows easily that $E\langle X_{r+h}^{(r)}\otimes X_r, X_h\otimes X_0\rangle_{\mathcal{S}}=0$. Hence the summands above are bounded by

$$\begin{aligned}|E\langle X_{r+h}\otimes X_r, X_h\otimes X_0\rangle_{\mathcal{S}}|&=\left|E\langle (X_{r+h}-X_{r+h}^{(r)})\otimes X_r, X_h\otimes X_0\rangle_{\mathcal{S}}\right|\\&\leq\nu_4^3(X_0)\nu_4(X_0-X_0^{(r)}).\end{aligned}$$

When $r\geq h$ we get

$$\begin{aligned}&\left|E\langle X_{r+h}\otimes X_r-C_h^X, X_h\otimes X_0-C_h^X\rangle_{\mathcal{S}}\right|\\&\quad=\left|E\langle X_{r+h}\otimes X_r-[X_{r+h}\otimes X_r]^{(r)}, X_h\otimes X_0-C_h^X\rangle_{\mathcal{S}}\right|\\&\quad\leq E\left[\|X_{r+h}\otimes X_r-X_{r+h}^{(r)}\otimes X_r^{(r-h)}\|_{\mathcal{S}}\left(\|X_h\otimes X_0\|_{\mathcal{S}}+\|C_h^X\|_{\mathcal{S}}\right)\right]\\&\quad\leq 2\left[E\|X_{r+h}\otimes X_r-X_{r+h}^{(r)}\otimes X_r^{(r-h)}\|_{\mathcal{S}}^2\right]^{1/2}\nu_4(X_0),\end{aligned}$$

where we used

$$\begin{aligned} E\left(\|X_h \otimes X_0\|_{\mathcal{S}} + \|C_h^X\|_{\mathcal{S}}\right)^2 &\leq 2E(\|X_h \otimes X_0\|_{\mathcal{S}}^2 + 2\|C_h^X\|_{\mathcal{S}}^2 \\ &\leq 2E\|X_h\|^2\|X_0\|^2 + 2E\|X_h \otimes X_h\|_{\mathcal{S}}^2 \\ &\leq 2\left(E\|X_h\|^4\right)^{1/2} E\left(\|X_0\|^4\right)^{1/2} + 2E\|X_h\|^2\|X_h\|^2 \\ &\leq 4\nu_4^2(X_0). \end{aligned}$$

Some further basic estimates show that

$$\begin{aligned} &\left[E\|X_{r+h} \otimes X_r - X_{r+h}^{(r)} \otimes X_r^{(r-h)}\|_{\mathcal{S}}^2\right]^{1/2} \\ &\quad \leq \sqrt{2}\nu_4(X_0)\left[\nu_4(X_0 - X_0^{(r-h)}) + \nu_4(X_0 - X_0^{(r)})\right]. \end{aligned}$$

Similar estimates can be obtained when $r < 0$, and the result follows from (2.6). □

It is convenient to introduce the following remainder terms:

$$\tau^X(h) = \sum_{|k|\geq h} \|C_k^X\|_{\mathcal{S}}; \quad \tau^{YX}(h) = \sum_{|k|\geq h} \|C_k^{YX}\|_{\mathcal{S}}; \quad \tau^b(h) = \sum_{|k|\geq h} \|b_k\|_{\mathcal{S}}.$$

*Proof of Lemma 4.1.* By repeated application of the triangle inequality, we obtain

$$\begin{aligned} &\sup_{\theta\in[-\pi,\pi]} \|\mathcal{F}_\theta^X - \widehat{\mathcal{F}}_\theta^X\|_{\mathcal{L}} \\ &\quad \leq \frac{1}{2\pi}\left\{\sum_{|h|\leq q} \|C_h^X - \widehat{C}_h^X\|_{\mathcal{L}} + \sum_{|h|\leq q} |1 - \omega_q(h)|\|C_h^X\|_{\mathcal{L}} + \tau^X(q)\right\}. \end{aligned}$$

Since by Lemma 6.4, $\sum_{|h|\leq q} E\|C_h^X - \widehat{C}_h^X\|_{\mathcal{L}} = O(qn^{-1/2})$, the first term tends to zero. The term $\sum_{|h|\leq q} |1 - \omega_q(h)|\|C_h^X\|_{\mathcal{L}}$ tends to zero by (2.3), Assumption 3.1 and by the dominated convergence. Again by (2.3), it follows that $\tau^X(q) \to 0$. For example, one may then chose

$$\psi_n^X = \left\{qn^{-1/2} + \sum_{|h|\leq q} |1 - \omega_q(h)|\|C_h^X\|_{\mathcal{L}} + \tau^X(q)\right\}^{1-\gamma}, \quad \gamma \in (0,1).$$

The same arguments apply to the spectral cross–density operators. □

*Proof of Theorem 4.1.* Since

$$\max_{h\in\mathbb{Z}} \|\hat{b}_h - b_h\|_{\mathcal{S}} \leq \frac{1}{2\pi}\int_{-\pi}^{\pi} \|\widehat{B}_\theta - B_\theta\|_{\mathcal{S}} d\theta,$$

we focus on the estimation of the frequency response operator $B_\theta$. Define

$$\widetilde{B}_\theta = \widetilde{B}_\theta(K) = \sum_{m\leq K} \mathcal{F}_\theta^{YX}\left(\frac{1}{\lambda_m(\theta)}\varphi_m(\theta) \otimes \varphi_m(\theta)\right).$$

Then

$$\frac{1}{2}\|\widehat{B}_\theta - B_\theta\|_{\mathcal{S}}^2 \le \|\widehat{B}_\theta - \widetilde{B}_\theta\|_{\mathcal{S}}^2 + \|\widetilde{B}_\theta - B_\theta\|_{\mathcal{S}}^2.$$

Since, by (3.3),

$$\sum_{\ell\ge 1}\frac{1}{\lambda_\ell^2(\theta)}\|\mathcal{F}_\theta^{YX}(\varphi_\ell(\theta))\|^2 = \|B_\theta\|_{\mathcal{S}}^2 \le \left(\sum_{k\in\mathbb{Z}}\|b_k\|_{\mathcal{S}}\right)^2 < \infty,$$

we see that

$$\|\widetilde{B}_\theta - B_\theta\|_{\mathcal{S}}^2 = \sum_{\ell>K}\frac{1}{\lambda_\ell^2(\theta)}\|\mathcal{F}_\theta^{YX}(\varphi_\ell(\theta))\|^2 \to 0, \quad K\to\infty.$$

Thus, it remains to prove that

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\|\widehat{B}_\theta - \widetilde{B}_\theta\|_{\mathcal{S}}d\theta \xrightarrow{\mathcal{P}} 0, \tag{6.1}$$

and that $K \xrightarrow{\mathcal{P}} \infty$. Condition (6.1) can be replaced by

$$\int_{-\pi}^{\pi}\|\widehat{B}_\theta - \widetilde{B}_\theta\|_{\mathcal{S}}d\theta \times I_{A_n} \xrightarrow{\mathcal{P}} 0, \tag{6.2}$$

where $A_n \subset \mathcal{A}$ is defined as

$$A_n := \{\sup_\theta \|\mathcal{F}_\theta^X - \widehat{\mathcal{F}}_\theta^X\| \le \psi_n^X\} \cap \{\sup_\theta \|\mathcal{F}_\theta^{YX} - \widehat{\mathcal{F}}_\theta^{YX}\| \le \psi_n^{YX}\}.$$

This is because by Lemma 4.1 we have that $P(A_n)\to 1$.

We have

$$\begin{aligned}
\widehat{B}_\theta - \widetilde{B}_\theta &= \sum_{m=1}^{K}\left[\mathcal{F}_\theta^{YX}\left(\frac{1}{\lambda_m(\theta)}\varphi_m(\theta)\otimes\varphi_m(\theta)\right) - \widehat{\mathcal{F}}_\theta^{YX}\left(\frac{1}{\hat\lambda_m(\theta)}\hat\varphi_m(\theta)\otimes\hat\varphi_m(\theta)\right)\right] \\
&= \sum_{m=1}^{K}\mathcal{F}_\theta^{YX}\left(\frac{1}{\lambda_m(\theta)}\varphi_m(\theta)\otimes\varphi_m(\theta) - \frac{1}{\hat\lambda_m(\theta)}\hat\varphi_m(\theta)\otimes\hat\varphi_m(\theta)\right) \\
&\quad + \sum_{m=1}^{K}\left(\mathcal{F}_\theta^{YX} - \widehat{\mathcal{F}}_\theta^{YX}\right)\left(\frac{1}{\hat\lambda_m(\theta)}\hat\varphi_m(\theta)\otimes\hat\varphi_m(\theta)\right).
\end{aligned}$$

Thus, using $\|F\circ G\|_{\mathcal{S}} \le \|F\|_{\mathcal{L}}\|G\|_{\mathcal{S}}$, we get

$$\begin{aligned}
\|\widehat{B}_\theta - \widetilde{B}_\theta\|_{\mathcal{S}} &= \|\mathcal{F}_\theta^{YX}\|_{\mathcal{L}}\left\|\sum_{m=1}^{K}\left(\frac{1}{\lambda_m(\theta)}\varphi_m(\theta)\otimes\varphi_m(\theta) - \frac{1}{\hat\lambda_m(\theta)}\hat\varphi_m(\theta)\otimes\hat\varphi_m(\theta)\right)\right\|_{\mathcal{S}} \\
&\quad + \|\mathcal{F}_\theta^{YX} - \widehat{\mathcal{F}}_\theta^{YX}\|_{\mathcal{L}}\left(\sum_{m=1}^{K}\frac{1}{\hat\lambda_m^2(\theta)}\right)^{1/2}.
\end{aligned}$$

Since we have $\sup_{\theta\in[-\pi,\pi]} \|\mathcal{F}_\theta^{YX}\|_{\mathcal{L}} \leq \frac{1}{\pi}\tau^{YX}(0)$, (6.2) follows from

$$\int_{-\pi}^{\pi} \left\| \sum_{m=1}^{K} \left( \frac{1}{\lambda_m(\theta)} \varphi_m(\theta) \otimes \varphi_m(\theta) - \frac{1}{\hat{\lambda}_m(\theta)} \hat{\varphi}_m(\theta) \otimes \hat{\varphi}_m(\theta) \right) \right\|_{\mathcal{S}} d\theta \times I_{A_n} = o_P(1) \tag{6.3}$$

and

$$\psi_n^{YX} \int_{-\pi}^{\pi} W_\lambda^K(\theta) d\theta = O_P(1). \tag{6.4}$$

Relation (6.4) is already immediate from the condition $K \leq K^{(2)}$.

Some routine estimates show that the integrand in (6.3) is bounded by

$$\left\{ 2 \sum_{m=1}^{K} \frac{1}{\hat{\lambda}_m(\theta)} \|\varphi_m(\theta) - \hat{c}_m(\theta)\hat{\varphi}_m(\theta)\| + \sum_{m=1}^{K} \frac{|\hat{\lambda}_m(\theta) - \lambda_m(\theta)|}{\hat{\lambda}_m(\theta)\lambda_m(\theta)} \right\} \times I_{A_n}, \tag{6.5}$$

where $\hat{c}_m(\theta)$ is given as in Section 3. By Lemma 3.2 in Hörmann and Kokoszka (2010) we have that

$$\|\varphi_m(\theta) - \hat{c}_m(\theta)\hat{\varphi}_m(\theta)\| \leq \frac{2\sqrt{2}}{\hat{\alpha}_m(\theta)} \sup_{\theta\in[-\pi,\pi]} \|\mathcal{F}_\theta^X - \widehat{\mathcal{F}}_\theta^X\|_{\mathcal{L}},$$

and

$$\sup_{\theta\in[-\pi,\pi]} \sup_{m\geq 1} |\hat{\lambda}_m(\theta) - \lambda_m(\theta)| \leq \sup_{\theta\in[-\pi,\pi]} \|\mathcal{F}_\theta^X - \widehat{\mathcal{F}}_\theta^X\|_{\mathcal{L}}. \tag{6.6}$$

Thus we obtain the bound

$$4\sqrt{2} \sum_{m=1}^{K} \frac{\psi_n^X}{\hat{\lambda}_m(\theta)} \left[ \frac{1}{\hat{\alpha}_m(\theta)} + \frac{1}{\lambda_m(\theta)} \right] \times I_{A_n} \tag{6.7}$$

for (6.5). We further remark that on $A_n$ we have that $\lambda_m(\theta) \geq \hat{\lambda}_m(\theta) - |\lambda_m(\theta) - \hat{\lambda}_m(\theta)| \geq \hat{\lambda}_m(\theta) - \psi_n^X$. Therefore, since $K \leq K^{(1)}$, we have that (6.7) is bounded by

$$4\sqrt{2} \sum_{m=1}^{K} \frac{\psi_n^X}{\hat{\lambda}_m(\theta)} \left[ \frac{1}{\hat{\alpha}_m(\theta)} + \frac{2}{\hat{\lambda}_m(\theta)} \right] \leq 4\sqrt{2}\psi_n^X \left( W_\lambda^K(\theta) W_\alpha^K(\theta) + 2\big(W_\lambda^K(\theta)\big)^2 \right), \tag{6.8}$$

where we have made use of the Cauchy-Schwarz inequality in the last step. Using $K \leq K^{(3)}$ and $K \leq K^{(4)}$ it is now easy to infer that (6.2) holds.

It remains to show that $K \xrightarrow{\mathcal{P}} \infty$, i.e. that $K^{(i)} \to \infty$ for $1 \leq i \leq 4$.

Fix a large $k$ and observe that $P(K^{(1)} \geq k) = P(\inf_\theta \hat{\lambda}_k(\theta) \geq 2\psi_n^X)$. Now define $B_{k;n} := \{\sup_\theta |\hat{\lambda}_k(\theta) - \lambda_k(\theta)| \leq \delta_k/2\}$ where $\delta_k := \inf_\theta \lambda_k(\theta)$. From Assumption 2.2 it follows that $\delta_k > 0$. Furthermore, it follows from Lemma 4.1 and (6.6) that $P(B_{k;n}) \to 1$ for $n \to \infty$. On the other hand $\inf_\theta \hat{\lambda}_k(\theta) \geq \inf_\theta \lambda_k(\theta) - \sup_\theta |\hat{\lambda}_k(\theta) - \lambda_k(\theta)|$, so that on $B_{k;n}$ we have $\inf_\theta \hat{\lambda}_k(\theta) \geq \delta_k/2$. And hence, for $n$ large enough, we have $\inf_\theta \hat{\lambda}_k(\theta) \geq 2\psi_n^X$ on $B_{k;n}$. Consequently $P(K^{(1)} \geq k) \to 1$ for $n \to \infty$, irrespective of how large $k$ was chosen.

Now we prove $K^{(4)} \to \infty$. Fix again a big $k$ and notice that is suffices to show that $P(\int_{-\pi}^{\pi} \min_{1\le m\le k} \hat{\alpha}_m^{-2}(\theta)d\theta > x_n) \to 0$, for any $x_n \to \infty$. Define $B'_{k;n} := \{\sup_\theta |\hat{\alpha}_k(\theta) - \alpha_k(\theta)| \le \delta'_k/2\}$ where $\delta'_k := \inf_\theta \alpha_k(\theta)$ and set $A_{k;n} = \cap_{m=1}^k B'_{k;n}$. Then for any fixed $k$ we have $P(A_{k;n}) \to 1$ and on $A_{k;n}$ it holds that $\min_{1\le m\le k} \hat{\alpha}_m(\theta) \le \min_{1\le m\le k} \delta'_m/2 = r_k$. By Assumption 4.1 $r_k > 0$ for any $k$. Hence, on $A_{k;n}$ the integral $\int_{-\pi}^{\pi} \min_{1\le m\le k} \hat{\alpha}_m^{-2}(\theta)d\theta$ is bounded by $4\pi/r_k$ and this is smaller than $x_n$ when $n$ is big enough. This proves $K^{(4)} \to \infty$.

The proof of $K^{(2)} \to \infty$ and $K^{(3)} \to \infty$ is similar and therefore omitted.

□

# A. Appendix

In this appendix, we derive the FPE method of selecting the dimension parameter $K$ used in Sections 3 and 4. In section A.1, we discuss the relation of our spectral approach to the time domain estimation in functional regression. This motivates the derivation of the FPE method in Section A.2. Section A.3 contains the proofs of two results stated in Sections A.1 and A.2.

## A.1. Relation to ordinary functional regression

As before we consider complex Hilbert spaces $H$ and $H'$ and define for elements $(a, f), (b, g) \in H' \times H$ define $[(a, f), (b, g)] = \langle a, b\rangle + \langle f, g\rangle$. This defines an inner product on $H' \times H$, and with it the space becomes a Hilbert space. Let us fix a frequency $\theta \in [-\pi, \pi]$, and define a zero mean complex random element $\Delta = (\Upsilon, \Xi) \in L^2_{H'\times H}$ such that

$$C^\Delta = E\Delta \otimes \Delta = \begin{pmatrix} C^\Upsilon & C^{\Upsilon\Xi} \\ C^{\Xi\Upsilon} & C^\Xi \end{pmatrix} = \begin{pmatrix} \mathcal{F}_\theta^Y & \mathcal{F}_\theta^{YX} \\ \mathcal{F}_\theta^{XY} & \mathcal{F}_\theta^X \end{pmatrix}. \tag{A.1}$$

Now we regress $\Upsilon$ on $\Xi$, i.e. we seek $h_0 \in \mathcal{L}(H, H')$ (the space of bounded linear operators from $H$ to $H'$) which satisfies

$$h_0 = \mathrm{argmin}_{h\in\mathcal{L}(H,H')} E\|\Upsilon - h(\Xi)\|^2.$$

Then by the usual projection arguments, $h_0$ solves the equation $C^{\Upsilon\Xi} = h_0 \circ C^\Xi$. By the definition of $C^{\Upsilon\Xi}$ and $C^\Xi$, it follows that $h_0$ is also the solution to (3.1) and hence, by Assumption 2.2, is equal to $B_\theta$. Consequently, $h_0$, or equivalently $B_\theta$, can also be estimated from a random sample $((\Upsilon_k, \Xi_k) \colon 1 \le k \le L)$ by standard methods known from functional linear models. A typical estimator (see e.g. Cardot et al. (1999)) is

$$\widehat{h}_{0;d}(f) = \sum_{\ell=1}^{d} \frac{\widehat{C}^{\Upsilon\Xi}(\hat{v}_\ell)}{\hat{\gamma}_\ell}\langle f, \hat{v}_\ell\rangle := \sum_{\ell=1}^{d} \hat{b}_\ell\langle f, \hat{v}_\ell\rangle, \tag{A.2}$$

where $\widehat{C}^{\Upsilon\Xi}(f) := \frac{1}{L}\sum_{k=1}^{L} \Upsilon_k\langle f, \Xi_k\rangle$ and $\hat{\gamma}_\ell$ and $\hat{v}_\ell$ are the eigenvalues and eigenvectors of $\widehat{C}^\Xi(f) := \frac{1}{L}\sum_{k=1}^{L} \Xi_k\langle f, \Xi_k\rangle$.

In practice we do not know $C^{\Delta}$, but, as we will see in Lemma 4.1, below, it can be consistently estimated from the data, which then in turn allows to generate a random sample $((\Upsilon_i, \Xi_i)\colon 1 \le i \le L)$ with a covariance which is asymptotically equal to $C^{\Delta}$. A more direct approach is to define the functional discrete Fourier transforms

$$\Upsilon_{k|p} = \frac{1}{\sqrt{2\pi p}} \sum_{t=p(k-1)+1}^{pk} Y_t e^{-\mathrm{i}(t-p(k-1))\theta} \quad \text{and} \quad \Xi_{k|p} = \frac{1}{\sqrt{2\pi p}} \sum_{t=p(k-1)+1}^{pk} X_t e^{-\mathrm{i}(t-p(k-1))\theta}.$$

If we denote $\widehat{C}_p^{\Upsilon\Xi}$ and $\widehat{C}_p^{\Xi}$ covariance and cross-covariance operators related to the sequence $((\Upsilon_{k|p}, \Xi_{k|p})\colon 1 \le k \le L)$, the following lemma holds:

**Lemma A.1.** *Consider the estimator $\widehat{\mathcal{F}}^X_{\theta|p}$ with the Bartlett weights $w_p(h) = 1 - |h|/p$. Under Assumption 2.3 we have $\|\widehat{\mathcal{F}}^X_{\theta|p} - \widehat{C}_p^{\Xi}\|^2_{\mathcal{S}} = O_P(p^3/n)$. Under the same conditions we have $\|\widehat{\mathcal{F}}^{YX}_{\theta|p} - \widehat{C}_p^{\Upsilon\Xi}\|^2_{\mathcal{S}} = O_P(p^3/n)$.*

The lemma, which we prove in Section A.3, confirms that computing (A.2) from the variables $(\Upsilon_{k|p}, \Xi_{k|p})$, which serve as an approximation to a random sample $(\Upsilon_k, \Xi_k)$, yields an estimator which resembles closely $\widehat{B}_{\theta|p,p,d}$ in (3.4).

## A.2. Description of the FPE approach

In order to keep this discussion short, we only consider the scalar response case. This is in line with our simulation study. Our starting point is the alternative interpretation of $B_\theta$ discussed in Section A.1. Suppose we have an estimator $\widehat{h}_{0;d}$ for $B_\theta$ from a sample $((\Upsilon_k, \Xi_k)\colon 1 \le k \le L)$. Now we pick $(\Upsilon, \Xi)$ independent of this sample and set $K = K_\theta = \operatorname{argmin}_{d\ge 0} E|\Upsilon - \widehat{h}_{0;d}(\Xi)|^2$. Note that here, by the Riesz representation theorem, $\widehat{h}_{0;d}(\Xi)$ is of the from $\langle \Xi, \widehat{h}_{0;d}\rangle$. With $d = K$ in (A.2) we minimize the mean squared prediction error in this functional regression. The related model selection criterion is commonly known as *final prediction error (FPE)* criterion. Of course, to compute $K$ explicitly is mathematically infeasible, and therefore we go for an approximation. For this purpose, we first note that that the coefficients $\hat{b}_k$ in (A.2) satisfy

$$\hat{\boldsymbol{b}}_d := (\hat{b}_1, \dots, \hat{b}_d)' = \operatorname{argmin}_{(b_1,\dots,b_d)\in\mathbb{C}^d} \sum_{i=1}^{L} |\Upsilon_i - \sum_{\ell=1}^{d} b_\ell \langle \Xi_i, \hat{v}_\ell\rangle|^2.$$

Our problem is greatly simplified if we replace the empirical principal component scores by the population ones and set

$$\tilde{\boldsymbol{b}}_d := (\tilde{b}_1, \dots, \tilde{b}_d)' = \operatorname{argmin}_{(b_1,\dots,b_d)\in\mathbb{C}^d} \sum_{i=1}^{L} |\Upsilon_i - \sum_{\ell=1}^{d} b_\ell \langle \Xi, v_\ell\rangle|^2,$$

and then define $\widetilde{h}_{0;d}(\Xi) = \sum_{\ell=1}^{d} \tilde{b}_\ell \langle \Xi, v_\ell\rangle$ and $K = \operatorname{argmin}_{d\ge 0} E|\Upsilon - \widetilde{h}_{0;d}(\Xi)|^2$.

**Proposition A.1.** *Suppose that the* $((\Upsilon_i, \Xi_i)\colon 1 \le i \le L)$ *constitute a Gaussian random sample, with circularly-symmetric observation, i.e.* $E\Delta[\Delta, (a, f)] = 0$ *for any* $(a, f) \in H' \times H$*. Then for* $L > d$ *we have*

$$E|\Upsilon - \widetilde{h}_{0;d}(\Xi)|^2 = \sigma_d^2 \times \frac{L}{L-d},$$

*where* $\sigma_d^2 = \frac{1}{L-d} E(\boldsymbol{\Upsilon} - \mathfrak{X}\tilde{\boldsymbol{b}}_d)^*(\boldsymbol{\Upsilon} - \mathfrak{X}\tilde{\boldsymbol{b}}_d)$ *and* $\mathfrak{X} = (\langle \Xi_i, v_\ell \rangle\colon 1 \le i \le L; 1 \le \ell \le d)$, $\boldsymbol{\Upsilon} = (\Upsilon_1, \ldots, \Upsilon_L)'$.

The proof of this proposition is given in Section A.3. Assuming Gaussianity is not a restriction, since our estimator only relies on the second order structure of the data. Furthermore, by Panaretos and Tavakoli (2013a) we know that under general dependence assumptions the discrete Fourier transforms $\Upsilon_{i|p}$ and $\Xi_{i|p}$ are asymptotically $(p \to \infty)$ complex normal random elements.

The proposition then suggests to choose $d$ such that $\sigma_d^2 \times \frac{L}{L-d}$ is minimized. An unbiased estimate for the unknown $\sigma_d^2$ is

$$\frac{1}{L-d}(\boldsymbol{\Upsilon} - \mathfrak{X}\hat{\boldsymbol{b}}_d)^*(\boldsymbol{\Upsilon} - \mathfrak{X}\hat{\boldsymbol{b}}_d).$$

Finally, replacing the theoretical scores leads to the following dimension selection:

$$K = \operatorname{argmin}_{0 \le d < L} \frac{L}{(L-d)^2}(\hat{\boldsymbol{\Upsilon}} - \hat{\mathfrak{X}}\hat{\boldsymbol{b}}_d)^*(\hat{\boldsymbol{\Upsilon}} - \hat{\mathfrak{X}}\hat{\boldsymbol{b}}_d), \tag{A.3}$$

where $\hat{\mathfrak{X}} = (\langle \Xi_{i|p}, \hat{v}_\ell \rangle\colon 1 \le i \le L; 1 \le \ell \le d)$ and $\hat{\boldsymbol{\Upsilon}} = (\Upsilon_{1|p}, \ldots, \Upsilon_{L|p})'$.

## A.3. Proofs of Lemma A.1 and Proposition A.1

*Proof Lemma A.1.* We define

$$\widetilde{C}_h^X = \frac{1}{Lp} \sum_{k=0}^{L-1} \Big( \sum_{t=1}^{p-h} X_{t+h+kp} \otimes X_{t+kp} \Big), \quad \text{for } 0 \le h < p,$$

and

$$\widetilde{C}_h^X = \frac{1}{Lp} \sum_{k=0}^{L-1} \Big( \sum_{t=|h|+1}^{p} X_{t-|h|+kp} \otimes X_{t+kp} \Big), \quad \text{for } -p < h < 0.$$

Direct verification shows that

$$\widehat{C}_\theta^\Xi = \frac{1}{2\pi} \sum_{|h|<p} \widetilde{C}_h^X e^{-\mathrm{i}h\theta}.$$

For two random operators $A_n$ and $B_n$ we write $A_n = B_n + O_p(m_n)$ if $\|A_n - B_n\|_{\mathcal{S}} = O_P(m_n)$. Then, for $p > h \ge 0$, we deduce with the help of Lemma 6.4 that

$$\begin{aligned} n\widehat{C}_h^X - Lp\widetilde{C}_h^X &= \sum_{k=0}^{L-1} \Big( \sum_{t=p-h+1}^{p} X_{t+h+kp} \otimes X_{t+kp} \Big) + \sum_{t=Lp+1}^{n-h} X_{t+h} \otimes X_t \\ &= Lh\Big(C_h^X + O_p(L^{-1/2})\Big) = Lh\Big(\widehat{C}_h^X + O_p(L^{-1/2})\Big). \end{aligned}$$

The same bound can be derived for $h < 0$. Thus,

$$\widetilde{C}_h^X = \left(1 - \frac{|h|}{p}\right) \widehat{C}_h^X + \left(\frac{n}{Lp} - 1\right) \widehat{C}_h^X + O_P(L^{-1/2}),$$

and since $\frac{n}{Lp} - 1 \leq \frac{p}{n-p}$ we have that

$$\widetilde{C}_h^X = \left(1 - \frac{|h|}{p}\right) \widehat{C}_h^X + O_P\big((p/n)^{1/2}\big).$$

We conclude that $\|\widehat{C}_\theta^\Xi - \widehat{\mathcal{F}}_\theta^X\|_\mathcal{S}^2 = O_P\big(p^3/n\big)$. A similar bound can be obtained for $\widehat{C}_\theta^{\Upsilon\Xi} - \widehat{\mathcal{F}}_\theta^{YX}$. This proves Lemma A.1. □

*Proof of Proposition A.1.* We have

$$E|\Upsilon - \widetilde{h}_{0;d}(\Xi)|^2 = E|\Upsilon - \sum_{\ell=1}^d \tilde{b}_\ell \langle \Xi, v_\ell \rangle|^2 = E|\sum_{\ell=1}^d (b_\ell - \tilde{b}_\ell)\langle \Xi, v_\ell \rangle + Z|^2, \qquad \text{(A.4)}$$

where $Z = (\Upsilon - \langle \Xi, h_0 \rangle) + \sum_{\ell > d} b_\ell \langle \Xi, v_\ell \rangle$. We set $\varepsilon = \Upsilon - \langle \Xi, h_0 \rangle$. By the projection theorem it follows that $\mathrm{Cov}(\varepsilon, \Xi) = 0$. Furthermore, since we assume that $\tilde{b}_\ell$ are independent of $\Xi$, and since principal components scores are orthogonal it follows that (A.4) equals

$$E|\sum_{\ell=1}^d (b_\ell - \tilde{b}_\ell)\langle \Xi, v_\ell \rangle|^2 + E|Z|^2 = \sum_{\ell=1}^d E|b_\ell - \tilde{b}_\ell|^2 \gamma_\ell + E|Z|^2.$$

With $\Gamma = \mathrm{diag}(\gamma_1, \ldots, \gamma_d)$ and $\boldsymbol{Z} = (Z_1, \ldots, Z_L)'$ and $Z_i = \varepsilon_i + \sum_{\ell > d} b_\ell \langle \Xi_i, v_\ell \rangle$ we get

$$\begin{aligned}
\sum_{\ell=1}^d E|b_\ell - \tilde{b}_\ell|^2 \gamma_\ell &= E\big[(\tilde{\boldsymbol{b}}_d - \boldsymbol{b}_d)^* \Gamma (\tilde{\boldsymbol{b}}_d - \boldsymbol{b}_d)\big] \\
&= E\big[\mathbf{Z}^* \mathfrak{X} (\mathfrak{X}^* \mathfrak{X})^{-1} \Gamma (\mathfrak{X}^* \mathfrak{X})^{-1} \mathfrak{X}^* \mathbf{Z}\big] \\
&= \mathrm{tr}\big(\Gamma E\big[(\mathfrak{X}^* \mathfrak{X})^{-1} \mathfrak{X}^* \mathbf{Z}\mathbf{Z}^* \mathfrak{X} (\mathfrak{X}^* \mathfrak{X})^{-1}\big]\big). \qquad \text{(A.5)}
\end{aligned}$$

We have $E[\mathbf{Z}\mathbf{Z}^*] = E|Z|^2 I_L$. The imposed circular-symmetry implies that

$$E\Upsilon\langle \Xi, f \rangle = 0 \quad \text{and} \quad E\langle \Xi, f \rangle \langle \Xi, g \rangle = 0 \quad \forall f, g \in H. \qquad \text{(A.6)}$$

Consequently, by Gaussianity it follows that $\mathbf{Z}$ and $\mathfrak{X}$ are independent. (Note that two complex Gaussian random variables $U_1$ and $U_2$, say, are independent if and only if $\mathrm{Cov}(U_1, U_2) = \mathrm{Cov}(U_1, \overline{U_2}) = 0$.) We can therefore conclude by a simple conditioning argument that (A.5) simplifies to

$$E|Z|^2 \,\mathrm{tr}\big(E\big[\big((\mathfrak{X}\Gamma^{-1/2})^* (\mathfrak{X}\Gamma^{-1/2})\big)^{-1}\big]\big) =: E|Z|^2 \,\mathrm{tr}\big(EW^{-1}\big).$$

The matrix $W^{-1}$ is an inverse complex Wishart matrix with expectation $EW^{-1} = \frac{I_d}{L-d}$. Thus $E|\Upsilon - \widetilde{h}_{0;d}(\Xi)|^2 = E|Z|^2 \times \frac{L}{L-d}$. □

# References


A. Aue, S. Hörmann, L. Horváth, and M. Reimherr. Break detection in the covariance structure of multivariate time series models. *The Annals of Statistics*, 37:4046–4087, 2009.

D. Bosq. *Linear Processes in Function Spaces*. Springer, 2000.

G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis: Forecasting and Control*. Prentice Hall, Englewood Cliffs, third edition, 1994.

D. R. Brillinger. *Time Series: Data Analysis and Theory*. Holt, New York, 1975.

T. Cai and P. Hall. Prediction in functional linear regression. *The Annals of Statistics*, 34: 2159–2179, 2006.

H. Cardot, F. Ferraty, and P. Sarda. Functional linear model. *Statistics and Probability Letters*, 45:11–22, 1999.

H. Cardot, F. Ferraty, A. Mas, and P. Sarda. Testing hypothesis in the functional linear model. *Scandinavian Journal of Statistics*, 30:241–255, 2003.

J-M. Chiou and H-G. Müller. Diagnostics for functional regression via residual processes. *Computational Statistics and Data Analysis*, 15:4849–4863, 2007.

F. Comte and J. Johannes. Adaptive functional linear regression. *The Annals of Statistics*, 40:2765–2797, 2012.

C. Crambes, A. Kneip, and P. Sarda. Smoothing splines estimators for functional linear regression. *The Annals of Statistics*, 37:35–72, 2009.

R. Gabrys, L. Horváth, and P. Kokoszka. Tests for error correlation in the functional linear model. *Journal of the American Statistical Association*, 105:1113–1125, 2010.

S. Hörmann and L. Kidziński. A note on estimation in Hilbertian linear models. *Scandinavian Journal of Statistics*, 2014. Forthcoming.

S. Hörmann and P. Kokoszka. Weakly dependent functional data. *The Annals of Statistics*, 38:1845–1884, 2010.

S. Hörmann and P. Kokoszka. Functional time series. In C. R. Rao and T. Subba Rao, editors, *Time Series*, volume 30 of *Handbook of Statistics*. Elsevier, 2012.

S. Hörmann, L. Horváth, and R. Reeder. A functional version of the ARCH model. *Econometric Theory*, 29:267–288, 2013.

S. Hörmann, L. Kidziński, and M. Hallin. Dynamic functional principal components. *Journal of the Royal Statistical Society: Series B*, 2014. Forthcoming.

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

G. M. James, J. Wang, and J. Zhu. Functional linear regression that's interpretable. *The Annals of Statistics*, 37:2083–2108, 2009.

P. Kokoszka and M. Reimherr. Predictability of shapes of intraday price curves. *The Econometrics Journal*, 16:285–308, 2013.

A. N. Kolmogorov. Interpolation und Extrapolation von stationären zufälligen Folgen. *Bull. Acad. Sci. U.S.S.R.*, 5:3–14, 1941.

Y. Li and T. Hsing. On rates of convergence in functional linear regression. *Journal of Multivariate Analysis*, 98:1782–1804, 2007.

I. McKeague and B. Sen. Fractals with point impacts in functional linear regression. *The Annals of Statistics*, 38:2559–2586, 2010.

H-G. Müller and U. Stadtmüller. Generalized functional linear models. *The Annals of Statistics*, 33:774–805, 2005.

V. M. Panaretos and S. Tavakoli. Fourier analysis of stationary time series in function space. *The Annals of Statistics*, 41:568–603, 2013a.

V. M. Panaretos and S. Tavakoli. Cramér–Karhunen–Loève representation and harmonic principal component analysis of functional time series. *Stochastic Processes and their Applications*, 123:2779–2807, 2013b.

M. B. Priestley. *Spectral Analysis and Time Series*. Academic Press, 1981.

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, 2005.

X. Shao and W. B. Wu. Asymptotic spectral theory for nonlinear time series. *The Annals of Statistics*, 35:1773–1801, 2007.

R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications with R Examples*. Springer, 2011.

N. Wiener. *The Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications*. Wiley, 1949.

W. Wu. *Nonlinear System Theory: Another Look at Dependence*, volume 102. The National Academy of Sciences of the United States, 2005.

F. Yao, H-G. Müller, and J-L. Wang. Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33:2873–2903, 2005.