Feature extraction for functional time series: Theory and application to NIR spectroscopy data

Yang Yang[a,*], Yanrong Yang[b], Han Lin Shang[c]

[a]Department of Econometrics and Business Statistics, Monash University, Melbourne, VIC 3145, Australia
[b]Research School of Finance, Actuarial Studies and Statistics, Australian National University, Canberra, ACT 2601, Australia
[c]Department of Actuarial Studies and Business Analytics, Macquarie University, Sydney, NSW 2109, Australia

Abstract

We propose a novel method to extract global and local features of functional time series. The global features concerning the dominant modes of variation over the entire function domain, and local features of function variations over particular short intervals within function domain, are both important in functional data analysis. Functional principal component analysis (FPCA), though a key feature extraction tool, only focus on capturing the dominant global features, neglecting highly localized features. We introduce a FPCA-BTW method that initially extracts global features of functional data via FPCA, and then extracts local features by block thresholding of wavelet (BTW) coefficients. Using Monte Carlo simulations, along with an empirical application on near-infrared spectroscopy data of wood panels, we illustrate that the proposed method outperforms competing methods including FPCA and sparse FPCA in the estimation functional processes. Moreover, extracted local features inheriting serial dependence of the original functional time series contribute to more accurate forecasts. Finally, we develop asymptotic properties of FPCA-BTW estimators, discovering the interaction between convergence rates of global and local features.

Keywords: Functional Principal Component Analysis, Long-run Covariance Estimation, Near-infrared Spectroscopy Data, Regularized Wavelet Approximation.

1. Introduction

The rapid improvements in automated data acquisition technology allow researchers to access functional data more frequently. Functional data sequentially recorded over time are often considered as finite realizations of a functional stochastic process $\{\mathcal{X}_t(u)\}_{t\in\mathbb{Z}}$, where the time parameter t is discrete, and the parameter u is a continuum bounded within a finite interval domain $[a, b]$. Observations $\{\mathcal{X}_t(u)\}_{t\in\mathbb{Z}}$ are commonly referred to as functional time series. Functional time series can arise when a continuous-time record is separated into natural consecutive time intervals. Examples include daily concentration curves of particulate matter with an aerodynamic diameter of less than 10 μm [e.g., 35] and monthly sea surface temperature in the "Niño region" [e.g., 64]. Alternatively, functional time series can arise when observations that are continuous functions in nature are repeatedly sampled in a period. For example, Figure 1a displays smoother near-infrared (NIR) spectra (see supplementary material for details of smoothing) recorded in monitoring glue curing process of wood panels in 72 experimental trials. The curves in the plot are ordered chronologically according to the colors of the rainbow [41].

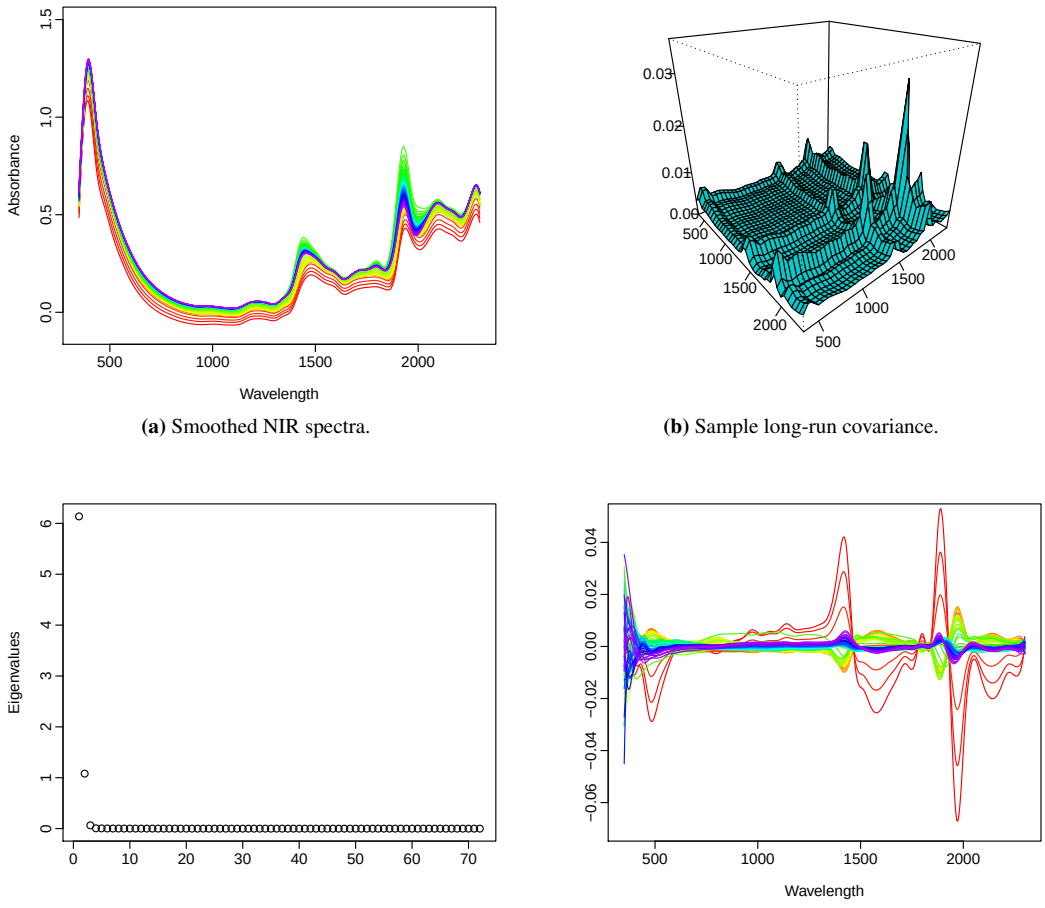
Functional data analysis (FDA) as a branch of statistics research has been popularized in the past two decades. Many recent advances in FDA draw much attention of the scientific community, for example classification of two-dimensional functional observations according to their shapes [see for instance 13], solving linear regressions when multivariate random functions are used as either predictors or responses or both [see e.g., 21, 57], detection of outlier curves based on function shapes and data depth [46], among many other interesting topics. We refer the audience to [30] for a detailed survey of recent trends in FDA, to [4] for a review of cutting-edge methods in FDA and high-dimensional

*Corresponding author. Email address: yang.yang3@monash.edu

statistics, and to [5] for most recent breakthroughs in functional statistics. Applications of most FDA methods, including those mentioned above, require accurate extractions of the main features of functions. However, many existing feature extraction methods such as approaches selecting impact points along trajectories [see,e.g., 6, 12, 22, 56, 57] concentrate on identifying a small subset of functions that capture the most important features, not paying enough attention to local or weak features of the data. This paper proposes a novel feature extraction method effectively capturing the most relevant information from a functional data set.

Most existing functional time series modeling methods [see, e.g., 9, 14, 15, 40, 43–45, 49] rely on functional principal component analysis (FPCA) to project the intrinsically infinite-dimensional functional objects onto directions of a small number of leading functional principal components. FPCA extracts only the dominant modes of variation of a functional object over its entire domain, with captured information referred to as the "main features" of the considered process. However, the "minor" components neglected by FPCA often have highly localized features possessing information on functional variations over particular short intervals within the function domain. A relatively recent dynamic FPCA introduced by [35] employs long-run covariance to include serial dependence of the data, but suffers the same problem of loss of local features in dimension reduction.

Attempts to capture local features of functional data include either restricting function domain [see, e.g., 27, 32]



(a) Smoothed NIR spectra.

(b) Sample long-run covariance.

(c) Eigenanalysis results.

(d) Dynamic FPCA residuals.

Fig. 1: NIR absorption spectra of wood panels and residual functions after extraction of the first two dynamic functional components associated with largest empirical eigenvalues of sample long-run covariance function. Using rainbow plots, curves from the distant past are shown in red, and the most recent curves are in violet.

or introducing sparseness penalty parameters [see, e.g., 2, 39] during dimension reduction. However, truncating function domains to specific intervals to enhance local feature extraction requires well aligned curves with most local features occurring in the same region. Thus, truncating methods are not suitable for analysis of NIR spectroscopy data that generally focus on identifying non-overlapping absorption spikes in observed spectra. In contrast, sparse FPCA methods impose sparsity penalties in regularized eigendecomposition to identify basis functions with local features. However, a single penalty parameter in practice is not sufficient to accommodate for local features of various magnitudes at different scales. As a result, solving optimization problems to identify the optimal penalty parameter can be tricky: a small penalty results in a significant amount of observation noise falsely identified as local features, while a large penalty fails to preserve peak heights of high-magnitude local features.

The problem of FPCA inadequately extracting local features is illustrated in Figure 1. Sample long-run covariance function of the observed NIR spectra shown in Figure 1b contains many local features. As suggested by Figure 1c, performing eigenanalysis on the sample long-run covariance function reveals that the first two leading dynamic functional components explain most functional variation of smoothed NIR spectra. The residual functions obtained by subtracting the empirical functional principal components from smoothed observations still contain sharp features around 1300 nm and 1900 nm of wavelength, as shown in Figure 1d. Thus, local features are important for the estimation of functional time series, typically in the study of NIR spectroscopy data that possess multiple significant local features.

Based on molecular overtones and combination vibrations of the investigated molecule, NIR spectroscopy generates complex absorption spectra over a region of the electromagnetic spectroscopy. Since many chemical compounds are known to have characteristic absorption bands over certain spectrum regions between 780–2500 nm, to determine composition materials of an object requires studying particular wavelength ranges (i.e., narrow bands with extreme absorption intensity) of the observed NIR spectrum together instead of examining absorptions one frequency at a time [16]. Thus, a computational method that can extract "local features" covering multiple frequencies of absorption spectrum is important for NIR spectroscopy analysis in practice. Use the wood panel NIR spectrum illustrated in Figure 1 above as an example. The observed local features between 350–2300 nm linked to composition materials, namely, the wood substrate, curing resin, and moisture content [19]. Subtle changes in experimental conditions such as temperature and pressure lead to variations of absorption bands over a series of trials. Hence, extracting and modeling local features are essential for monitoring the glue curing process of wood panels. Moreover, local features inheriting serial dependence of the original NIR curves can be used to make forecasts for future experiments. In this paper, we aim at developing a methodology for recovering local features that are ignored by FPCA and for using these extracted local features to make more accurate estimations and forecasts for functional time series.

Unlike the feature extraction methods mentioned above, [42] considered extracting principal components of high-dimensional data in wavelet domains. The wavelet bases are considered to be natural for uncovering sparse local features in the signal for the following four reasons. First, wavelet transform is a spatially varying decomposition that adapts its effective "window width" to magnitudes of local oscillations in FPCA residual functions. As a result, wavelet-based algorithms can accurately estimate local features at various scales. Second, orthonormal bases of compactly supported wavelets are particularly good at estimating sharp, highly localized features. This character of wavelet transform allows effective detection of local features associated with chemicals that have very narrow absorption bands (i.e., short intervals of wavelength frequencies) but high intensities (i.e., large absorbance coefficients) in NIR spectroscopy data [16]. Third, the wavelet transform is computationally efficient. For a given orthonormal wavelet basis, feature extraction can be completed in one step of matrix multiplication known as the "discrete wavelet transform" (for further detail on discrete wavelet transform, see [23, 66]). Fourth and the most important, many types of functional forms encountered in practice, including NIR absorption spectrum, can be sparsely and uniquely represented by a series of wavelet coefficients. Thus, wavelet transform allows a parsimonious representation of local features using only a relatively small number of estimated coefficients.

We propose a two-step algorithm that captures global and local features of functional time series sequentially. Initially, dynamic FPCA is applied to extract global features from the smoothed functional time series. Residuals of dynamic FPCA are then transformed into wavelet domains and block thresholding of wavelet (BTW) coefficients are conducted. Advantages of the FPCA-BTW method over sparse FPCA methods in relation to local feature extraction are demonstrated using simulated data in Section 5.1, and via an empirical application in Section 6. It should be noticed that neither conducting the BTW alone, or conducting the BTW before dynamic FPCA, would effectively capture most global and local features of functional time series in a parsimonious set of estimated wavelet coefficients: First, wavelet

approximations requires a fairly large number of coefficients (e.g., 2^{11} for the wood panel spectra, and the number of coefficients would increase if more spectrum frequencies are considered) to summarize all global and local features of a continuous function consisting of non-zero signals over its entire domain. More details of wavelet approximations will be presented in Section 2.3 later. Second, implementing BTW leads to a trade-off between preserving the overall smoothness and attaining to fine details of the true signal [see, page 942, Figure 1 in 8, for a depiction of this trade-off]. As a result, in practice many local features need to be sacrificed to minimize estimation errors measured by an L^2 norm for functional time series. In contrast, after conducting FPCA in the initial step of our proposed FPCA-BTW method isolates significant local features in the format of sparse "spikes" over short segments of a function that contains no signal but noise elsewhere. Then, performing the BTW in the second step yields only a small number of non-zero estimated wavelet coefficients containing information on local features as the thresholding algorithm reduces the remaining least important coefficients to zero.

To the best of our knowledge, there is no precedent research focusing on improving FPCA estimation performance via adequately extracting local features contained in "minor" functional components. The principal orthogonal complement thresholding method of [26] for the estimation of a high dimensional covariance with a conditional sparsity structure is closely analogous to our work as both methods attempt to produce improved estimation performance for processes consisting of finite common global features and sparse local features.

The rest of the paper is organized as follows. In Section 2, we provide necessary background on FPCA and wavelet approximation, before introducing the FPCA-BTW feature extraction method. Implementation details of the proposed method in estimation and forecasting of functional time series are given in Section 3. Section 4 presents asymptotic properties of FPCA-BTW estimators. In Section 5, we use Monte Carlo simulations to illustrate finite sample performances of FPCA-BTW estimators regarding estimation and forecasting of functional time series. Section 6 presents real data applications on NIR spectroscopy data of wood panels. Finally, Section 7 concludes the paper and provides some discussion and directions for future research.

2. Methodology

2.1. Notations

We start by fixing the notations used in this paper. Let $\{\mathcal{X}_t(u)\}_{t\in\mathbb{Z}}$ denote random functions defined on some common probability space $(\Omega, \mathcal{A}, P)$. Observations $\{\mathcal{X}_t(u)\}_{t\in\mathbb{Z}}$ are elements of the Hilbert space $H = L^2([0,1])$ equipped with the inner product $\langle x, y\rangle = \int_0^1 x(u)y(u)du$. Each $\mathcal{X}_t$ is a square integrable function satisfying $\|\mathcal{X}_t\|^2 = \int_0^1 \mathcal{X}_t^2(u)du < \infty$, where the standard norm on $L^2([0,1])$ is defined as $\|x\| = \langle x, x\rangle^{1/2}$. Define a notation $\mathcal{X} \in L_H^p(\Omega, \mathcal{A}, P)$ such that, for some $p > 0$, $\mathrm{E}\,\|\mathcal{X}\|^p < \infty$.

We consider functional time series $\{\mathcal{X}_t(u)\}_{t\in\mathbb{Z}}$ with a general representation given by

$$\mathcal{X}_t(u) = \mu(u) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(u) + Z_t(u) + \varepsilon_t(u), \qquad u \in [0,1], \tag{1}$$

where $\mu(u) = \mathrm{E}[\mathcal{X}(u)]$ is the mean function; $\{\phi_k(u)\}_{k=1}^K$ are real-valued orthogonal functions with K a fixed positive integer; a set of pairwise uncorrelated real numbers $\{\beta_{t,k}\}_{k=1}^K = \{\beta_{t,1}, \ldots, \beta_{t,K}\}$ satisfy that $\mathrm{var}(\beta_{t,i}, \beta_{t,j}) = 0$ for any $i \neq j$; $\{Z_t(u)\}_{t\in\mathbb{Z}}$ is a set of functions uncorrelated with $\{\phi_k(u)\}_{k=1}^K$; $\{\varepsilon_t(u)\}_{t\in\mathbb{Z}}$ is H-white noise with $\mathrm{E}\,\{\varepsilon_t(u)\} = 0$. (See Chapter 3 of [14] for further detail about strong white noise function in Hilbert space.) The $\sum_{k=1}^K \beta_{t,k}\phi_k(u)$ in (1) containing dominant modes of variation of $\{\mathcal{X}_t(u)\}_{t\in\mathbb{Z}}$ are referred to as "global features", whereas $\{Z_t(u)\}_{t\in\mathbb{Z}}$ with sparse localized spikes over the function domain $[0,1]$ are referred to as "local features". We assume that all eigenvalues of long-run covariance of local features are bounded, and the first K eigenvalues of long-run covariance function of global features decrease at the rate of $O(1)$. Extraction of global features and local features from functional time series $\{\mathcal{X}_t(u)\}_{t\in\mathbb{Z}}$ are introduced in Sections 2.2 and 2.3, respectively.

2.2. Extraction of global features

A weakly stationary functional time series $\{\mathcal{X}_t(u)\}_{t\in\mathbb{Z}}$ satisfies that, for all $t \in \mathbb{Z}$, (a) $\mathcal{X}_t(u) \in L^2([0,1])$, (b) $\mathrm{E}[\mathcal{X}_t(u)] = \mathrm{E}[\mathcal{X}_0(u)] = \mu(u)$, and (c) for all $\ell \in \mathbb{Z}$ and $u, s \in [0,1]$,

$$c_\ell(u,s) = \mathrm{cov}[\mathcal{X}_0(u), \mathcal{X}_\ell(s)] = \mathrm{cov}[\mathcal{X}_{t+\ell}(u), \mathcal{X}_t(s)], \tag{2}$$

with $\mathrm{cov}[\mathcal{X}(u), \mathcal{X}(s)] = \mathrm{E}[\{\mathcal{X}(u) - \mu(u)\}\{\mathcal{X}(s) - \mu(s)\}]$. C_ℓ induces an operator $L^2([0,1]) \to L^2([0,1])$ given by

$$C_\ell(x)(u) = \int_0^1 C_\ell(u,s)x(s)\,\mathrm{d}s, \qquad x \in L^2([0,1]), \quad u, s \in [0,1].$$

When $\ell = 0$, the autocovariance operator C_ℓ has a special case of covariance operator C_0 defined by $C_0(x) = \int_0^1 C_0(u,s)x(s)\,\mathrm{d}s$ for $x \in L^2([0,1])$ and $u, s \in [0,1]$.

In practice, $\{\mathcal{X}_t(u)\}_{t\in\mathbb{Z}}$ often consists of serially correlated observed trajectories. To incorporate serial dependence carried by lagged observations, recent studies [see, e.g., 62, 63] suggest computing a long-run covariance function $C(u,s)$ as

$$C(u,s) = \sum_{\ell=-\infty}^{\infty} c_\ell(u,s), \qquad u, s \in [0,1].$$

A long-run covariance operator C is then defined as

$$C(x)(u) = \int_0^1 C(u,s)x(s)\,\mathrm{d}s \qquad x \in L^2([0,1]), \quad u, s \in [0,1].$$

The symmetric positive-definite Hilbert–Schmidt operator C admits a decomposition as

$$C(x) = \sum_{k=1}^{\infty} \lambda_k \langle \phi_k\,, x\rangle \phi_k, \qquad x \in L^2([0,1]),$$

where $\{\lambda_k\}_{k\in\mathbb{Z}^+}$ are the nonincreasing eigenvalues, and $\{\phi_k\}_{k\in\mathbb{Z}^+}$ the corresponding orthonormal eigenfunctions such that $C(\phi_k) = \lambda_k\phi_k$, and $\langle \phi_i\,, \phi_j\rangle = 1$ iff $i = j$. The Karhunen–Loève expansion of a stochastic process $\mathcal{X}_t(u)$ is then given by

$$\mathcal{X}_t(u) = \mu(u) + \sum_{k=1}^{\infty} \beta_{t,k}\phi_k(u),$$

where the kth functional component score $\beta_{t,k}$ is a projection of $\overline{\mathcal{X}}_t(u) = \mathcal{X}_t(u) - \mu(u)$ in the direction of the kth eigenfunction $\phi_k(u)$, that is, $\beta_k = \langle \overline{\mathcal{X}}_t\,, \phi_k\rangle$.

According to (1), the main features of the infinite-dimensional $\{\mathcal{X}_t(u)\}_{t\in\mathbb{Z}}$ can be summarized by its first K leading components as

$$\mathcal{X}_t(u) = \mu(u) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(u) + e_t(u),$$

where $\{e_t(u)\}_{t\in\mathbb{Z}}$ are error functions after truncation. According to Theorem 2 of [35], the linear combination of $\sum_{k=1}^{K} \beta_{t,k}\phi_k(u)$ obtained by dynamic FPCA satisfies that, for any other orthonormal basis $\{\varphi_k\}_{k\in\mathbb{Z}^+}$ of the Hilbert space H,

$$\mathrm{E}\left[\left\|\overline{\mathcal{X}}_t - \sum_{k=1}^{K} \beta_{t,k}\phi_k\right\|^2\right] \le \mathrm{E}\left[\left\|\overline{\mathcal{X}}_t - \sum_{k=1}^{K} \langle \mathcal{X}_t\,, \varphi_k\rangle\varphi_k\right\|^2\right]. \tag{3}$$

In rare cases, functional time series $\{\mathcal{X}_t(u)\}_{t\in\mathbb{Z}}$ may possess weak serial dependence. The significance of serial dependence can be determined according to the hypothesis test of [38]. Functional observations are treated as independent if c_ℓ of (2) at all lags apart from $\ell = 0$ are tested to be negligible. A process that decomposes the covariance operator C_0 to extract global features is often referred to as static FPCA to distinguish it from dynamic FPCA. In the remaining of this paper, we present feature extraction results obtained by dynamic FPCA and include feature extraction results associated with static FPCA in supplementary material. The aim of this paper is to demonstrate the proposed local feature extraction method can be applied to improve performances of static FPCA and dynamic FPCA, instead of comparing performances of the two versions of FPCA.

It can be seen from (3) that dynamic FPCA can find an optimal representation of global features of $\{\mathcal{X}_t(u)\}_{t\in\mathbb{Z}}$, but ignores most local features $\{Z_t(u)\}_{t\in\mathbb{Z}}$. Our proposed two-step feature extraction method will continue to capture any

remaining local features from FPCA residuals, as described in Section 2.3.

2.3. Extraction of local features

To extract sharp and highly localized features from FPCA residuals, we consider an orthonormal system of wavelet functions. Wavelet functions combine compact support with various degrees of smoothness, which enables the extraction of signals at a variety of different scales. It has been tested that wavelets can effectively isolate signals from noisy functions in statistical applications [see, e.g., 7, 58]. Most recent wavelet applications in statistics adopt the approach of [23] to define two related and specially selected orthonormal parent wavelet functions: the scaling function ψ and the mother wavelet Ψ. Wavelets can then be generated by dilation and translation as

$$\psi_{j,p} = 2^{j/2}\psi(2^j t - p), \quad \Psi_{j,p} = 2^{j/2}\Psi(2^j t - p), \qquad (j \in \mathbb{Z}^+, \quad p = 1, \ldots, 2^j),$$

where the index j represents resolution level in wavelet decomposition. This wavelet system produces wavelet functions forming an orthonormal wavelet basis in $L^2([0,1])$. With a primary decomposition level $j_0 \geq 0$, local features $\{Z_t(u)\}_{t\in\mathbb{Z}}$ admit a decomposition given by

$$Z_t(u) = \sum_{p=1}^{2^{j_0}} D'_{j_0,p,t}\psi_{j_0,p}(u) + \sum_{j=j_0}^{\infty}\sum_{p=1}^{2^j} D_{j,p,t}\Psi_{j,p}(u), \tag{4}$$

where wavelet coefficients are defined as

$$D'_{j_0,p,t} = \int_0^1 Z_t(u)\psi_{j_0,p}(u)\,\mathrm{d}u, \quad D_{j,p,t} = \int_0^1 Z_t(u)\Psi_{j,p}(u)\,\mathrm{d}u.$$

"Approximations" and "details" of $Z_t(u)$ are stored in wavelet coefficients $D'_{j_0,p,t}$ and $D_{j,p,t}$, respectively [52].

According to (1), residual functions $\{e_t(u)\}_{t\in\mathbb{Z}}$ consist of highly localized features $Z_t(u)$ and random noise $\varepsilon_t(u)$ given by

$$e_t(u) = Z_t(u) + \varepsilon_t(u).$$

The wavelet transform of $e_t(u)$ can be expressed as

$$e_t(u) = \sum_{p=1}^{2^{j_0}} \widetilde{D}'_{j_0,p,t}\psi_{j_0,p}(u) + \sum_{j=j_0}^{\infty}\sum_{p=1}^{2^j} \widetilde{D}_{j,p,t}\Psi_{j,p}(u),$$

where the empirical wavelet coefficients $\widetilde{D}'_{j_0,p,t}$ and $\widetilde{D}_{j,p,t}$ are given by

$$\widetilde{D}'_{j_0,p,t} = \int_0^1 e_t(u)\psi_{j_0,p}(u)\,\mathrm{d}u, \quad \widetilde{D}_{j,p,t} = \int_0^1 e_t(u)\Psi_{j,p}(u)\,\mathrm{d}u.$$

Wavelet coefficients related to detailed structure of $e_t(u)$ and $Z_t(u)$ thus satisfy that, for any $t \in \mathbb{Z}$,

$$\widetilde{D}_{j,p,t} = D_{j,p,t} + \epsilon_{j,p,t}, \qquad (j \in \mathbb{Z}^+, \quad p = 1, \ldots, 2^j), \tag{5}$$

where $\epsilon_{j,p,t} = \int_0^1 \varepsilon_t(u)\Psi_{j,p}(u)\,\mathrm{d}u$ represents a wavelet transform of contamination noise. Since local features $\{Z_t(u)\}_{t\in\mathbb{Z}}$ are sparse, a vector of wavelet coefficients $D_t = \{D_{j_0,1,t}, \ldots, D_{j_0,2^{j_0},t}, \ldots\}$ contains many zeros. Extracting local features is then equivalent to determining non-zero wavelet coefficients $D_{j,p,t}$. From a statistical modeling perspective, the denoising problem of (5) has been commonly approached by shrinking the empirical wavelet coefficients $\{\widetilde{D}_{j,p,t}\}_{j\in\mathbb{Z}^+,p=1,\ldots,2^j}$ one by one [see, e.g., 8, 25]. However, local features of functional data often occur over short intervals within the function domain that correspond to several consecutive wavelet coefficients at fine resolution levels. To determine chemical content of an object by NIR spectroscopy, simultaneously considering the non-zero wavelet coefficients corresponding to certain distinctive absorption bands of known chemical compounds provides more accurate composition results than examining absorption value at any single frequency. For example, local features

depicting extreme absorption bands of approximately 1900 nm, shown in Figure 1a, are summarized into 21 consecutive empirical wavelet coefficients at the resolution level $j = 11$. Thus, to enhance extraction of local features, adjacent wavelet coefficients should be modeled together as a group. For this purpose, we adopt a block thresholding approach of [17] to make simultaneous selection of empirical wavelet coefficients in groups as follows. At each resolution level j, divide the empirical wavelet coefficients $\widetilde{D}_{j,p,t}$ into non-overlapping blocks of length L. Denote indices of the coefficients in the ath block at level j by j_a, i.e., $j_a = \{(j,p) : (a-1)L + 1 \le p \le aL\}$. Let $S^2_{j_a} = \sum_{p\in j_a} \widetilde{D}^2_{j,p,t}$ denote the sum of squares of the empirical coefficients in the block. A block is significant if its $S^2_{j_a}$ is larger than a threshold $T_w = \lambda^\star L\sigma^2/2^J$, where $\lambda^\star$ is a threshold constant and σ is the noise level. Retaining significant wavelet coefficients while discarding the remaining negligible coefficients leads to a local feature estimator as

$$\widehat{Z}_t(u) = \sum_{k=0}^{2^{j_0}} \widetilde{D}'_{j_0,p,t}\psi_{j_0,p}(u) + \sum_{j=j_0}^{J-1}\sum_{a}\left(\sum_{p\in j_a}\widetilde{D}_{j,p,t}\Psi_{j,p}(u)\mathbb{1}(S^2_{j_a} > T_w)\right), \tag{6}$$

where $\mathbb{1}(\cdot)$ represents the binary indicator function. It can be seen that a depends on j and p, while p depends on j as well. Since a varies for different resolution levels, we omit the upper limit of $\sum_a$ in (6).

In (6), the block length L and the threshold constant T_w together control global and local adaptivity of the estimator $\widehat{Z}_t(u)$. A global adaptive estimator adjusts to the overall regularity of the target function, and a locally adaptive estimator focuses on optimally adapting to subtle and highly localized features along the curve. The optimal selection of parameters L and T_w, together with other implementation details about the FPCA-BTW feature extraction method, are described in Section 3.

3. Implementation details

3.1. Long-run covariance estimation

We first present technical details of extracting global features of a finite sample functional time series. To consider serial dependence of stationary functional observations $\{\mathcal{X}_t(u)\}_{t=1}^T$, we compute the empirical long-run covariance function as

$$\widehat{C}_{h,q}(u,s) = \sum_{\ell=-T}^{T} W_q\left(\frac{\ell}{h}\right)\widehat{c}_\ell(u,s), \tag{7}$$

where W_q is a symmetric weight function with bounded support of order q, and h is a bandwidth parameter; the estimator of $c_\ell(u,s)$ is defined in the form of

$$\widehat{c}_\ell(u,s) = \begin{cases} \frac{1}{T}\sum_{j=1}^{T-\ell}\left[\mathcal{X}_j(u) - \widehat{\mu}(u)\right]\left[\mathcal{X}_{j+\ell}(s) - \widehat{\mu}(s)\right], & \ell \ge 0; \\ \frac{1}{T}\sum_{j=1-\ell}^{T}\left[\mathcal{X}_j(u) - \widehat{\mu}(u)\right]\left[\mathcal{X}_{j+\ell}(s) - \widehat{\mu}(s)\right], & \ell < 0. \end{cases}$$

The optimal bandwidth parameter h is selected via the "plug-in" algorithm proposed in [62]. More details about estimating the corresponding $\widehat{C}_{\widehat{h}_{\text{opt}}}(u,s)$ are provided in supplementary material S2.1. The empirical long-run covariance operator is then given by

$$\widehat{C}(x)(u) = \int_0^1 \widehat{C}_{\widehat{h}_{\text{opt}}}(u,s)x(s)\,\mathrm{d}s, \qquad x \in L^2([0,1]).$$

Performing eigendecomposition on the empirical long-run covariance operator yields

$$\widehat{C}(x) = \sum_{k=1}^{\infty}\widehat{\lambda}_k\langle\widehat{\phi}_k, x\rangle\widehat{\phi}_k, \qquad x \in L^2([0,1]),$$

where $\{\widehat{\phi}_k\}_{k\in\mathbb{Z}^+}$ are the empirical eigenfunctions, and $\{\widehat{\lambda}_k\}_{k\in\mathbb{Z}^+}$ are associated eigenvalues. To facilitate dimension reduction, the dimension of global features $\widehat{K}$ need to be empirically determined. Existing functional time series methods generally select $\widehat{K}$ by requiring that retained functional components should explain a certain level of the total variance, approximately 85% [see, e.g., 20, 35, 63]. However, this criterion of cumulative percentage of explained

variation has the disadvantage of incorrectly selecting too many components as global features when fast-diverging eigenvalues are present in FPCA analysis. To precisely extract global features, following [49], the value of K is determined as the integer minimizing ratios of two adjacent empirical eigenvalues given by

$$\widehat{K} = \underset{1\le k\le k_{max}}{\operatorname{argmin}} \left\{ \frac{\widehat{\lambda}_{k+1}}{\widehat{\lambda}_k} \times \mathbb{1}\left(\frac{\widehat{\lambda}_k}{\widehat{\lambda}_1} \ge \tau\right) + \mathbb{1}\left(\frac{\widehat{\lambda}_k}{\widehat{\lambda}_1} < \tau\right) \right\}, \tag{8}$$

where k_{max} is a prespecified positive integer, τ is a prespecified small positive number, and $\mathbb{1}(\cdot)$ is the indicator function. When without priori information about a possible maximum of K, it is unproblematic to choose a relatively large k_{max}, e.g., $k_{max} = \#\{k|\widehat{\lambda}_k \ge \sum_{k=1}^{T} \widehat{\lambda}_k/T, k \ge 1\}$ [1]. Given that the small empirical eigenvalues $\widehat{\lambda}_k$ for some $K < k < k_{max}$ are likely to be practically zero, we adopt the threshold constant $\tau = 1/\ln(\max\{\widehat{\lambda}_1, T\})$ to ensure consistency of $\widehat{K}$.

As described in Section 2.2, it is possible to have nearly independent $\{\mathcal{X}_t(u)\}_{t=1}^{T}$ in practice. The sample covariance operator of independent observations is computed as

$$\widehat{C}_0(x)(u) = \int_0^1 \widehat{c}_0(u,s)x(s)\,\mathrm{d}s, \qquad x \in L^2([0,1]),$$

where $\widehat{c}_0(u,s) = \frac{1}{T}\sum_{t=1}^{T}[\mathcal{X}_t(u) - \widehat{\mu}(u)][\mathcal{X}_t(s) - \widehat{\mu}(s)]$ is the empirical covariance kernel with the empirical mean function $\widehat{\mu}(u) = \frac{1}{T}\sum_{t=1}^{T}\mathcal{X}_t(u)$. For such data static FPCA is applied to extract global features from $\widehat{C}_0(x)(u)$ using the same criterion as (8).

With FPCA results, functional time series can be estimated by

$$\widehat{\mathcal{X}}_t(u) = \widehat{\mu}(u) + \sum_{k=1}^{\widehat{K}} \widehat{\beta}_{t,k}\widehat{\phi}_k(u),$$

where $\sum_{k=1}^{K}\widehat{\beta}_{t,k}\widehat{\phi}_k(u)$ represents the extracted global features, with the empirical principal component scores defined by $\widehat{\beta}_{t,k} = \int_0^1 [\mathcal{X}_t(u) - \widehat{\mu}(u)]\,\widehat{\phi}_k(u)\,\mathrm{d}u$. Removing the estimated mean function and the extracted global features from functional observations leaves residual functions given by

$$\widehat{e}_t(u) = \mathcal{X}_t(u) - \widehat{\mu}(u) - \sum_{k=1}^{\widehat{K}} \widehat{\beta}_{t,k}\widehat{\phi}_k(u).$$

In Section 3.2, we present details of recovering local features from $\{\widehat{e}_t(u)\}_{t=1}^{T}$ through block thresholding of wavelet coefficients.

3.2. Estimation of wavelet coefficients

The continuous wavelet transform formalized by [31] can be implemented in computer software such as R [61] to extract local features from FPCA residuals $\{\widehat{e}_t(u)\}_{t=1}^{T}$. However, in practice, we most likely only observe discretized values $\{\mathcal{X}_t(u_i)\}_{i=1}^{n_t}$, with n_t denoting the number of grid points in the tth curve. Removing global features evaluated at each grid point leaves discrete residuals $\{\widehat{e}_t(u_i)\}_{i=1}^{n_t}$. When equally spaced grids satisfy $n_t = 2^J$ for $t = 1,\ldots,T$, wavelet transform of $\{\widehat{e}_t(u)\}_{t=1}^{T}$ can be performed in $O(2^J)$ operations [51].

In situations when functional observations have nondyadic, varying or unequally spaced grid points, the non-linear regularized Sobolev interpolator of [8] is adopted to perform the wavelet transform. Local feature extraction can then be completed in the following steps. First, select an orthonormal wavelet family to obtain an orthogonal DWT base matrix W with dimension $N \times N$, where $N = 2^J \ge \max(n_1,\ldots,n_T)$ is a dyadic integer. There are many discrete wavelet families available in the literature. We follow [68] and consider the Daubechies least asymmetric wavelets with 10 vanishing moments in the analysis of NIR spectroscopy data. Denote A as a matrix of dimension $n_t \times N$ whose ith row corresponds to the row of the matrix W^{T}. We interpolate the vector $\widehat{e}_t = [\widehat{e}_t(u_1),\ldots,\widehat{e}_t(u_{n_t})]^{\mathrm{T}}$ as

$$\widetilde{D}_t = A^{\mathrm{T}}\widehat{e}_t, \tag{9}$$

where $\widetilde{D}_t = [\widetilde{D}'_{j_0,1,t}, \ldots, \widetilde{D}'_{j_0,2^{j_0},t}, \widetilde{D}_{j_0,1,t}, \ldots, \widetilde{D}_{j_0,2^{j_0},t}, \ldots, \widetilde{D}_{J-1,1,t}, \ldots, \widetilde{D}_{J-1,2^{J-1},t}]^{\mathrm{T}}$ is a vector of size N [8]. The optimal parameters for block thresholding are then selected according to [17]. Specifically, for the block size L and the threshold constant $\lambda^\star$ in (6), $L = 2^{\lfloor \log_2(\ln(2^J)) \rfloor}$ and $\lambda^\star = 4.5052$ are chosen. Noise level of residual functions are estimated by taking the median absolute deviation (MAD) as

$$\widehat{\sigma} = \frac{\mathrm{MAD}\{\widetilde{D}_{J-1,p,t}/v_{J-1,p,t}^{1/2} : v_{J-1,p,t} > 0.0001\}}{0.6745},$$

where $\{\widetilde{D}_{J-1,p,t}\}_{p=1,\ldots,2^{J-1}}$ are the empirical wavelet coefficients at the resolution level $J-1$, and $\{v_{J-1,p,t}\}_{p=1,\ldots,2^{J-1}}$ are diagonal elements of the matrix $V = A^{\mathrm{T}}A$ [8]. Next, the first-round block thresholding is implemented according to (6), and intermediate results are denoted as $\widetilde{D}_t^*$. Subtracting the inverse transform of $\widetilde{D}_t^*$ from discrete residuals gives

$$\widehat{e}_t^* = \widehat{e}_t - A\widetilde{D}_t^*.$$

The second round empirical wavelet coefficients are then computed as

$$\widetilde{D}_t^\dagger = \widetilde{D}_t^* + A^{\mathrm{T}}\widehat{e}_t^*.$$

Finally, performing block thresholding again on $\widetilde{D}_t^\dagger$ yields the final BTW coefficients $\widehat{D}_t = [\widehat{D}'_{j_0,1,t}, \ldots, \widehat{D}'_{j_0,2^{j_0},t}, \widehat{D}_{j_0,1,t}, \ldots, \widehat{D}_{j_0,2^{j_0},t}, \ldots, \widehat{D}_{J-1,1,t}, \ldots, \widehat{D}_{J-1,2^{J-1},t}]^{\mathrm{T}}$ with many zero entries reflecting the sparseness of the local features. Note that we keep the "approximation" wavelet coefficients unchanged as $\widehat{D}'_{j_0,p,t} = \widetilde{D}'_{j_0,p,t}$ for all $p = 1, \ldots, 2^{j_0}$. According to [65], implementing the estimation method of [8] through a two-round block thresholding process simplifies computation. Applying the above procedure to each discrete residual function $\widehat{e}_t$ leads to a sparse $N \times T$ matrix of BTW coefficients $\widehat{D} = [\widehat{D}_1, \ldots, \widehat{D}_T]$. The extracted local features are then given by a product $A\widehat{D}$.

Using the extracted global and local features, we can make improved estimation of the considered functional process and its covariance structure, and produce more accurate forecasts. We demonstrate applications of the proposed feature extraction method using simulated samples in Section 5 and real NIR spectroscopy data in Section 6. Additional technical details about long-run covariance estimation and applications of the FPCA-BTW method are provided in supplementary material.

4. Asymptotic properties

Before presenting assumptions and asymptotic results of long-run covariance based FPCA-BTW estimators, we introduce some notations. Let $\mathcal{L} = \mathcal{L}(H, H)$ be the space of bounded linear operators from H to H. We define the operator norm $\|A\|_{\mathcal{L}} = \sup_{\|x\|\leq 1} \|A(x)\|$ for $A \in \mathcal{L}$. The operator A is compact if there exists two orthonormal bases $\{v_k\}$ and $\{v_k\}$, and a real sequence $\{\lambda_k\}$ converging to zero, such that

$$A(x) = \sum_{k=1}^{\infty} \lambda_k \langle x, v_k \rangle v_k, \qquad x \in H.$$

A compact operator is said to be a Hilbert–Schmidt operator if $\sum_{k=1}^{\infty} \lambda_k^2 < \infty$. We denote the Hilbert–Schmidt norm by $\|A\|_{\mathcal{S}}$. For any Hilbert–Schmidt operator A, one can show that $\|A\|_{\mathcal{S}}^2 = \sum_{k\geq 1} \lambda_k^2$ and $\|A\|_{\mathcal{L}} \leq \|A\|_{\mathcal{S}}$ [37, Chapter 2].

Assumption 1. *Functions $\{\mathcal{X}_t(u), u \in [0,1]\}_{t\in\mathbb{Z}}$ are L^4 - m-approximable, taking values in $L^2([0,1])$, satisfying the following conditions:*

1. *X_t admits the representation $\mathcal{X}_t = f(\delta_t, \delta_{t-1}, \delta_{t-2} \ldots, \delta_{t-m+1}, \delta_{t-m}, \delta_{t-m-1}, \ldots)$ with δ_i i.i.d. elements taking values in a measurable space S and a measurable function $f : \mathcal{S}^\infty \to H$.*
2. *$E\|\mathcal{X}_0\|^{4+d} < \infty$ for some $d > 0$, and*
3. *$\{\mathcal{X}_t(u), u \in [0,1]\}_{t\in\mathbb{Z}}$ can be approximated by m-dependent sequences*

$$\mathcal{X}_t^{(m)} = f(\delta_t, \delta_{t-1}, \delta_{t-2}, \ldots, \delta_{t-m+1}, \delta_{t,t-m}^{(m)}, \delta_{t,t-m-1}^{(m)}, \ldots),$$

where $\{\delta_{t,i}^{(m)}\}$ *are independent copies of sequence* $\{\delta_t\}_{-\infty<t<\infty}$ *defined on the same measurable space* S *such that* $\sum_{m=1}^{\infty} \upsilon_4(\mathcal{X}_t - \mathcal{X}_t^{(m)}) < \infty$ *with* $\upsilon_4(\mathcal{X}_t - \mathcal{X}_t^{(m)}) = \left\{E\left\|\mathcal{X}_t - \mathcal{X}_t^{(m)}\right\|^4\right\}^{1/4}$.

Remark 1. Assumption 1 follows the dependence concept for functional time series introduced in [36]. This assumption is often considered as equivalent conditions to the classic mixing conditions in function spaces [see, e.g., 11, 38, 62]. Condition (3) specifies the level of dependence that is allowed within process $\{\mathcal{X}_t(u)\}_{t\in\mathbb{Z}}$ in relation to how well it can be approximated by finite m-dependent processes. Condition (3) can also be satisfied when $\upsilon_4(\mathcal{X}_t - \mathcal{X}_t^{(m)}) = O(m^{-\rho})$ for some $\rho > 4$. Roughly speaking, the $\mathcal{X}_t^{(m)}$ defined by the coupling construction in Condition (3) can be determined by the first m elements $\delta_t, \delta_{t-1}, \ldots, \delta_{t-m+1}$. When the measurable space S coincides with H, the sequence $\{\widetilde{\mathcal{X}}_t^{(m)}\}$ given by

$$\widetilde{\mathcal{X}}_t^{(m)} = f(\delta_t, \delta_{t-1}, \delta_{t-2}, \ldots, \delta_{t-m+1}, 0, 0, \ldots)$$

is also strictly stationary and m-dependent, satisfying $\sum_{m=1}^{\infty} \upsilon_4(\mathcal{X}_t - \widetilde{\mathcal{X}}_t^{(m)}) < \infty$.

Assumption 2. *The kernel function* $W_q(\cdot)$ *in* (7) *satisfies the following standard conditions:*

$$W_q(0) = 1, \quad W_q(u) \le 1, \quad W_q(u) = W_q(-u), \quad W_q(u) = 0 \text{ if } |u| > g \text{ for some constant } g > 0,$$

and $W_q(u)$ *is Lipschitz continuous on* $[-g, g]$.

There exists a $q > 0$ *such that*

$$0 < \lim_{u\to 0} \frac{W_q(u) - 1}{|u|^q} = \mathcal{W}_q < \infty,$$

and there exists $q' > q$ *such that*

$$\sum_{\ell=-\infty}^{\infty} |\ell|^{q'} \|c_\ell\| < \infty,$$

where c_ℓ *is the lag-*ℓ *autocovariance function defined in* (2).

Remark 2. Assumption 2 limits the growing rate of $W_q(u)$ at $u = 0$, with q referred to as the characteristic exponent of the kernel function by [59]. The smoother the kernel $W_q(u)$ at zero, the larger the value of q for which $\mathcal{W}_q$ is finite. This assumption has been widely adopted in studies on limit behaviors of the long-run covariance estimator [e.g., 11, 62].

The conditions in Assumptions 1 and 2 can be easily verified for most stationary time series models based on independent innovations (see supplementary material S1.3 for an example). To ensure the consistency of the long-run covariance estimator $\widehat{C}_{h,q}$ in (7), we impose the following condition on the bandwidth parameter h.

Assumption 3. *The bandwidth parameter* h *of long-run covariance estimator in* (7) *satisfies*

$$h = h(T) \to \infty \text{ and } \frac{h(T)}{T} \to 0, \text{ as } T \to \infty.$$

We use the optimal value of h *selected according to the plug-in bandwidth selection procedure of [62] in developing asymptotic results and conducting simulation and empirical studies in this paper.*

Remark 3. Assumption 3 is a weaker and more standard condition compared to the one used in the Theorem 4.2 of [36], that is, $h^2/T \to 0$. Details of the plug-in algorithm are provided in supplementary material S2.

Assumption 4. *The eigenvalues of the long-run covariance operator* C *are finite, positive, and distinctive, i.e.,* $\infty > \lambda_1 > \lambda_2 > \ldots$. *There exists a positive integer* K *such that*

$$\frac{\sum_{k=K+1}^{\infty} \lambda_k}{\sum_{k=1}^{K} \lambda_k} = o(1). \tag{10}$$

Remark 4. Distinctive eigenvalues of covariance operators are commonly adopted in the literature to ensure identification of eigenfunctions [see, e.g., 35, 36]. Assumption 4 requires that the sum of the "insignificant" eigenvalues $\{\lambda_{K+1}, \lambda_{K+2} \ldots\}$ tend to zero sufficiently rapidly. Thus, the K-dimensional global feature contains "most information" of $\mathcal{X}_t(u)$ [see e.g., 10, 34]. Roughly speaking, Assumption 4 requires that the first K eigenvalues $\{\lambda_1, \ldots \lambda_K\}$ have greater orders than the remaining eigenvalues in the sense of (10). For example, denoting $a/b \to 1$ by "$a \sim b$", [49] proposed that eigenvalues of long-run covariance function satisfying the conditions (a) $\lambda_k \sim \rho_k T^{3-2\alpha_k}$ for $k = 1, \ldots, K$ with coefficients $\infty > \rho_1 \geq \rho_2 \geq \ldots \rho_K > 0$ and $1/2 < \alpha_1 < \alpha_2 < \ldots < \alpha_K < 1$, and (b) $\sum_{k=K+1}^{\infty} \lambda_k = O(T)$. Given that $T^{3-2\alpha_1} > T^{3-2\alpha_2} > \ldots T^{3-2\alpha_K} > T$ for a fixed K, the sum of $\sum_{k=1}^{K} \lambda_k$ has an order of $T^{3-2\alpha_1}$. It can then be readily seen that $\sum_{k=K+1}^{\infty} \lambda_k / \sum_{k=1}^{K} \lambda_k = O(T)/T^{3-2\alpha_1} = o(1)$ as $T \to \infty$. Hence, Assumption 4 is satisfied with non-zero "insignificant" eigenvalues $\{\lambda_{K+1}, \lambda_{K+2} \ldots\}$. We are going to identify K and estimate the dynamic space $\mathcal{M}$ spanned by the (deterministic) eigenfunctions $\phi_1(u), \ldots, \phi_K(u)$.

Assumption 5. *The dynamic FPC scores $\{\beta_{t,k}\}$ are uncorrelated across k at all different lags, i.e., $cov(\beta_{t,i}, \beta_{t+h,j})$ with $i \neq j$, $i, j = 1, \ldots, K$, and $h \in \mathbb{Z}$.*

Remark 5. Assumption 5 specifies the uncorrelatedness of dynamic FPC scores, which is considered as one of the important properties of dynamic FPC scores [see, e.g., page 329, proposition 3(b) in 35].

Assumption 6. *The empirical eigenfunctions are in the same direction of the true eigenfunction, i.e., $\langle \phi_k, \widehat{\phi}_k \rangle > 0$.*

Remark 6. Under Assumption 4, the empirical eigenfunctions $\widehat{\phi}_k$ recovered are in the same direction, or in the opposite direction, with the true eigenfunction ϕ_k, i.e., $\text{sign}(\langle \widehat{\phi}_k, \phi_k \rangle) = \pm 1$. With Assumption 6, the derivations of equations and proofs are simplified. Note that Assumption 6 is optional for conducting the Karhunen–Loève expansion of a stochastic process $\mathcal{X}(u)$ given that $\langle \mathcal{X}, \phi \rangle \phi$ and $\langle \mathcal{X}, -\phi \rangle (-\phi)$ are identical.

Assumption 7. *Let n denote the number of observations on each curve. Let $N = 2^J \geq n$ be a dyadic integer. As $N, n \to \infty$, we assume that $(n \log^a n)^{-1} N$ tends to a constant for some $a > 0$. Let G_n be the empirical distribution function of the grid points $\{u_1, \ldots, u_n\}$. Suppose that there exists a distribution $G(u)$ with density $g(u)$, which is bounded away from 0 and infinity such that*

$$G_n(u) \to G(u) \text{ for all } t \in [0, 1] \text{ as } n \to \infty.$$

Further, $g(u)$ has the αth bounded derivative.

Remark 7. Assumption 7 specifies technical conditions ensuring the estimator of (9) is closely approximate to the true signal over the Besov space $B^{\alpha}_{P,Q}$ (see Appendix A.1). The same assumption was adopted in [8] in the development of their Theorem 6. Functional data $\{\mathcal{X}_t(u_1), \ldots, \mathcal{X}_t(u_n)\}$ measured at dense grids $\{u_1, \ldots, u_n\}$ can easily satisfy Assumption 7. Since the global feature extraction is conducted before the local feature extraction, selections of N and n have no impact on convergence of global feature estimators.

Proposition 1. *Under Assumptions 1 to 6, as $T \to \infty$ it happens that $\widehat{K}$ satisfies the following property*

$$Pr(\widehat{K} = K) \to 1,$$

where $\widehat{K}$ is determined by (8).

Remark 8. The estimation approach of (8) has one similarity with the "scree plot" method of [20]: the estimated dimension of functional principal component is chosen to be the point at which the ordered eigenvalues drop substantially. Similar decision rules are often used to estimate the number of factors for high-dimensional factor models; see [47, 48], and [1]. For functional time series with short memory, [10] adopted an estimator similar to (8) in analysis of the lagged autocovariance operator C_ℓ for $\ell \neq 0$ of the K-dimensional functions satisfying $\lambda_K = 0$ and $\lambda_{K+1} = 0$. Most recently, [49] used an estimator similar to (8) to identify the dimension of the dominant subspace in the long memory functional time series. We fill in the literature gap by using the estimator of (8) when estimating the dimension of long-run covariance operator for short memory functional time series.

We are now ready to present consistency properties of global and local feature estimators in the following theorems.

Theorem 1. *Denote the kth empirical functional component by $\widehat{\beta}_{t,k}$ and its associated score by $\widehat{\phi}_k$. Under Assumptions 1 to 6, as $T \to \infty$,*

$$\left\| \sum_{k=1}^{K} \beta_{t,k}\phi_k(u) - \sum_{k=1}^{\widehat{K}} \widehat{\beta}_{t,k}\widehat{\phi}_k(u) \right\| = O_P\left(T^{-2/5}\right).$$

Remark 9. The convergence rate of global feature estimators depends on the weight function W_q and the bandwidth h in (7). We use a flat-top weight function with quadratic spectral kernel (more details see supplementary material S2.1) that has been considered by [3], together with the optimal bandwidth selected according to the plug-in method of [62]. The order of $T^{-2/5}$ associated with the selected W_q matches findings of [60] when the optimal bandwidth is used.

Theorem 2. *Denote the number of dyadic points in wavelet transform by N. Under Assumptions 1 to 7, as $N, T \to \infty$, and for some $\alpha > 0$,*

$$\left\| Z_t(u) - \widetilde{Z}_t(u) \right\| = O_P(N^{-\alpha/(1+2\alpha)} + T^{-2/5}),$$

where $Z_t(u)$ and $\widetilde{Z}_t(u)$ are defined in (4) and (6), respectively.

Remark 10. Theorem 1 states the convergence rate for FPCA-based global feature estimators when the optimal bandwidth selected by the plug-in algorithm of [62] is used. Here, α indicates the degree of smoothness of the true signal of local features in a Besov ball $B^{\alpha}_{P,Q}$ (see 8 in Appendix A.1 for the definition of $B^{\alpha}_{P,Q}$). Loosely speaking, the true signal in the Besov space $B^{\alpha}_{P,Q}$ has α bounded derivatives in L^P space, with finer gradation of smoothness further controlled by the parameter Q [see, e.g., 53, for definitions and properties of Besov spaces]. Given that local features are estimated after the extraction of global features, convergence of local feature estimators should depend on global feature estimators. This conjecture is confirmed by the term of $O_P\left(T^{-2/5}\right)$, i.e., the convergence rate of FPCA global feature estimators, in the derived convergence rate for BTW local feature estimators.

Theorem 3. *Under Assumptions 1 to 7, as $N, T \to \infty$,*

$$\left\| X_t(u) - \widehat{X}_t(u) \right\| = O_P(N^{-\alpha/(1+2\alpha)} + T^{-2/5}).$$

Remark 11. Theorem 3 indicates the estimation error for $X_t(u)$ includes a component from the estimation of global features, and another component from the estimation of local features. As $N, T \to \infty$, both components converge to zero, and we have the total estimation error subsequently converges to zero.

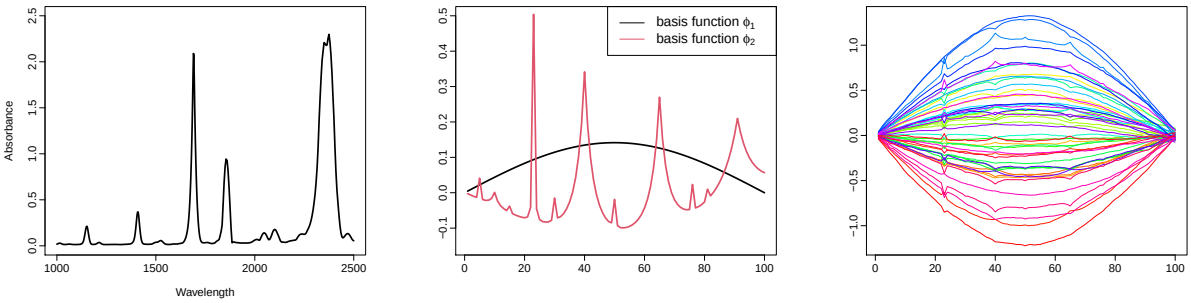
5. Monte Carlo experiments

Finite sample performances of FPCA-BTW estimators are examined through two Monte Carlo experiments. The FPCA-BTW method is applied to make estimation of the functional process and its covariance structure, and produce out-of-sample forecasts. The data generating process for each experiment is calibrated according to real NIR data. Throughout this section, the dimension of global features K is a fixed integer estimated by (8).

5.1. Experiment 1

To generate functional data imitating NIR spectroscopy spectra of a common chemical compound shown in Figure 2a, we extend the "bumps" function of [25] to simulate local features. Figure 2b presents the orthonormalized basis functions used in this experiment. Figure 2c shows an example of simulated trajectories when $T = 50$ where curves with small t indices are shown in red and large t indices are shown in purple. More details about data generation in this experiment, including location and scale local features, can be found in supplementary material S3.1.

For each sample size $T \in \{25, 50, 100\}$, we use simulated signal processes X_t^{TRUE} containing the true global and local features to generate functional time series $\{X_t(u) : X_t(u) = X_t^{\text{TRUE}} + \varepsilon_t(u)\}_{t=1}^{T}$ for $u \in [0, 1]$, where $\varepsilon_t(u)$ is independent noise. We apply dynamic FPCA with $\widehat{K} = 1$ to extract global features, with obtained results denoted by $\widehat{g}_t^{FPCA}(u)$. The BTW method is then applied to extract local features from FPCA residuals. Averaged over 1000



(a) NIR spectrum of chloroform. **(b)** Orthonormalized basis functions. **(c)** Simulated data.

Fig. 2: Motivation data, designed basis functions and a set of 50 simulated trajectories for Experiment 1.

simulation replications, we estimated 10.11, 10.16, and 10.14 non-zero empirical wavelet coefficients at high-resolution levels for $T = 25$, $T = 50$, and $T = 100$ settings, respectively. The number of retained wavelet coefficients after thresholding matches the 11 designed local spikes on each trajectory shown in Figure 2c.

We also fit the unified sparse and functional PCA (SFPCA) method of [2] and the two-way FPCA (TWFPCA) method of [39] together with FPCA to provide benchmarks for the proposed FPCA-BTW method. SFPCA and TWFPCA are implemented with a grid search parameter selection approach provided by the `MoMa` package [67] in R. Although a greedy "coordinate-wise" Bayesian Information Criterion (BIC) optimization scheme by [2] can significantly reduce computation time for SFPCA and TWFPCA, the BIC optimization approach fails to detect most local features in this experiment.

Estimation accuracy is assessed by relative squared error (RSE) defined in a simple Riemann sum as

$$\text{RSE} = \sum_{t=1}^{T} \frac{\left\| \mathcal{X}_t^{\text{TRUE}} - \widehat{g}_t^{\text{FPCA}} - \widehat{Z}_t \right\|^2}{\left\| \mathcal{X}_t^{\text{TRUE}} - \widehat{g}_t^{\text{FPCA}} \right\|^2} = \sum_{t=1}^{T} \sum_{i=1}^{100} \frac{\left| \mathcal{X}_t^{\text{TRUE}}(u_i) - \widehat{g}_t^{\text{FPCA}}(u_i) - \widehat{Z}_t(u_i) \right|^2}{\left| \mathcal{X}_t^{\text{TRUE}}(u_i) - \widehat{g}_t^{\text{FPCA}}(u_i) \right|^2},$$

where $i = \{1, \ldots, 100\}$ denote equally spaced discrete realizations over $[0, 1]$. Given that the denominator of RSE corresponds to the reconstruction accuracy of the FPCA estimator, any estimation method with RSE < 1 has a more accurate estimation performance than the conventional FPCA method. Moreover, the numerator of RSE is proportional to mean squared estimation error defined by $T^{-1} \sum_{t=1}^{T} \left\| \mathcal{X}_t(u) - \widehat{\mathcal{X}}_t(u) \right\|^2$. Thus, small RSE indicates an efficient local feature extraction method.

Table 1: Mean RSE and running time of various local feature extraction methods (standard errors in parentheses) together with p-values of hypothesis tests. The bold entries highlighting the best performing method for each setting.

Sample size		FPCA-SFPCA	FPCA-TWFPCA	FPCA-BTW
$T = 25$	RSE	0.700 (0.081)	0.758 (0.059)	**0.673** (0.088)
	Time	20.795 (0.871)	27.030 (5.931)	**0.098** (0.032)
	p-value	4.57×10^{-13}	$< 2.20 \times 10^{-16}$	NA
$T = 50$	RSE	0.656 (0.065)	0.735 (0.044)	**0.637** (0.064)
	Time	40.746 (1.513)	24.497 (4.226)	**0.139** (0.017)
	p-value	2.22×10^{-11}	$< 2.20 \times 10^{-16}$	NA
$T = 100$	RSE	0.642 (0.047)	0.724 (0.031)	**0.621** (0.047)
	Time	108.136 (3.164)	24.657 (3.159)	**0.225** (0.022)
	p-value	$< 2.20 \times 10^{-16}$	$< 2.20 \times 10^{-16}$	NA

Table 1 presents RSE averaged over 1000 replications for three considered local feature extraction methods, together with averaged computation time (in seconds) for one replication in R on an AMD Ryzen Threadripper 1950X CPU at

3.40GHz. It can be seen that the BTW local feature extraction method consistently outperforms competing methods in estimation accuracy and computation efficiency. All three methods report RSE significantly less than 1, indicating that extracting local features after FPCA dramatically improves estimation accuracy. We conduct t-tests hypotheses "H_0: mean RSE of FPCA-SFPCA (or FPCA-TWFPCA) is the same as mean RSE of FPCA-BTW" against alternatives "H_1: mean RSE of FPCA-SFPCA (or FPCA-TWFPCA) is larger than mean RSE of FPCA-BTW" and report p-values in Table 1. Null hypotheses are rejected for every $T \in \{25, 50, 100\}$ suggesting BTW-FPCA producing the most accurate results of extracted signals. Results relevant to static FPCA are included in supplementary material S3.1.

5.2. Experiment 2

Significant local features of functional time series are often visible in surface plots of long-run covariance functions. To assess the BTW method extracting large "bumps" in the covariance surface, we simulate data following a true theoretical long-run covariance function that has a "pyramid-shaped" local feature, as shown in Figure 3b. Details of calibration of data generating processes for Experiment 2 can be found in supplementary material S3.2.

We apply the FPCA-BTW method to extract global and local features from the simulated functional time series, and use the extracted features to reconstruct long-run covariance functions. The dynamic FPCA estimators are considered as comparison benchmarks. Estimation accuracy for covariance is assessed according to relative error (RE) given by

$$\text{RE} = \sqrt{\sum_{i=1}^{40}\sum_{j=1}^{40} \frac{\left|C(u_i, s_j) - \widehat{C}(u_i, s_j)\right|^2}{\left|C(u_i, s_j)\right|^2}},$$

where $C(u, s)$ is the theoretical long-run covariance function, and $\widehat{C}(u, s)$ is the reconstructed estimator using extracted features; $i, j = \{1, \ldots, 40\}$ denote equally spaced grid points over $[0, 1]$.

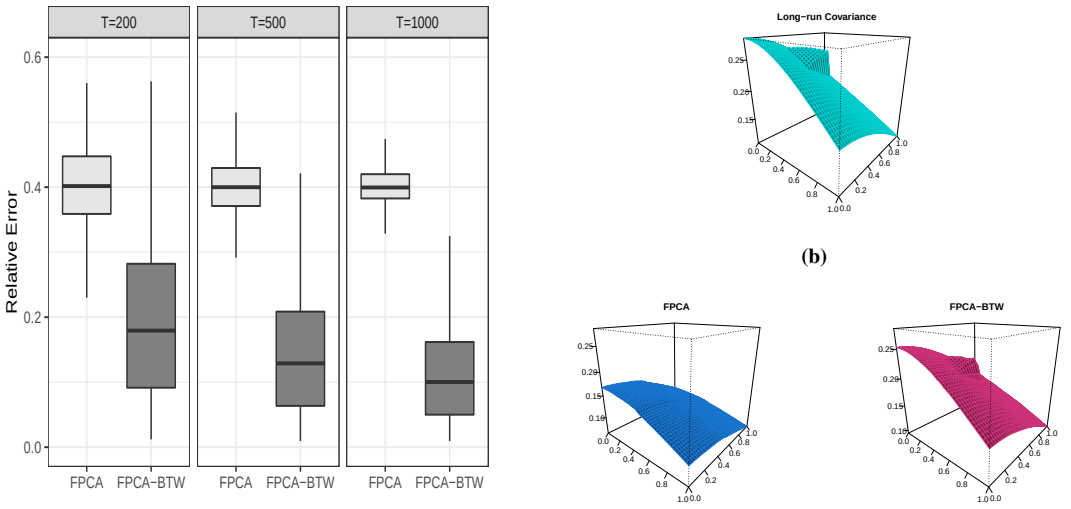
For each sample size $T \in \{200, 500, 1000\}$, we replicate the experiment 1000 times. Throughout the experiment, the empirical dimension of global features is determined to be $\widehat{K} = 1$ by (8). Figure 3a shows that the FPCA-BTW method produces smaller reconstruction errors than the FPCA method. Hence, the extracted local features are tested to improve long-run covariance estimation accuracy. Finally, it can be easily observed that both FPCA and FPCA-BTW methods report smaller estimation errors when sample sizes increase.

Figure 3 visualizes the advantage of FPCA-BTW estimators in long-run covariance estimation when sample size $T = 200$. Estimators depicted by Figure 3d fail to capture the "bump" of local features. In contrast, FPCA-BTW estimators shown in Figure 3c successfully recover most information about local features in the presence of intentionally added noise. We again use t-statistic to test hypotheses "H_0: mean RE of dynamic FPCA estimation is the same as mean RE of dynamic FPCA-BTW estimation" against "H_1: mean RE of dynamic FPCA estimation is greater than mean RE of dynamic FPCA-BTW estimation". We obtain p-values $< 2.20 \times 10^{-16}$ for $T \in \{200, 500, 1000\}$, indicating the superior estimation performance of FPCA-BTW over the classic FPCA method. This experiment shows that local features are essential for the estimation of the long-run covariance function of functional time series.

Monte Carlo experiments introduced in Section 5 prove that FPCA-BTW produces the best feature extraction performance among considered methods. We also design experiments to show that local features extracted by FPCA-BTW help to improve point forecast accuracy, with details included in supplementary material S3.3. In the next section, advantages of the FPCA-BTW method at feature extraction and forecasting functional time series are demonstrated using the empirical wood panel NIR spectroscopy data.

6. Empirical application

The wood panel NIR spectroscopy data illustrated in Figure 1 consists of spectra of absorbance (in negative base ten logarithm of the transmittance) recorded at wavelengths from 350 to 2500 nm in 1 nm intervals in a series of 72 experimental trials. Removing observations from 2301 to 2500 nm because of considerable noise gives $n = 1951$ discrete realizations on each curve. Details of smoothing the raw NIR spectra functions are provided in supplementary material S3.4. We assess the FPCA-BTW method's out-of-sample forecasting performance on the smoothed wood panel NIR data.



(a) **(b)** **(c)** **(d)**

Fig. 3: (a) Relative errors of long-run covariance estimators. (b)–(d) Surface plots of the theoretical long-run covariance function (cyan), along with mean long-run covariance estimators over 1000 simulations obtained by FPCA (blue) and FPCA-BTW (red) for $T = 200$.

Since the smoothed functions $\{\mathcal{X}_t(u)\}_{t=1}^{72}$ are serially correlated, a training set needs to select observations consecutively in time. Hence, the conventional cross-validation strategy of sequentially rotating complementary subsets as testing sets are not appropriate. We adopt the expanding window approach of [69] to gradually increase (decrease) the training (testing) data size in iterations. Specifically, in the initial round a training set $\{\mathcal{X}_1(u), \dots, \mathcal{X}_{62}(u)\}$ and a testing set $\{\widehat{\mathcal{X}}_{63}(u), \dots, \widehat{\mathcal{X}}_{72}(u)\}$ are selected. We apply the FPCA-BTW method to the training set, and use the obtained global and local features to make 10 out-of-sample forecasts. Next, we increase the training set by one observation to $\{\mathcal{X}_1(u), \dots, \mathcal{X}_{63}(u)\}$, and assess the obtained 9 out-of-sample forecasts by a testing set of $\{\widehat{\mathcal{X}}_{64}(u), \dots, \widehat{\mathcal{X}}_{72}(u)\}$. The process is repeated until the training set in the last iteration contains data $\{\mathcal{X}_1(u), \dots, \mathcal{X}_{71}(u)\}$. In total we produce ten one-step-ahead forecasts, nine two-step-ahead forecasts, and so on, up to one 10-step-ahead forecast. Point forecasts obtained without considering local features under the same expanding window setting serve as comparison benchmarks in this application.

To accelerate computation, we pick $n = 500$ equally spaced grids on each $\mathcal{X}_t(u)$, and compute the mean absolute forecast error (MAFE) as

$$\text{MAFE}(h) = \frac{1}{500 \times (11-h)} \sum_{\varsigma=h}^{10} \sum_{i=1}^{500} |\mathcal{X}_{62+\varsigma}(u_i) - \widehat{\mathcal{X}}_{62+\varsigma|62+\varsigma-h}(u_i)|,$$

and the root mean squared forecast error (RMSFE) as

$$\text{RMSFE}(h) = \sqrt{\frac{1}{500 \times (11-h)} \sum_{\varsigma=h}^{10} \sum_{i=1}^{500} \left\{\mathcal{X}_{62+\varsigma}(u_i) - \widehat{\mathcal{X}}_{62+\varsigma|62+\varsigma-h}(u_i)\right\}^2},$$

where $\mathcal{X}_{62+\varsigma}(u_i)$ represents the actual holdout sample at the ith wavelength of the ςth curve, and $\widehat{\mathcal{X}}_{62+\varsigma}(u_i)$ is the corresponding point forecasts. Averaging over ten forecast horizons, we obtain summary statistics given by

$$\text{Median (MAFE)} = \frac{1}{2}[\text{MAFE}(h=5) + \text{MAFE}(h=6)], \quad \text{and} \quad \text{Mean (RMSFE)} = \frac{1}{10}\sum_{h=1}^{10} \text{RMSFE}(h).$$

The median statistic is suitable for handling the absolute error MAFE while the mean statistic is good at handling the squared error RMSFE [28].

Point forecast evaluation results are reported in Tables 2. The forecasts constructed using only global features are shown in the columns with the heading "None", with the remaining columns reporting forecasts produced with global and local features extracted by various methods. It can be easily seen that forecasts produced with local features are consistently more accurate. This result highlights the importance of incorporating local features in forecasting NIR spectroscopy spectra time series. Further, it can be seen that BTW consistently outperforms the competing methods in recovering local features relevant to forecasting. Thus, we recommend FPCA-BTW method in modeling and forecasting functional time series in practice. In addition, a comparison with point forecast evaluation results shown in supplementary material S2.2 indicates that dynamic FPCA produces more accurate point forecasts than static FPCA for the NIR spectroscopy data. This finding indicates that incorporating serial dependence carried by lagged NIR spectroscopy observations improves point forecast accuracy.

Table 2: Mean MAFEs and RMSFEs of point forecasts . The bold entries highlight the feature extraction method with higher forecast accuracy.

	MAFE				RMSFE			
h	FPCA	FPCA-BTW	FPCA-SFPCA	FPCA-TWFPCA	FPCA	FPCA-BTW	FPCA-SFPCA	FPCA-TWFPCA
1	0.482	**0.430**	0.450	0.450	0.870	**0.837**	0.841	0.846
2	0.502	**0.449**	0.473	0.475	0.882	**0.841**	0.852	0.857
3	0.528	**0.475**	0.498	0.502	0.910	**0.872**	0.878	0.884
4	0.537	**0.486**	0.511	0.516	0.918	**0.870**	0.884	0.891
5	0.543	**0.491**	0.513	0.518	0.939	**0.891**	0.902	0.910
6	0.580	**0.533**	0.556	0.566	0.988	**0.938**	0.951	0.962
7	0.598	**0.552**	0.575	0.592	1.016	**0.959**	0.976	0.994
8	0.645	**0.596**	0.627	0.650	1.041	**0.972**	1.003	1.030
9	0.704	**0.646**	0.685	0.717	1.115	**1.044**	1.072	1.105
10	0.593	**0.531**	0.548	0.577	1.144	**1.082**	1.109	1.133
Mean	0.571	**0.519**	0.544	0.556	0.982	**0.931**	0.942	0.961
Median	0.561	**0.511**	0.530	0.542	0.963	**0.915**	0.927	0.936

Table 2 shows that the FPCA-BTW method produces the most accurate point forecasts. Therefore, we do not further consider other competing feature extraction methods. To access the forecast uncertainty of FPCA-BTW method, we adapt the approach of [9] and compute pointwise prediction intervals at the $100(1-a)\%$ nominal coverage probability. Technical details of interval forecasts are provided in supplementary material S2.3. Pointwise predictions intervals are evaluated using the interval score of [29] given by

$$S_a\left[\widehat{\mathcal{X}}^{\text{lb}}_{T+h}(u_i),\widehat{\mathcal{X}}^{\text{ub}}_{T+h}(u_i);\mathcal{X}_{T+h}(u_i)\right]=\left[\widehat{\mathcal{X}}^{\text{ub}}_{T+h}(u_i)-\widehat{\mathcal{X}}^{\text{lb}}_{T+h}(u_i)\right]+\frac{2}{a}\left[\widehat{\mathcal{X}}^{\text{lb}}_{T+h}(u_i)-\mathcal{X}_{T+h}(u_i)\right]\mathbb{1}\left\{\mathcal{X}_{T+h}(u_i)<\widehat{\mathcal{X}}^{\text{lb}}_{T+h}(u_i)\right\}$$
$$+\frac{2}{a}\left[\mathcal{X}_{T+h}(u_i)-\widehat{\mathcal{X}}^{\text{ub}}_{T+h}(u_i)\right]\mathbb{1}\left\{\mathcal{X}_{T+h}(u_i)>\widehat{\mathcal{X}}^{\text{ub}}_{T+h}(u_i)\right\},$$

where $\widehat{\mathcal{X}}^{\text{lb}}_{T+h}(u_i)$ and $\widehat{\mathcal{X}}^{\text{ub}}_{T+h}(u_i)$ denote lower and upper bounds of a symmetric $100(1-a)\%$ prediction interval, and the level of significance is customarily selected as $a=0.2$. To accelerate computation, we again pick $n=500$ equally spaced grids on each $\mathcal{X}_t(u)$. Averaging over different points in a curve and different forecast horizons, the mean interval score is defined as

$$\bar{S}_\alpha(h)=\frac{1}{500\times(11-h)}\sum_{\varsigma=h}^{10}\sum_{i=1}^{500}S_a\left[\widehat{\mathcal{X}}^{\text{lb}}_{T+h}(u_i),\widehat{\mathcal{X}}^{\text{ub}}_{T+h}(u_i);\mathcal{X}_{T+h}(u_i)\right],$$

where $S_a\left[\widehat{\mathcal{X}}^{\text{lb}}_{T+h}(u_i),\widehat{\mathcal{X}}^{\text{ub}}_{T+h}(u_i);\mathcal{X}_{T+h}(u_i)\right]$ denotes the interval score at the ςth curve in the testing set. The interval scores summarized in Figure 4 confirm that incorporating the local features produces more accurate interval forecasts. Moreover, extracting local features after retaining only one empirical functional component gives decent interval

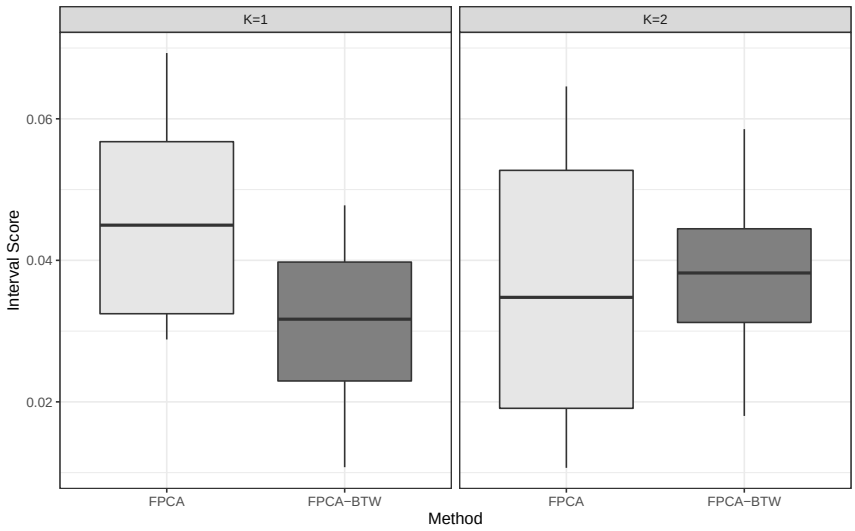


Fig. 4: Scores of pointwise interval forecasts produced by FPCA method and FPCA-BTW method.

forecasts, which further highlights that the BTW method can effectively recover nearly all relevant information of functional data.

7. Conclusion

We propose a novel feature extraction method for functional time series. The proposed FPCA-BTW method improves the feature extraction performance of FPCA by recovering sharp, highly localized features from dimension reduction residuals. Local features extracted by BTW possess information of functional variations over particular short intervals within function domain, contributing to improved estimation results and more accurate forecasts. Theoretical properties of FPCA-BTW method are developed. Superior estimation and forecasting performances of FPCA-BTW estimators in finite samples are verified by Monte Carlo experiments and an empirical application to wood panel NIR spectroscopy data.

There are several ways in which the present paper can be further extended. First, this paper employed Cai's (2002) parametric blockwise threshold approach to select wavelet coefficients. To make the proposed FPCA-BTW a nonparametric feature extraction method, the block size and threshold level at different resolution levels need to be selected based on characteristics of observations. A possible extension of the current method is adopting the data-driven block thresholding approach of [18] to enhance extraction of local features. Moreover, this paper considered forecasting functional time series with extracted linear features. Non-linear extensions of functional regression, for example the continuously additive model of [54], provide enhanced flexibility and structural stability. Another possible extension of the FPCA-BTW method may consider functional additive model [55] as the main dimension reduction tool. Since the inspirational work on functional manifold models by [24], functional manifold models have witnessed increasing contributions in methodology and applications [see, e.g., 50].

Acknowledgment

The authors thank Professor Jiguo Cao from Simon Fraser University for providing us with the lumber data set.

Appendix A. Proofs of the main results

In Appendix A.1, we provide proofs of consistency of FPCA global feature estimators. In Appendix A.2, we present proofs of consistency of BTW local feature estimators. Finally, Appendix A.3 provides proof of consistency of

FPCA-BTW estimators for functional time series. To save space, the preliminary lemmas, together with additional notations facilitating development, are presented in the supplementary material.

A.1. Consistency of global feature estimators

Proof of Proposition 1. Denote a complete set of orthonormal basis on $L^2([0,1])$ by $\{\phi_k\}_{k\in\mathbb{Z}^+}$. The long-run covariance operator C and its estimator $\widehat{C}_{h,q}$ are both positive-definite Hilbert–Schmidt operators, and thus admit the decomposition

$$C(x) = \sum_{k=1}^{\infty} \lambda_k \langle x\,, \phi_k\rangle \phi_k, \quad \text{and} \quad \widehat{C}_{h,q}(x) = \sum_{k=1}^{\infty} \widehat{\lambda}_k \langle x\,, \phi_k\rangle \phi_k, \qquad x \in L^2([0,1]).$$

We then have

$$\sum_{k=1}^{T} \widehat{\lambda}_k = \sum_{k=1}^{\infty} \langle \phi_k\,, \widehat{C}_{h,q}(\phi_k)\rangle, \tag{A.1}$$

$$\sum_{k=1}^{\infty} \lambda_k = \sum_{k=1}^{\infty} \langle \phi_k\,, C(\phi_k)\rangle. \tag{A.2}$$

Deducting (A.2) from (A.1) gives

$$\sum_{k=1}^{T} (\widehat{\lambda}_k - \lambda_k) = \sum_{k=1}^{\infty} \langle \phi_k\,, (\widehat{C}_{h,q} - C)(\phi_k)\rangle + \sum_{k=T+1}^{\infty} \lambda_k. \tag{A.3}$$

By definition, the long-run covariance operator C can be expressed as

$$\begin{aligned} C(x) &= \sum_{\ell=-\infty}^{\infty} \mathrm{E}[\langle \mathcal{X}_t\,, x\rangle \mathcal{X}_{t+\ell}] = \mathrm{E}[\langle \mathcal{X}_t\,, x\rangle \mathcal{X}_t] + \sum_{|\ell|\geq 1}^{\infty} \mathrm{E}[\langle \mathcal{X}_t\,, x\rangle \mathcal{X}_{t+\ell}] + \sum_{|\ell|\geq 1}^{\infty} \mathrm{E}[\langle \mathcal{X}_t\,, x\rangle \mathcal{X}_{t-\ell}] \\ &=: C_0(x) + \sum_{|\ell|\geq 1}^{\infty} C_\ell(x) + \sum_{|\ell|\geq 1}^{\infty} C_{-\ell}(x). \end{aligned}$$

It can be seen that C is self adjoint because, by a direct verification,

$$C_\ell^*(x) \equiv \mathrm{E}[\langle \mathcal{X}_t\,, x\rangle \mathcal{X}_{t-\ell}] = C_{-\ell}(x),$$

where the superscript $*$ denotes the adjoint operator. Similarly, we can prove that $\widehat{C}_{h,q}$ is also self adjoint. In addition, using the relations $C(\phi_k) = \lambda_k \phi_k$ and $\widehat{C}_{h,q}(\widehat{\phi}_k) = \widehat{\lambda}_k \widehat{\phi}_k$, we have for $k = 1, \ldots, K$,

$$\begin{aligned} \left|\langle \phi_k\,, (\widehat{C}_{h,q} - C)(\widehat{\phi}_k)\rangle - (\widehat{\lambda}_k - \lambda_k)\right| &= \left|\langle \phi_k\,, \widehat{C}_{h,q}(\widehat{\phi}_k)\rangle - \langle \phi_k\,, C(\widehat{\phi}_k)\rangle - (\widehat{\lambda}_k - \lambda_k)\right| \\ &= \left|\langle \phi_k\,, \widehat{C}_{h,q}(\widehat{\phi}_k)\rangle - \langle C(\phi_k)\,, \widehat{\phi}_k\rangle - (\widehat{\lambda}_k - \lambda_k)\right| \quad \text{(self adjoint } C\text{)} \\ &= \left|(\widehat{\lambda}_k - \lambda_k)(\langle \phi_k\,, \widehat{\phi}_k\rangle - 1)\right| = \left|\widehat{\lambda}_k - \lambda_k|\langle \phi_k\,, \widehat{\phi}_k\rangle - 1)\right| \\ &= \left|\widehat{\lambda}_k - \lambda_k|\langle \phi_k\,, \widehat{\phi}_k - \phi_k\rangle\right| \\ &\leq \left|\widehat{\lambda}_k - \lambda_k\right| \|\phi_k\| \left\|\widehat{\phi}_k - \phi_k\right\| = \left|\widehat{\lambda}_k - \lambda_k\right| \left\|\widehat{\phi}_k - \phi_k\right\|. \end{aligned}$$

By Lemma 4.2 in [14], $\sup_{k\geq 1} |\widehat{\lambda}_k - \lambda_k| \leq \left\|\widehat{C}_{h,q} - C\right\|_{\mathcal{S}}$. When the optimal bandwidth $\widehat{h}_{\text{opt}}$ is used, by Lemma 2 and Lemma 3 (see supplementary material S1), we then have

$$\widehat{\lambda}_k = \lambda_k + O_P\left(T^{-2/5}\right), \qquad k = 1, \ldots, K. \tag{A.4}$$

Now consider $k = K+1, \ldots, k_{max}$. For $x \in L^2([0,1])$, define the following operators

$$Q_1(x) := \sum_{k=K+1}^{k_{max}} \langle x, \phi_k \rangle \phi_k, \quad Q_2(x) := \sum_{k=1}^{K} \langle x, \phi_k \rangle \phi_k, \quad \widehat{Q}_1(x) := \sum_{k=K+1}^{k_{max}} \langle x, \widehat{\phi}_k \rangle \widehat{\phi}_k.$$

Then Q_1, Q_2, the long-run covariance operator C, and the difference of operators $\widehat{C}_{h,q} - C$ correspond to X, Y_A, A and E in Lemma 7 (see supplementary material S1). Using the fact that $CQ_2 = \sum_{k=1}^{K} \lambda_k \phi_k$ and $\langle \phi_j, \phi_k \rangle = 0 \, \forall j \neq k$, we can get

$$\|B_{12}\|_{\mathcal{L}} := \left\| Q_1^* C Q_2 \right\|_{\mathcal{L}} = \left\| Q_1^* \Big(\sum_{k=1}^{K} \lambda_k \phi_k \Big) \right\|_{\mathcal{L}} = 0.$$

Next, by results of Lemma 2 and Lemma 3 (see supplementary material S1),

$$\|E_{11}\|_{\mathcal{L}} := \left\| Q_1^* (\widehat{C}_{h,q} - C) Q_1 \right\|_{\mathcal{L}} \leq \|Q_1\|_{\mathcal{L}}^2 \left\| \widehat{C}_{h,q} - C \right\|_{\mathcal{S}} \leq O_P(T^{-2/5}),$$

where the last inequality is due to

$$\|Q_1\|_{\mathcal{L}}^2 = \sup_{\|x\| \leq 1} \left\{ \sum_{k=K+1}^{k_{max}} \langle x, \phi_k \rangle^2 \right\} < \sup_{\|x\| \leq 1} \left\{ \sum_{k=1}^{\infty} \langle x, \phi_k \rangle^2 \right\} = \sup_{\|x\| \leq 1} \|x\|^2 .$$

In Lemma 7 (see supplementary material S1), $\mathscr{Y}$ is the closure of $\mathscr{Y}_A$ and Y the extension of Y_A to $\mathscr{Y}$. We define a $Q_3(x) := \sum_{k=1}^{K} \langle x, \phi_k \rangle \phi_k + \sum_{k=k_{max}+1}^{\infty} \langle x, \phi_k \rangle \phi_k$, with $x \in L^2([0,1])$ corresponding to Y in the lemma. It can be easily seen that

$$\|Q_3\|_{\mathcal{L}}^2 = \left\| \sum_{k=1}^{K} \langle x, \phi_k \rangle \phi_k + \sum_{k=k_{max}+1}^{\infty} \langle x, \phi_k \rangle \phi_k \right\|_{\mathcal{L}}^2 \leq \sup_{\|x\| \leq 1} \left\{ \sum_{k=1}^{K} \langle x, \phi_k \rangle^2 + \sum_{k=k_{max}+1}^{\infty} \langle x, \phi_k \rangle^2 \right\} < \sup_{\|x\| \leq 1} \left\{ \sum_{k=1}^{\infty} \langle x, \phi_k \rangle^2 \right\} = \sup_{\|x\| \leq 1} \|x\|^2 .$$

We then have

$$\begin{aligned}
\|E_{21}\|_{\mathcal{L}} &:= \left\| Q_3^* (\widehat{C}_{h,q} - C) Q_1 \right\|_{\mathcal{L}} \leq \left\| Q_3^* \right\|_{\mathcal{L}} \left\| \widehat{C}_{h,q} - C \right\|_{\mathcal{S}} \|Q_1\|_{\mathcal{L}} = O_P(T^{-2/5}), \\
\|E_{12}\|_{\mathcal{L}} &:= \left\| Q_1^* (\widehat{C}_{h,q} - C) Q_3 \right\|_{\mathcal{L}} \leq \|Q_1\|_{\mathcal{L}} \left\| \widehat{C}_{h,q} - C \right\|_{\mathcal{S}} \|Q_3\|_{\mathcal{L}} = O_P(T^{-2/5}), \quad \text{and} \\
\|E_{22}\|_{\mathcal{L}} &:= \left\| Q_3^* (\widehat{C}_{h,q} - C) Q_3 \right\|_{\mathcal{L}} \leq \|Q_3\|_{\mathcal{L}}^2 \left\| \widehat{C}_{h,q} - C \right\|_{\mathcal{S}} = O_P(T^{-2/5}).
\end{aligned}$$

We have previously checked that the long-run covariance operator C is self-adjoint. This corresponds to the well-known fact that if C is Hermitian and ϕ is an approximate normalized eigenvector, then $\phi^* C \phi$ is an approximate eigenvalue. Thus we have B_{11} and $B_{22} := Q_2^* C Q_2$. The separation between of B_{11} and B_{22} satisfies

$$\operatorname{sep}(B_{11}, B_{22}) \geq \min_{\lambda_i \in \lambda(B_{11}), \lambda_j \in \lambda(B_{22})} \left| \lambda_i - \lambda_j \right| \geq |\lambda_K - \lambda_{K+1}| > 0$$

where the last inequality due to Assumption 4 requiring that $\lambda_K > 0$ and $\lambda_{K+1}/\lambda_K = o(1)$.

Now we readily have the condition in Lemma 7 (see supplementary material S1) satisfied such that

$$\begin{aligned}
\theta^{-2} \eta \|E_{21}\|_{\mathcal{L}} &\leq \frac{\left(\left\| Q_1^* C Q_2 \right\|_{\mathcal{L}} + \left\| Q_1^* (\widehat{C}_{h,q} - C) Q_3 \right\|_{\mathcal{L}} \right) \left\| Q_3^* (\widehat{C}_{h,q} - C) Q_1 \right\|_{\mathcal{L}}}{\left(|\lambda_{K+1} - \lambda_K| - \left\| Q_1^* (\widehat{C}_{h,q} - C) Q_1 \right\|_{\mathcal{L}} - \left\| Q_3^* (\widehat{C}_{h,q} - C) Q_3 \right\|_{\mathcal{L}} \right)^2} \\
&\leq \frac{\left(O_P(T^{-2/5}) + O_P(T^{-2/5}) \right) O_P(T^{-2/5})}{\left(|\lambda_{K+1} - \lambda_K| - O_P(T^{-2/5}) - O_P(T^{-2/5}) \right)^2} < \frac{1}{4}.
\end{aligned}$$

By Lemma 7 (see supplementary material S1), we can then write $\widehat{Q}_1 = (Q_1 + Q_2P)(I + P^*P)^{-1/2}$, with

$$\begin{aligned}
\|P\|_{\mathcal{L}} &\leq \frac{2\,\|E_{21}\|_{\mathcal{L}}}{\mathrm{sep}(Q_1^* C Q_1, Q_2^* C Q_2) - \|E_{11}\|_{\mathcal{L}} - \|E_{22}\|_{\mathcal{L}}} \\
&\leq \frac{2\left\|Q_3^*(\widehat{C}_{h,q} - C)Q_1\right\|_{\mathcal{L}}}{|\lambda_{K+1} - \lambda_K| - \left\|Q_1^*(\widehat{C}_{h,q} - C)Q_1\right\|_{\mathcal{L}} - \left\|Q_3^*(\widehat{C}_{h,q} - C)Q_3\right\|_{\mathcal{L}}} \qquad \text{(A.5)} \\
&\leq \frac{2 \times O_P(T^{-2/5})}{|\lambda_{K+1} - \lambda_K| - O_P(T^{-2/5}) - O_P(T^{-2/5})} = O_P(T^{-2/5}). \qquad \text{(A.6)}
\end{aligned}$$

We can then compute the difference between Q_1 and its estimator as

$$\begin{aligned}
\left\|\widehat{Q}_1 - Q_1\right\|_{\mathcal{L}} &= \left\|(Q_1 + Q_2P)(I + P^*P)^{-1/2} - Q_1\right\|_{\mathcal{L}} \\
&= \left\|\left[Q_1 + Q_2P - Q_1(I + P^*P)^{1/2}\right](I + P^*P)^{-1/2}\right\|_{\mathcal{L}} \\
&\leq \left\|Q_1\left[I - (I + P^*P)^{1/2}\right](I + P^*P)^{-1/2}\right\|_{\mathcal{L}} + \left\|Q_2P(I + P^*P)^{-1/2}\right\|_{\mathcal{L}} \\
&\leq \left\|[I - (I + P^*P)^{1/2}](I + P^*P)^{-1/2}\right\|_{\mathcal{L}} + \left\|P(I + P^*P)^{-1/2}\right\|_{\mathcal{L}} \\
&\leq \left\|I - (I + P^*P)^{1/2}\right\|_{\mathcal{L}} + \|P\|_{\mathcal{L}} \leq 2\,\|P\|_{\mathcal{L}} = O_P(T^{-2/5}). \qquad \text{(A.7)}
\end{aligned}$$

Using the linearity and symmetric properties of inner product, and the fact that $C(\phi_{K+j}) = \lambda_{K+j}\phi_{K+j}$, for $j = 1, \ldots k_{max} - K$, we have

$$\begin{aligned}
\left|\widehat{\lambda}_{K+j}\right| &= \left|\langle\widehat{\phi}_{K+j}, \widehat{C}_{h,q}(\widehat{\phi}_{K+j})\rangle\right| \\
&= \left|\langle\widehat{\phi}_{K+j} - \phi_{K+j} + \phi_{K+j}, (\widehat{C}_{h,q} - C + C)(\widehat{\phi}_{K+j} - \phi_{K+j} + \phi_{K+j})\rangle\right| \\
&= \Big|\langle\widehat{\phi}_{K+j} - \phi_{K+j}, (\widehat{C}_{h,q} - C)(\widehat{\phi}_{K+j} - \phi_{K+j})\rangle + \langle\widehat{\phi}_{K+j} - \phi_{K+j}, C(\widehat{\phi}_{K+j} - \phi_{K+j})\rangle \\
&\quad + 2\langle\widehat{\phi}_{K+j} - \phi_{K+j}, (\widehat{C}_{h,q} - C)(\phi_{K+j})\rangle + 2\langle\widehat{\phi}_{K+j} - \phi_{K+j}, C(\phi_{K+j})\rangle\Big| + \left|\langle\phi_{K+j}, C(\phi_{K+j})\rangle\right| \\
&\leq \left|\lambda_{K+j}\right| + \left|\langle\widehat{\phi}_{K+j} - \phi_{K+j}, (\widehat{C}_{h,q} - C)(\widehat{\phi}_{K+j} - \phi_{K+j})\rangle\right| + \left|\langle\widehat{\phi}_{K+j} - \phi_{K+j}, C(\widehat{\phi}_{K+j} - \phi_{K+j})\rangle\right| \\
&\quad + 2\left|\langle\widehat{\phi}_{K+j} - \phi_{K+j}, (\widehat{C}_{h,q} - C)(\phi_{K+j})\rangle\right| + 2\left|\langle\widehat{\phi}_{K+j} - \phi_{K+j}, C(\phi_{K+j})\rangle\right| \quad \text{(triangle inequality)} \\
&\leq \left|\lambda_{K+j}\right| + \left\|\widehat{\phi}_{K+j} - \phi_{K+j}\right\|^2 \left\|\widehat{C}_{h,q} - C\right\|_{\mathcal{S}} + \left\|\widehat{\phi}_{K+j} - \phi_{K+j}\right\|^2 \|C\|_{\mathcal{S}} \\
&\quad + 2\left\|\widehat{\phi}_{K+j} - \phi_{K+j}\right\| \left\|\phi_{K+j}\right\| \left\|\widehat{C}_{h,q} - C\right\|_{\mathcal{S}} + 2\left|\lambda_{K+j}\right| \left\|\phi_{K+j}\right\| \left\|\widehat{\phi}_{K+j} - \phi_{K+j}\right\| \quad \text{(Cauchy–Schwarz inequality)} \\
&= \left|\lambda_{K+j}\right| + O_P(T^{-4/5}), \qquad \text{(A.8)}
\end{aligned}$$

where the last inequality follows from (A.7) and Lemmas 2 to 5 (see supplementary material S1).

Denoting $a \asymp b$ if $a = O_P(b)$ and $b = O_P(a)$, conditions in Assumption 4 indicate that $\lambda_{k+1}/\lambda_k \asymp 1$ for $k = 1, \ldots, K-1$, and $\lambda_{K+1}/\lambda_K \asymp o(1)$, with $\lambda_K > 0$. By (A.4), for $k = 1, \ldots, K-1$, we then have

$$\frac{\widehat{\lambda}_{k+1}}{\widehat{\lambda}_k} = \frac{\lambda_{k+1} + O_P\left(T^{-2/5}\right)}{\lambda_k + O_P\left(T^{-2/5}\right)} \asymp 1. \qquad \text{(A.9)}$$

Similarly, when $k = K$, by (A.4) and (A.8), we have

$$\frac{\widehat{\lambda}_{K+1}}{\widehat{\lambda}_K} = \frac{\lambda_{K+1} + O_P\left(T^{-4/5}\right)}{\lambda_K + O_P\left(T^{-2/5}\right)} \xrightarrow{P} 0, \quad \frac{\widehat{\lambda}_K}{\widehat{\lambda}_1} = \frac{\lambda_K + O_P\left(T^{-2/5}\right)}{\lambda_1 + O_P\left(T^{-2/5}\right)} \asymp 1. \qquad \text{(A.10)}$$

Since $\lambda_{K+1} > \lambda_{K+2} > \ldots > \lambda_{k_{max}}$, by (A.8) we have, for $k = K+1, \ldots, k_{max}$,

$$\frac{\widehat{\lambda}_k}{\widehat{\lambda}_1} \leq \frac{\lambda_{K+1} + O_P\left(T^{-4/5}\right)}{\lambda_1 + O_P\left(T^{-2/5}\right)} = o_P(1), \tag{A.11}$$

which is less than the threshold τ in (8). With (A.9), (A.10), and (A.11), we complete the proof of Proposition 1. □

Proof of Theorem 1. We prove this theorem firstly assuming that K is known. We then have

$$\begin{aligned}
&\left\| \sum_{k=1}^{K} \beta_{t,k}\phi_k(u) - \sum_{k=1}^{K} \widehat{\beta}_{k,t}\widehat{\phi}_k(u) \right\| \\
&\leq \sum_{k=1}^{K} \left\| \beta_{t,k}\phi_k(u) - \widehat{\beta}_{k,t}\widehat{\phi}_k(u) \right\| = \sum_{k=1}^{K} \left\| \beta_{t,k}\phi_k(u) + \beta_{t,k}\widehat{\phi}_k(u) - \beta_{t,k}\widehat{\phi}_k(u) - \widehat{\beta}_{k,t}\widehat{\phi}_k(u) \right\| \\
&\leq \sum_{k=1}^{K} \left\{ |\beta_{t,k}| \left\| \phi_k(u) - \widehat{\phi}_k(u) \right\| + |\beta_{t,k} - \widehat{\beta}_{t,k}| \left\| \widehat{\phi}_k(u) \right\| \right\} \quad \text{(triangle inequality)} \\
&\leq \sum_{k=1}^{K} \sqrt{|\beta_{t,k}|^2 \left\| \phi_k(u) - \widehat{\phi}_k(u) \right\|^2} + \sum_{k=1}^{K} \sqrt{\{\beta_{t,k} - \widehat{\beta}_{t,k}\}^2 \left\| \widehat{\phi}_k(u) \right\|^2} \quad \text{(Cauchy–Schwarz inequality)} \\
&= O_P\left(T^{-2/5}\right).
\end{aligned}$$

In the above derivation, we have used the fact that the estimated eigenfunctions have unit length due to normalization, i.e., $\left\| \widehat{\phi}_k(u) \right\|^2 = 1$, and results of Lemma 5 and Lemma 6 (see supplementary material S1) in the last step.

By Proposition 1, $\Pr(\widehat{K} = K) \to 1$. We readily have the unconditional arguments, completing the proof of Theorem 1. □

A.2. Consistency of local feature estimators

Proof. Proof of Theorem 2
The residual function after FPCA is given by

$$\widehat{e}_t(u) = \mathcal{X}_t(u) - \widehat{\mu}(u) - \sum_{k=1}^{\widehat{K}} \widehat{\beta}_{t,k}\widehat{\phi}_k(u) = Z_t(u) + \mu(u) - \widehat{\mu}(u) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(u) - \sum_{k=1}^{\widehat{K}} \widehat{\beta}_{t,k}\widehat{\phi}_k(u) + \varepsilon_t(u)$$

The NRSI initially interpolates residual observations on grids $\{u_1, \ldots, u_{n_t}\}$ into a vector $\widehat{\boldsymbol{e}}_t = [\widehat{e}_t(u_1), \ldots, \widehat{e}_t(u_N)]$ with $N = 2^J > n_t$ equally spaced points. According to [8], approximation errors in this step caused by moving nondyadic points to dyadic points are negligible. Hence, discretized FPCA residuals can be expressed as

$$\widehat{e}_t(u_i) = Z_t(u_i) + \mu(u_i) - \widehat{\mu}(u_i) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(u_i) - \sum_{k=1}^{\widehat{K}} \widehat{\beta}_{t,k}\widehat{\phi}_k(u_i) + \varepsilon_t(u_i), \tag{A.12}$$

where $i = 1, \ldots, N$, $u_i = i/N$. It follows from Definition 3.1 in [14] that strong H-white noise process (independently and identically distributed sequence of random variables with mean 0 and constant variance taking values in H) can be expressed as

$$\varepsilon_t(u_i) = (\mathscr{W}(u_i) - \mathscr{W}(u_{i-1})) \cdot \sigma, \qquad 0 \leq u_{i-1} < u_i \leq 1, \quad t \in \mathbb{Z},$$

where $\mathscr{W}(u)$, for $u \geq 0$ with $\mathscr{W}(0) = 0$ is a measurable bilateral Wiener process, and σ is the noise level. By definition, the Wiener process has independent Gaussian increments, i.e., $\mathscr{W}(u) - \mathscr{W}(0) \sim \text{Normal}(0, u)$, and for $0 \leq u_a < u_b < u_c < u_d \leq 1$, $\mathscr{W}(u_b) - \mathscr{W}(u_a)$ independent of $\mathscr{W}(u_d) - \mathscr{W}(u_c)$. Since $u_i = i/N$ for $i = 1, \ldots, N$, we have equally sized increments $u_i - u_{i-1} = 1/N$. The sequence $\varepsilon_t(u_1), \ldots, \varepsilon_t(u_N)$ therefore follows an i.i.d. Normal$(0, \sigma^2/N)$ distribution.

We consider least asymmetric wavelets $\{\psi, \Psi\}$ constructed by [23]. Using the "subband filtering schemes" discussed by Daubechies [23, Chapter 5], the true function $Z_t(u)$ can be approximated by discretized observations as $\widetilde{Z}_t(u) = \sum_{i=1}^{N} N^{-1/2} Z_t(u_i)\psi_{J,i}(u)$. Let $D'_{j_0,p,t}$ and $D_{j,p,t}$ denote the true wavelet coefficients of $Z_t(u)$, i.e., $D'_{j_0,p,t} = \langle Z_t\,, \psi_{j_0,p}\rangle$ and $D_{j,p,t} = \langle Z_t\,, \Psi_{j,p}\rangle$. Plugging $Z_t(u_i)$ from (A.12) into the last equation, we have

$$
\begin{aligned}
\widetilde{Z}_t(u) &= \sum_{i=1}^{N} N^{-1/2} \left\{ \widehat{e}_t(u_i) - [\mu(u_i) - \widehat{\mu}(u_i)] - \left[\sum_{k=1}^{K} \beta_{t,k}\phi_k(u_i) - \sum_{k=1}^{\widehat{K}} \widehat{\beta}_{t,k}\widehat{\phi}_k(u_i) \right] - \varepsilon_t(u_i) \right\} \psi_{J,i}(u) \\
&= \sum_{i=1}^{N} \left\{ D'_{J,i,t} + [N^{-1/2}\widehat{e}_t(u_i) - D'_{J,i,t}] - N^{-1/2}[\epsilon_i N^{-1/2}\sigma] \right\} \psi_{J,i}(u) \\
&\quad + \sum_{i=1}^{N} \left[\widehat{\mu}(u_i) - \mu(u_i) + \sum_{k=1}^{\widehat{K}} \widehat{\beta}_{t,k}\widehat{\phi}_k(u_i) - \sum_{k=1}^{K} \beta_{t,k}\phi_k(u_i) \right] \psi_{J,i}(u) \\
&= \sum_{p=1}^{2^{j_0}} \left\{ D'_{j_0,p,t} + a'_{j_0,p,t} + \sigma\epsilon_{j_0,p,t}/N \right\} \psi_{j_0,p}(u) + \sum_{j=j_0}^{J-1} \sum_{p=1}^{2^j} \left\{ D_{j,p,t} + a_{j,p,t} + \sigma\epsilon_{j,p,t}/N \right\} \Psi_{j,p}(u) \\
&\quad + \sum_{i=1}^{N} \left[\widehat{\mu}(u_i) - \mu(u_i) + \sum_{k=1}^{\widehat{K}} \widehat{\beta}_{t,k}\widehat{\phi}_k(u_i) - \sum_{k=1}^{K} \beta_{t,k}\phi_k(u_i) \right] \psi_{J,i}(u),
\end{aligned} \tag{A.13}
$$

where ϵ_i's are i.i.d. Normal$(0,1)$ such that $\mathrm{var}(\varepsilon_t(u_i)) = \mathrm{var}(\epsilon_i N^{-1/2}\sigma) = \sigma^2/N$. In (A.13), $D'_{j0,p,t}$ and $D_{j,p,t}$ are the orthogonal transform of $\{D'_{J,i,t}\}_{i=1}^{N}$ via the DWT base matrix $\boldsymbol{W}$, likewise $a'_{j_0,p,t}$ and $a_{j,p,t}$ the transform of $\{N^{-1/2}\widehat{e}_t(u_i) - D'_{J,i,t}\}_{i=1}^{N}$, and $\epsilon_{j_0,p,t}$ and $\epsilon_{j,p,t}$ the transform of $\{\epsilon_i\}_{i=1}^{N}$. The $\epsilon_{j_0,p,t}$ and $\epsilon_{j,p,t}$ are i.i.d. Normal$(0,1)$ since ϵ_i's are i.i.d. Normal$(0,1)$. By Lemma 10 (see supplementary material S1), the approximation errors satisfy

$$
\sum_{p=1}^{2^{j_0}} (a'_{j_0,p,t})^2 + \sum_{j=j_0}^{J-1} \sum_{p=1}^{2^j} a^2_{j,p,t} = \sum_{i=1}^{N} [N^{-1/2}\widehat{e}_t(u_i) - D'_{J,i,t}]^2 = o(N^{-2\alpha/(1+2\alpha)}). \tag{A.14}
$$

More details about the derivation of this result can be found in Page 43 of [33].

Let $\widehat{D}'_{j_0,p,t} = D'_{j_0,p,t} + a'_{j_0,p,t} + \sigma\epsilon_{j_0,p,t}/N$ and $\widehat{D}_{j,p,t} = D_{j,p,t} + a_{j,p,t} + \sigma\epsilon_{j,p,t}/N$ denote the NRSI wavelet coefficients. By Lemma 8 (see supplementary material S1), $\widehat{\boldsymbol{D}}_t \in B^{\alpha}_{P,Q}$. According to Definition 1, the Besov space $B^{\alpha}_{P,Q}$ is a subset of the function class $\mathcal{H}(\alpha_1, \alpha, \gamma, M_1, M_2, M_3, r, v)$ (more details see Example 3.1 in [33]). Thus, we can apply Lemma 10 and Lemma 11 (see supplementary material S1) in the following derivations involving NRSI estimator $\widehat{\boldsymbol{D}}_t$. Denoting the wavelet coefficients after blockwise thresholding as $\widetilde{\boldsymbol{D}}_t$ as in Section 3.2, according to (6) we have $\widetilde{D}'_{j_0,p,t} = \widehat{D}'_{j_0,p,t}$ and $\widetilde{D}_{j,p,t} = \widehat{D}_{j,p,t}\mathbb{1}(S^2_{j_a} > \lambda^* L^* \sigma^2/N)$ for $(j,p) \in j_a$. The orthonormal wavelet functions satisfy $\|\psi\| = \|\Psi\| = 1$. By the isometry of the function norm and the sequence norm, then by triangle inequality we have

$$
\begin{aligned}
\mathrm{E}\left\|Z_t(u) - \widetilde{Z}_t(u)\right\|^2 \le \mathfrak{c}_0 \Bigg\{ &\sum_{p=1}^{2^{j_0}} \mathrm{E}(D'_{j_0,p,t} - \widetilde{D}'_{j_0,p,t})^2 + \sum_{j=j_0}^{J-1} \sum_{p=1}^{2^j} \mathrm{E}(D_{j,p,t} - \widetilde{D}_{j,p,t})^2 + \sum_{j=J}^{\infty} \sum_{p=1}^{2^j} D^2_{j,p,t} \\
&+ \sum_{i=1}^{N} \mathrm{E}[\widehat{\mu}(u_i) - \mu(u_i)]^2 + \sum_{i=1}^{N} \mathrm{E}\left[\sum_{k=1}^{\widehat{K}} \widehat{\beta}_{t,k}\widehat{\phi}_k(u_i) - \sum_{k=1}^{K} \beta_{t,k}\phi_k(u_i) \right]^2 \Bigg\},
\end{aligned} \tag{A.15}
$$

where $\mathfrak{c}_0$ is a constant.

We need to show that $\mathrm{E}\left\|Z_t(u) - \widetilde{Z}_t(u)\right\|^2 \to 0$ as $T, N \to \infty$. The result of convergence can be easily confirmed since each summand in (A.15) converges to zero:

- By Lemma 10 and (A.14) (see supplementary material S1),

$$\sum_{p=1}^{2^{j_0}} \mathrm{E}(D'_{j_0,p,t} - \widetilde{D}'_{j_0,p,t})^2 + \sum_{j=J}^{\infty}\sum_{p=1}^{2^j} D^2_{j,p,t} = o(N^{-2\alpha/(1+2\alpha)}).$$

- By Lemma 11 and (A.14) (see supplementary material S1),

$$\sum_{j=j_0}^{J-1}\sum_{p=1}^{2^j} \mathrm{E}(D_{j,p,t} - \widetilde{D}_{j,p,t})^2 = O(N^{-2\alpha/(1+2\alpha)}).$$

- By Lemma 1 (2) (see supplementary material S1),

$$\sum_{i=1}^{N} \mathrm{E}[\widehat{\mu}(u_i) - \mu(u_i)]^2 < \mathrm{E}\left\|\widehat{\mu}(u) - \mu(u)\right\|^2 = O(1/T).$$

- By Theorem 1,

$$\sum_{i=1}^{N} \mathrm{E}\left[\sum_{k=1}^{\widehat{K}} \widehat{\beta}_{t,k}\widehat{\phi}_k(u_i) - \sum_{k=1}^{K} \beta_{t,k}\phi_k(u_i)\right]^2 < \mathrm{E}\left\|\sum_{k=1}^{\widehat{K}} \widehat{\beta}_{t,k}\widehat{\phi}_k(u) - \sum_{k=1}^{K} \beta_{t,k}\phi_k(u)\right\|^2 = O(T^{-4/5}).$$

Hence, the MSE of the estimator $\widetilde{Z}_t(u)$ satisfies

$$\mathrm{E}\left\|Z_t(u) - \widetilde{Z}_t(u)\right\|^2 = O(N^{-2\alpha/(1+2\alpha)} + T^{-4/5}).$$

The Chebyshev's inequality then implies that,

$$\left\|Z_t(u) - \widetilde{Z}_t(u)\right\| = O_P(N^{-\alpha/(1+2\alpha)} + T^{-2/5}) = o_P(1).$$

□

A.3. Consistency of functional time series estimators

Proof. Proof of Theorem 3
This theorem can be easily proved with results of Theorems 1 and 2. By triangle inequality, we have

$$\begin{aligned}
\mathrm{E}\left\|\mathcal{X}_t(u) - \widehat{\mathcal{X}}_t(u)\right\|^2 &= \left\|\sum_{k=1}^{K} \beta_{t,k}\phi_k(u) + Z_t(u) + \varepsilon_t(u) - \left(\sum_{k=1}^{\widehat{K}} \widehat{\beta}_{t,k}\widehat{\phi}_k + \widehat{Z}_t(u)\right)\right\|^2 \\
&\leq \mathrm{E}\left\|\sum_{k=1}^{K} \beta_{t,k}\phi_k(u) - \sum_{k=1}^{\widehat{K}} \widehat{\beta}_{t,k}\widehat{\phi}_k\right\|^2 + \left\|Z_t(u) - \widehat{Z}_t(u)\right\|^2 \\
&= O_P(T^{-4/5}) + O(N^{-2\alpha/(1+2\alpha)} + T^{-4/5}).
\end{aligned}$$

The Chebyshev's inequality then implies that

$$\left\|\mathcal{X}_t(u) - \widehat{\mathcal{X}}_t(u)\right\| = O_P(T^{-2/5}) + O_P(N^{-\alpha/(1+2\alpha)} + T^{-2/5}).$$

Since N is a positive integer, we have $N^{-\alpha/(1+2\alpha)} > 0$ for $\alpha > 0$. Thus,

$$\left\|\mathcal{X}_t(u) - \widehat{\mathcal{X}}_t(u)\right\| = O_P(N^{-\alpha/(1+2\alpha)} + T^{-2/5}).$$

□

Appendix B. Supplementary material

Supplementary material related to this article can be found online. The supplementary material contains detailed proofs of preliminary lemmas used for derivation of theoretical results, technical details of implementing the FPCA-BTW method, and additional simulation results.

References

[1] Ahn, S. C. and Horenstein, A. R. [2013], 'Eigenvalue ratio test for the number of factors', *Econometrica* **81**(3), 1203–1227.

[2] Allen, G. I. and Weylandt, M. [2019], Sparse and functional principal components analysis, *in* '2019 IEEE Data Science Workshop (DSW)', IEEE, pp. 11–16.

[3] Andrews, D. [1991], 'Heteroskedasticity and autocorrelation consistent covariant matrix estimation', *Econometrica* **59**(3), 817–858.

[4] Aneiros, G., Cao, R., Fraiman, R., Genest, C. and Vieu, P. [2019], 'Recent advances in functional data analysis and high-dimensional statistics', *Journal of Multivariate Analysis* **170**, 3–9.

[5] Aneiros, G., Horová, I., Hušková, M. and Vieu, P. [2020], *Functional and High-Dimensional Statistics and Related Fields*, Springer, Cham, Switzerland.

[6] Aneiros, G. and Vieu, P. [2014], 'Variable selection in infinite-dimensional problems', *Statistics & Probability Letters* **94**, 12–20.

[7] Antoniadis, A. [2007], 'Wavelet methods in statistics: Some recent developments and their applications', *Statistics Surveys* **1**, 16–55.

[8] Antoniadis, A. and Fan, J. [2001], 'Regularization of wavelet approximations', *Journal of the American Statistical Association: Theory and Methods* **96**(455), 939–967.

[9] Aue, A., Norinho, D. D. and Hörmann, S. [2015], 'On the prediction of stationary functional time series', *Journal of the American Statistical Association: Theory and Methods* **110**(509), 378–392.

[10] Bathia, N., Yao, Q. and Ziegelmann, F. [2010], 'Identifying the finite dimensionality of curve time series', *The Annals of Statistics* **38**(6), 3352–3386.

[11] Berkes, I., Horváth, L. and Rice, G. [2016], 'On the asymptotic normality of kernel estimators of the long run covariance of functional time series', *Journal of Multivariate Analysis* **144**, 150–175.

[12] Berrendero, J. R., Bueno-Larraz, B. and Cuevas, A. [2019], 'An rkhs model for variable selection in functional linear regression', *Journal of Multivariate Analysis* **170**, 25–45.

[13] Berrendero, J. R., Cuevas, A. and Pateiro-López, B. [2016], 'Shape classification based on interpoint distance distributions', *Journal of Multivariate Analysis* **146**, 237–247.

[14] Bosq, D. [2000], *Linear Processes in Function Spaces*, Springer Science+Business Media, New York.

[15] Bosq, D. and Blanke, D. [2007], *Inference and Prediction in Large Dimensions*, John Wiley & Sons, West Sussex, England.

[16] Burns, D. A. and Ciurczak, E. W. [2007], *Handbook of Near-Infrared Analysis*, CRC press, Boca Raton, Florida.

[17] Cai, T. T. [2002], 'On block thresholding in wavelet regression: Adaptivity, block size, and threshold level', *Statistica Sinica* **12**, 1241–1273.

[18] Cai, T. T. and Zhou, H. H. [2009], 'A data-driven block thresholding approach to wavelet estimation', *The Annals of Statistics* **37**(2), 569–595.

[19] Cao, J., Sang, P., Groves, K., Feng, M. and FPInnovations [2018], Stopping time detection in functional time series: An application to wood panel glue curing process. Joint Statistical Meetings 2018, Vancouver, Canada.

[20] Chiou, J.-M. [2012], 'Dynamical functional prediction and classification with application to traffic flow prediction', *The Annals of Applied Statistics* **6**(4), 1588–1614.

[21] Chiou, J.-M., Yang, Y.-F. and Chen, Y.-T. [2016], 'Multivariate functional linear regression and prediction', *Journal of Multivariate Analysis* **146**, 301–312.

[22] Cuevas, A., Febrero, M. and Fraiman, R. [2007], 'Robust estimation and classification for functional data via projection-based depth notions', *Computational Statistics* **22**(3), 481–496.

[23] Daubechies, I. [1992], *Ten Lectures on Wavelets*, Society for Industrial and Applied Mathematics, Philadelphia, PA.

[24] Donoho, D. L. and Grimes, C. [2005], 'Image manifolds which are isometric to euclidean space', *Journal of mathematical imaging and vision* **23**(1), 5–24.

[25] Donoho, D. L. and Johnstone, J. M. [1994], 'Ideal spatial adaptation by wavelet shrinkage', *Biometrika* **81**(3), 425–455.

[26] Fan, J., Liao, Y. and Mincheva, M. [2013], 'Large covariance estimation by thresholding principal orthogonal complements', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**(4), 603–680.

[27] Gellar, J. E., Colantuoni, E., Needham, D. M. and Crainiceanu, C. M. [2014], 'Variable-domain functional regression for modeling ICU data', *Journal of the American Statistical Association: Applications and Case Studies* **109**(508), 1425–1439.

[28] Gneiting, T. [2011], 'Making and evaluating point forecasts', *Journal of the American Statistical Association: Review Article* **106**(494), 746–762.

[29] Gneiting, T. and Raftery, A. E. [2007], 'Strictly proper scoring rules, prediction and estimation', *Journal of the American Statistical Association: Review Article* **102**(477), 359–378.

[30] Goia, A. and Vieu, P. [2016], 'An introduction to recent advances in high/infinite dimensional statistics', *Journal of Multivariate Analysis* **146**, 1–6.

[31] Grossmann, A. and Morlet, J. [1984], 'Decomposition of hardy functions into square integrable wavelets of constant shape', *SIAM Journal on Mathematical Analysis* **15**(4), 723–736.

[32] Hall, P. and Hooker, G. [2016], 'Truncated linear models for functional data', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78**(3), 637–653.

[33] Hall, P., Kerkyacharian, G. and Picard, D. [1999], 'On the minimax optimality of block thresholded wavelet estimators', *Statistica Sinica* **9**(1), 33–49.
[34] Hall, P. and Vial, C. [2006], 'Assessing the finite dimensionality of functional data', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(4), 689–705.
[35] Hörmann, S., Kidziński, Ł. and Hallin, M. [2015], 'Dynamic functional principal components', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77**(2), 319–348.
[36] Hörmann, S. and Kokoszka, P. [2010], 'Weakly dependent functional data', *The Annals of Statistics* **38**(3), 1845–1884.
[37] Horváth, L. and Kokoszka, P. [2012], *Inference for Functional Data with Applications*, Vol. 200, Springer Science & Business Media, New York.
[38] Horváth, L., Rice, G. and Whipple, S. [2016], 'Adaptive bandwidth selection in the long run covariance estimator of functional time series', *Computational Statistics & Data Analysis* **100**, 676–693.
[39] Huang, J. Z., Shen, H. and Buja, A. [2009], 'The analysis of two-way functional data using two-way regularized singular value decompositions', *Journal of the American Statistical Association: Theory and Methods* **104**(488), 1609–1620.
[40] Hyndman, R. J. and Shang, H. L. [2009], 'Forecasting functional time series (with discussions)', *Journal of the Korean Statistical Society* **38**(3), 199–221.
[41] Hyndman, R. J. and Shang, H. L. [2010], 'Rainbow plots, bagplots, and boxplots for functional data', *Journal of Computational and Graphical Statistics* **19**(1), 29–45.
[42] Johnstone, I. M. and Lu, A. Y. [2009], 'On consistency and sparsity for principal components analysis in high dimensions', *Journal of the American Statistical Association: Theory and Methods* **104**(486), 682–693.
[43] Klepsch, J. and Klüppelberg, C. [2017], 'An innovations algorithm for the prediction of functional linear processes', *Journal of Multivariate Analysis* **155**, 252–271.
[44] Klepsch, J., Klüppelberg, C. and Wei, T. [2017], 'Prediction of functional ARMA processes with an application to traffic data', *Econometrics and Statistics* **1**, 128–149.
[45] Kokoszka, P. and Reimherr, M. [2013], 'Determining the order of the functional autoregressive model', *Journal of Time Series Analysis* **34**(1), 116–129.
[46] Kuhnt, S. and Rehage, A. [2016], 'An angle-based multivariate functional pseudo-depth for shape outlier detection', *Journal of Multivariate Analysis* **146**, 325–340.
[47] Lam, C. and Yao, Q. [2012], 'Factor modeling for high-dimensional time series: inference for the number of factors', *The Annals of Statistics* **40**(2), 694–726.
[48] Lam, C., Yao, Q. and Bathia, N. [2011], 'Estimation of latent factors for high-dimensional time series', *Biometrika* **98**(4), 901–918.
[49] Li, D., Robinson, P. M. and Shang, H. L. [2020], 'Long-range dependent curve time series', *Journal of the American Statistical Association: Theory and Methods* **115**(530), 957–971.
[50] Lin, Z. and Yao, F. [2019], 'Intrinsic riemannian functional data analysis', *The Annals of Statistics* **47**(6), 3533–3577.
[51] Mallat, S. G. [1989], 'A theory for multiresolution signal decomposition: the wavelet representation', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**(7), 674–693.
[52] Mallat, S. G. [2009], *A Wavelet Tour of Signal Processing: the Sparse Way*, 3rd edn, Elsevier/Academic Press, Amsterdam; Boston.
[53] Meyer, Y. [1992], *Wavelets and operators*, Vol. 1, Cambridge University Press, Cambridge.
[54] Müller, H.-G., Wu, Y. and Yao, F. [2013], 'Continuously additive models for nonlinear functional regression', *Biometrika* **100**(3), 607–622.
[55] Müller, H.-G. and Yao, F. [2008], 'Functional additive models', *Journal of the American Statistical Association: Theory and Methods* **103**(484), 1534–1544.
[56] Novo, S., Aneiros, G. and Vieu, P. [2019], 'Automatic and location-adaptive estimation in functional single-index regression', *Journal of Nonparametric Statistics* **31**(2), 364–392.
[57] Novo, S., Aneiros, G. and Vieu, P. [2021], 'A knn procedure in semiparametric functional data analysis', *Statistics & Probability Letters* **171**, 109028.
[58] Ogden, T. [1997], *Essential Wavelets for Statistical Applications and Data Analysis*, Springer, Boston.
[59] Parzen, E. [1957], 'On consistent estimates of the spectrum of a stationary time series', *The Annals of Mathematical Statistics* **28**(2), 329–348.
[60] Politis, D. N. and Romano, J. P. [1996], 'On flat-top kernel spectral density estimators for homogeneous random fields', *Journal of Statistical Planning and Inference* **51**(1), 41–53.
[61] R Core Team [2020], *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. **URL:** http://www.R-project.org/
[62] Rice, G. and Shang, H. L. [2017], 'A plug-in bandwidth selection procedure for long-run covariance estimation with stationary functional time series', *Journal of Time Series Analysis* **38**(4), 591–609.
[63] Shang, H. L. [2019], 'Dynamic principal component regression: Application to age-specific mortality forecasting', *ASTIN Bulletin: The Journal of the IAA* **49**(3), 619–645.
[64] Shang, H. L. and Hyndman, R. J. [2011], 'Nonparametric time series forecasting with dynamic updating', *Mathematics and Computers in Simulation* **81**(7), 1310–1324.
[65] Solo, V. [2001], 'Regularization of wavelet approximations: Discussion', *Journal of the American Statistical Association: Theory and Methods* **96**(455), 963–964.
[66] Strang, G. [1989], 'Wavelets and dilation equations: A brief introduction', *SIAM review* **31**(4), 614–627.
[67] Weylandt, M., Allen, G. and Liao, L. [2018], *MoMA: MoMA - Modern Multivariate Analysis in R*. R package version 0.1. **URL:** https://github.com/DataSlingers/MoMA
[68] Zhao, Y., Ogden, R. T. and Reiss, P. T. [2012], 'Wavelet-based lasso in functional linear regression', *Journal of Computational and Graphical Statistics* **21**(3), 600–617.
[69] Zivot, E. and Wang, J. [2006], *Modeling Financial Time Series with S-PLUS*, Springer, New York.