Statistica Sinica **33** (2023), 1-27
doi:https://doi.org/10.5705/ss.202020.0365

TWO-SAMPLE TESTS FOR RELEVANT DIFFERENCES IN THE EIGENFUNCTIONS OF COVARIANCE OPERATORS

Alexander Aue, Holger Dette and Gregory Rice

University of California, Ruhr-Universität Bochum and University of Waterloo

Abstract: This study examines two-sample tests for functional time series data, which have become widely available with the advent of modern complex observation systems. Here, we evaluate whether two sets of functional time series observations share the shape of their primary modes of variation, as encoded by the eigenfunctions of the respective covariance operators. To this end, a novel testing approach is introduced that adds to existing literature in two main ways. First, tests are set up in the relevant testing framework, where interest is not in testing an exact null hypothesis, but rather in detecting deviations deemed sufficiently relevant, with relevance determined by the practitioner and perhaps guided by domain experts. Second, the proposed test statistics rely on a self-normalization principle that helps to avoid the notoriously difficult task of estimating the long-run covariance structure of the underlying functional time series. The main theoretical result of this study is the derivation of the large-sample behavior of the proposed test statistics. Empirical evidence, which indicates that the proposed procedures work well in finite samples and compare favorably with competing methods, is provided through a simulation study and an application to annual temperature data.

Key words and phrases: Functional data, functional time series, relevant tests, self-normalization.

1. Introduction

This study develops testing tools for two independent sets of functional observations, explicitly allowing for temporal dependence within each set. Functional data analysis has become a mainstay for dealing with those complex data sets that may conceptually be viewed as being comprised of curves. Monographs detailing many of the available statistical procedures for functional data are provided by Ramsay and Silverman (2005) and Horváth and Kokoszka (2012). This type of data naturally arises in contexts such as environmental data (Aue, Dubart Norinho and Hörmann (2015)), molecular biophysics (Tavakoli and Panaretos (2016)), climate science (Zhang et al. (2011); Aue, Rice and Sönmez (2018)), and

Corresponding author: Gregory Rice, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada. E-mail: grice@uwaterloo.ca.

economics (Kowal, Matteson and Ruppert (2019)). Most of these examples intrinsically contain a time series component, because successive curves are expected to depend on each other. As a result, the literature on functional time series has grown steadily; see, for example, Hörmann and Kokoszka (2010), Panaretos and Tavakoli (2013) and the references therein.

Our main goal is to develop two-sample tests for comparing the second-order properties of functional time series data. Two-sample inference and testing methods for curves have been developed by several authors. Hall and Van Keilegom (2007) examined the effect of pre-processing discrete data into functions on two-sample testing procedures. Horváth, Kokoskza and Reeder (2013) investigated two-sample tests for the equality of the means of two functional time series taking values in the Hilbert space of square integrable functions. Dette, Kokot and Aue (2020) introduced multiplier bootstrap-assisted two-sample tests for functional time series taking values in the Banach space of continuous functions. Panaretos, Kraus and Maddocks (2010), Fremdt et al. (2013), Pigoli et al. (2014), Paparoditis and Sapatinas (2016), and Guo, Zhou and Zhang (2018) provided procedures for testing the equality of covariance operators in functional samples.

Although general differences between covariance operators can be attributed to differences in the eigenfunctions of the operators, eigenvalues of the operators, or perhaps both, we focus here on constructing two-sample tests that focus on differences in the eigenfunctions. The eigenfunctions of covariance operators hold a special place in functional data analysis owing to their near ubiquitous use in dimension reduction via functional principal component analysis (FPCA). FPCA is the basis of most inferential procedures for functional data. In fact, an assumption common to a number of such procedures is that observations from different samples/populations share a common eigenbasis generated by their covariance operators; see Benko, Härdle and Kneip (2009) and Pomann, Staicu and Ghosh (2016). FPCA is arguably even more crucial to the analysis of functional time series, because it underlies most forecasting and change-point methods; see, for example, Aue, Dubart Norinho and Hörmann (2015), Hyndman and Shang (2009), and Aston and Kirch (2012). The tests proposed here are useful both for determining the plausibility that two samples share similar eigenfunctions, or whether or not one should pool together data observed in different samples for a joint analysis of their principal components. We illustrate these applications in Section 4 in an analysis of annual temperature profiles recorded at several locations, for which the shape of the eigenfunctions can help in the interpretation of geographical differences in the primary modes of temperature variation over time. A more detailed argument for the usefulness and impact of such tests on

validating climate models is given in the introduction of Zhang and Shao (2015).

The procedures introduced in this paper are noteworthy in at least two respects. First, unlike in the existing literature, they are phrased in the relevant testing framework. In this paradigm, deviations from the null are deemed of interest only if they surpass a minimum threshold set by the practitioner. Classical hypothesis tests are included in this approach if the threshold is chosen to be equal to zero. There are several advantages of the relevant framework. In general, it avoids Berkson's consistency problem (Berkson (1938)) that any consistent test will reject for arbitrarily small differences if the sample size is sufficiently large. More specific to functional data, the L^2-norm sample mean curve differences might not be close to zero, even if the underlying population mean curves coincide. Adopting the relevant framework typically comes at the cost of having to invoke involved theoretical arguments. A recent review of methods for testing relevant hypotheses in two-sample problems with one-dimensional data from a biostatistics perspective can be found in Wellek (2010), while Section 2 specifies the details important here.

Second, the proposed two-sample tests are built using self-normalization, a recent concept for studentizing test statistics introduced originally for univariate time series in Shao (2010) and Shao and Zhang (2010). When conducting inference with time series data, one frequently encounters the problem of having to estimate the long-run variance in order to scale the fluctuations of test statistics. This is typically done using estimators that rely on tuning parameters that ideally should adjust to the strength of the autocorrelation present in the data. In practice, the success of such methods can vary widely. As a remedy, self-normalization is a tuning parameter-free method that achieves standardization, typically through recursive estimates. The advantages of such an approach for testing relevant hypotheses of parameters of functional time series, which can be defined directly in terms of expectations such as the mean or covariance operator, were recently recognized in Dette, Kokot and Volgushev (2020). However, because of its implicit definition, the direct application of these ideas for statistical inference of the eigensystem is not possible.

In this study, we develop a statistical methodology for the problem of testing relevant differences between the eigenfunctions of two covariance operators in functional data. The asymptotic variance of the common estimate of the difference of the eigenvalues depends on all other eigenvalues of both covariance operators in an intricate way (see Hall and Hosseini-Nasab (2006), among others) that is difficult, if not impossible, to estimate with sufficient precision. A similar comment applies to estimates of the eigenfunctions, making the concept

of self-normalization quite desirable in this context. To develop such a concept, we study the sequential processes of the eigenvalues and eigenfunctions of a sequential empirical covariance operator, for which we prove weak convergence. Self-normalized statistics can then be defined as continuous functionals of these processes.

Zhang and Shao (2015) is the work most closely related to our study, because it pertains to self-normalized two-sample tests for eigenfunctions and eigenvalues in functional time series. However, an important difference is that the methods proposed here do not require a dimension reduction of the eigenfunctions but compare the functions directly with respect to a norm in the L^2-space. A further crucial difference is that their work is in the classical testing setup, whereas ours is in the strictly relevant setting, so that the contributions are not directly comparable on the same footing; even though we report the outcomes from both tests on the same simulated curves in Section 3. There, we find that, despite the proposed test being constructed to detect relevant differences, it appears to compare favorably with the test of Zhang and Shao (2015) when the difference in the eigenfunctions is large. In this sense, both tests can be seen as complementing each other.

The rest of the paper is organized as follows. Section 2 introduces the framework, provides the model assumptions, and gives the two-sample test procedures and their theoretical properties. Section 3 reports the results of a comparative simulation study. In Section 4, we apply the proposed tests to Australian temperature curves obtained at different locations during the past century or so. Section 5 concludes the paper. Finally, some technical details used in the arguments of Section 2.2 are given in an online Supplementary Material.

2. Testing the Similarity of Two Eigenfunctions

Let $L^2([0,1])$ denote the common space of square integrable functions $f\colon [0,1] \to \mathbb{R}$ with inner product $\langle f_1, f_2\rangle = \int_0^1 f_1(t)f_2(t)dt$ and norm $\|f\| = (\int_0^1 f^2(t)dt)^{1/2}$. Consider two independent stationary functional time series $(X_t)_{t\in\mathbb{Z}}$ and $(Y_t)_{t\in\mathbb{Z}}$ in $L^2([0,1])$ and assume that each X_t and Y_t is centered and square integrable, that is $\mathbb{E}[X_t] = 0$, $\mathbb{E}[Y_t] = 0$ and $\mathbb{E}[\|X_t\|^2] < \infty$, $\mathbb{E}[\|Y_t\|^2] < \infty$, respectively. In practice centering can be achieved by subtracting the sample mean function estimate, which will not change our results. Denote by

$$C^X(s,t) = \sum_{j=1}^{\infty} \tau_j^X v_j^X(s) v_j^X(t), \tag{2.1}$$

$$C^Y(s,t) = \sum_{j=1}^{\infty} \tau_j^Y v_j^Y(s) v_j^Y(t) \tag{2.2}$$

the corresponding covariance operators; see Section 2.1 of Bücher, Dette and Heinrichs (2020) for a detailed discussion of expected values in Hilbert spaces. The eigenfunctions of the kernel integral operators with kernels C^X and C^Y, corresponding to the ordered eigenvalues $\tau_1^X \geq \tau_2^X \geq \cdots$ and $\tau_1^Y \geq \tau_2^Y \geq \cdots$, are denoted by $v_1^X, v_2^X, \ldots$ and $v_1^Y, v_2^Y, \ldots$, respectively. We are interested in testing the similarity of the covariance operators C^X and C^Y by comparing their eigenfunctions v_j^X and v_j^Y of order j for some $j \in \mathbb{N}$. This is framed as the relevant hypothesis testing problem

$$H_0^{(j)} \colon \|v_j^X - v_j^Y\|^2 \leq \Delta_j \qquad \text{versus} \qquad H_1^{(j)} \colon \|v_j^X - v_j^Y\|^2 > \Delta_j, \tag{2.3}$$

where $\Delta_j > 0$ is a prespecified constant representing the maximal value for the squared distances $\|v_j^X - v_j^Y\|^2$ between the eigenfunctions which can be accepted as scientifically insignificant. In order to make the comparison between the eigenfunctions meaningful, we assume throughout this paper that $\langle v_j^X, v_j^Y \rangle \geq 0$ for all $j \in \mathbb{N}$. The choice of the threshold $\Delta_j > 0$ depends on the specific application and is essentially defined by the change size one is really interested in from a scientific viewpoint. In particular, the choice $\Delta_j = 0$ gives the classical hypotheses $H_0^c \colon v_j^X = v_j^Y$ versus $H_1^c \colon v_j^X \neq v_j^Y$. We argue, however, that often it is well known that the eigenfunctions, or other parameters for that matter, from different samples will not precisely coincide. Further there is frequently no actual interest in arbitrarily small differences between the eigenfunctions. For this reason, $\Delta_j > 0$ is assumed throughout.

Observe also that a similar hypothesis testing problem could be formulated for relevant differences of the eigenvalues $\tau_j^X - \tau_j^Y$ of the covariance operators. We studied the development of such tests alongside those presented below for the eigenfunctions, and found, interestingly, that they generally are less powerful empirically. An elaboration and explanation of this is detailed in Remark 1 below. The arguments presented there are also applicable to tests based on direct long-run variance estimation.

The proposed approach is based on an appropriate estimate, say $\hat{D}_{m,n}^{(j)}$, of the squared L^2-distance $\|v_j^X - v_j^Y\|^2$ between the eigenfunctions, and the null hypothesis in (2.3) is rejected for large values of this estimate. It turns out that the

(asymptotic) distribution of this distance depends sensitively on all eigenvalues and eigenfunctions of the covariance operators C^X and C^Y and on the dependence structure of the underlying processes. To address this problem we propose a self-normalization of the statistic $\hat{D}_{m,n}^{(j)}$. Self-normalization is a well-established concept in the time series literature and was introduced in two seminal papers by Shao (2010) and Shao and Zhang (2010) for the construction of confidence intervals and change point analysis, respectively. More recently, it has been developed further for the specific needs of functional data by Zhang et al. (2011) and Zhang and Shao (2015); see also Shao (2015) for a recent review on self-normalization. In the present context, where one is interested in hypotheses of the form (2.3), a non-standard approach of self-normalization is necessary to obtain a distribution-free test, which is technically demanding due to the implicit definition of the eigenvalues and eigenfunctions of the covariance operators. For this reason, we first present the main idea of our approach in Section 2.1 and defer a detailed discussion to the subsequent Section 2.2.

2.1. Testing for relevant differences between eigenfunctions

If $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are the two samples, then

$$\hat{C}_m^X(s,t) = \frac{1}{m}\sum_{i=1}^{m} X_i(s)X_i(t), \quad \hat{C}_n^Y(s,t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(s)Y_i(t) \tag{2.4}$$

are the common estimates of the covariance operators (Ramsay and Silverman (2005); Horváth and Kokoszka (2012)). Denote by $\hat{\tau}_j^X, \hat{\tau}_j^Y$ and $\hat{v}_j^X, \hat{v}_j^Y$ the corresponding eigenvalues and eigenfunctions. Together, these define the canonical estimates of the respective population quantities in (2.1) and (2.2). See Chapter 3 of Horváth and Kokoszka (2012) for details on how to calculate these quantities in practice. Again, to make the comparison between the eigenfunctions meaningful, it is assumed throughout this paper that the inner product of $\langle \hat{v}_j^X, \hat{v}_j^Y \rangle$ is nonnegative for all j, which can be achieved in practice by changing the sign of one of the eigenfunction estimates if needed. We use the statistic

$$\hat{D}_{m,n}^{(j)} = \|\hat{v}_j^X - \hat{v}_j^Y\|^2 = \int_0^1 (\hat{v}_j^X(t) - \hat{v}_j^Y(t))^2 dt \tag{2.5}$$

to estimate the squared distance

$$D^{(j)} = \|v_j^X - v_j^Y\|^2 = \int_0^1 (v_j^X(t) - v_j^Y(t))^2 dt \tag{2.6}$$

between the jth population eigenfunctions. The null hypothesis will be rejected for large values of $\hat{D}^{(j)}_{m,n}$ compared to Δ_j. In the following, a self-normalized test statistic based on $\hat{D}^{(j)}_{m,n}$ will be constructed; see Dette, Kokot and Volgushev (2020). Specifically, let $\lambda \in [0,1]$ and define

$$\hat{C}^X_m(s,t,\lambda) = \frac{1}{\lfloor m\lambda \rfloor} \sum_{i=1}^{\lfloor m\lambda \rfloor} X_i(s)X_i(t), \quad \hat{C}^Y_n(s,t,\lambda) = \frac{1}{\lfloor n\lambda \rfloor} \sum_{i=1}^{\lfloor n\lambda \rfloor} Y_i(s)Y_i(t) \tag{2.7}$$

as the sequential version of the estimators in (2.4), noting that the sums are defined as 0 if $\lfloor m\lambda \rfloor < 1$. Observe that, under suitable assumptions detailed in Section 2.2, the statistics $\hat{C}^X_m(\cdot,\cdot,\lambda)$ and $\hat{C}^Y_n(\cdot,\cdot,\lambda)$ are consistent estimates of the covariance operators C^X and C^Y, respectively, whenever $0 < \lambda \leq 1$. The corresponding sample eigenfunctions of $\hat{C}^X_m(\cdot,\cdot,\lambda)$ and $\hat{C}^Y_n(\cdot,\cdot,\lambda)$ are denoted by $\hat{v}^X_j(t,\lambda)$ and $\hat{v}^Y_j(t,\lambda)$, respectively, assuming throughout that $\langle \hat{v}^X_j, \hat{v}^Y_j \rangle \geq 0$. Define the stochastic process

$$\hat{D}^{(j)}_{m,n}(t,\lambda) = \lambda(\hat{v}^X_j(t,\lambda) - \hat{v}^Y_j(t,\lambda)), \quad t \in [0,1]\ ,\ \lambda \in [0,1], \tag{2.8}$$

and note that the statistic $\hat{D}^{(j)}_{m,n}$ in (2.5) can be represented as

$$\hat{D}^{(j)}_{m,n} = \int_0^1 (\hat{D}^{(j)}_{m,n}(t,1))^2 dt. \tag{2.9}$$

Self-normalization is enabled through the statistic

$$\hat{V}^{(j)}_{m,n} = \left(\int_0^1 \left(\int_0^1 (\hat{D}^{(j)}_{m,n}(t,\lambda))^2 dt - \lambda^2 \int_0^1 (\hat{D}^{(j)}_{m,n}(t,1))^2 dt \right)^2 \nu(d\lambda) \right)^{1/2}, \tag{2.10}$$

where ν is a probability measure on the interval $(0,1]$. Note that, under appropriate assumptions, the statistic $\hat{V}^{(j)}_{m,n}$ converges to 0 in probability. However, it can be proved that its scaled version $\sqrt{m+n}\hat{V}^{(j)}_{m,n}$ converges in distribution to a random variable, which is positive with probability 1. More precisely, it is shown in Theorem 2 below that, under an appropriate set of assumptions,

$$\sqrt{m+n}\big(\hat{D}^{(j)}_{m,n} - D^{(j)}, \hat{V}^{(j)}_{m,n}\big) \xrightarrow{\mathcal{D}} \left(\zeta_j \mathbb{B}(1), \left\{ \zeta_j^2 \int_0^1 \lambda^2 (\mathbb{B}(\lambda) - \lambda \mathbb{B}(1))^2 \nu(d\lambda) \right\}^{1/2} \right) \tag{2.11}$$

as $m,n \to \infty$, where $D^{(j)}$ is defined in (2.6). Here $\{\mathbb{B}(\lambda)\}_{\lambda\in[0,1]}$ is a Brownian motion on the interval $[0,1]$ and $\zeta_j \geq 0$ is a constant, which is assumed to be strictly positive if $D^{(j)} > 0$ (the square ζ_j^2 is akin to a long-run variance

parameter). Consider then the test statistic

$$\hat{\mathbb{W}}_{m,n}^{(j)} := \frac{\hat{D}_{m,n}^{(j)} - \Delta_j}{\hat{V}_{m,n}^{(j)}}. \tag{2.12}$$

Based on this, the null hypothesis in (2.3) is rejected whenever

$$\hat{\mathbb{W}}_{m,n}^{(j)} > q_{1-\alpha}, \tag{2.13}$$

where $q_{1-\alpha}$ is the $(1-\alpha)$-quantile of the distribution of the random variable

$$\mathbb{W} := \frac{\mathbb{B}(1)}{\{\int_0^1 \lambda^2 (\mathbb{B}(\lambda) - \lambda \mathbb{B}(1))^2 \nu(d\lambda)\}^{1/2}}. \tag{2.14}$$

The quantiles of this distribution do not depend on the long-run variance, but on the measure ν in the statistic $\hat{V}_{m,n}^{(j)}$ used for self-normalization. An approximate P-value of the test can be calculated as

$$p = \mathbb{P}(\mathbb{W} > \hat{\mathbb{W}}_{m,n}^{(j)}). \tag{2.15}$$

The following theorem shows that the test just constructed keeps a desired level in large samples and has power increasing to one with the sample sizes.

Theorem 1. *If the weak convergence in* (2.11) *holds, then the test* (2.13) *has asymptotic level* α *and is consistent for the relevant hypotheses in* (2.3)*. In particular,*

$$\lim_{m,n\to\infty} \mathbb{P}(\hat{\mathbb{W}}_{m,n}^{(j)} > q_{1-\alpha}) = \begin{cases} 0 & \text{if } D^{(j)} < \Delta_j. \\ \alpha & \text{if } D^{(j)} = \Delta_j. \\ 1 & \text{if } D^{(j)} > \Delta_j. \end{cases} \tag{2.16}$$

Proof. If $D^{(j)} > 0$, the continuous mapping theorem and (2.11) imply

$$\frac{\hat{D}_{m,n}^{(j)} - D^{(j)}}{\hat{V}_{m,n}^{(j)}} \xrightarrow{\mathcal{D}} \mathbb{W}\ , \tag{2.17}$$

where the random variable $\mathbb{W}$ is defined in (2.14). Consequently, the probability of rejecting the null hypothesis is given by

$$\mathbb{P}(\hat{\mathbb{W}}_{m,n}^{(j)} > q_{1-\alpha}) = \mathbb{P}\bigg(\frac{\hat{D}_{m,n}^{(j)} - D^{(j)}}{\hat{V}_{m,n}^{(j)}} > \frac{\Delta_j - D^{(j)}}{\hat{V}_{m,n}^{(j)}} + q_{1-\alpha} \bigg). \tag{2.18}$$

It follows moreover from (2.11) that $\hat{V}_{m,n}^{(j)} \xrightarrow{\mathbb{P}} 0$ as $m, n \to \infty$ and therefore (2.17) implies (2.16), thus completing the proof in the case $D^{(j)} > 0$. If $D^{(j)} = 0$ it follows from the proof of (2.11) (see Proposition 2 below) that $\sqrt{m+n}\hat{D}_{m,n}^{(j)} = o_{\mathbb{P}}(1)$ and $\sqrt{m+n}\hat{V}_{m,n}^{(j)} = o_{\mathbb{P}}(1)$. Consequently,

$$\mathbb{P}(\hat{\mathbb{W}}_{m,n}^{(j)} > q_{1-\alpha}) = \mathbb{P}\big(\sqrt{m+n}\hat{D}_{m,n}^{(j)} > \sqrt{m+n}\Delta_j + \sqrt{m+n}\hat{V}_{m,n}^{(j)} q_{1-\alpha}\big) = o(1),$$

which completes the proof.

The main difficulty in the proof of Theorem 1 is hidden by postulating the weak convergence in (2.11). A proof of this statement is technically demanding. The precise formulation is given in the following section.

Remark 1. (Estimation of the long-run variance, power, and relevant differences in the eigenvalues).

(1) The parameter ζ_j^2 is essentially a long-run variance parameter. Therefore it is worthwhile to mention that on a first glance the weak convergence in (2.11) provides a very simple test for the hypotheses (2.3) if a consistent estimator, say $\hat{\zeta}_{n,j}^2$, of the long-run variance would be available. To this end, note that in this case it follows from (2.11) that $\sqrt{m+n}(\hat{D}_{m,n}^{(j)} - D^{(j)})/\hat{\zeta}_{n,j}$ converges weakly to a standard normal distribution. Consequently, using the same arguments as in the proof of Theorem 1,we obtain that rejecting the null hypothesis in (2.3), whenever

$$\frac{\sqrt{m+n}(\hat{D}_{m,n}^{(j)} - \Delta_j)}{\hat{\zeta}_{n,j}} > u_{1-\alpha}\ , \tag{2.19}$$

yields a consistent and asymptotic level α test. However a careful inspection of the representation of the long-run variance in equations (6.11)–(6.15) in the Supplementary Material suggests that it would be extremely difficult, if not impossible, to construct a reliable estimate of the parameter ζ_j in this context, due to its complicated dependence on the covariance operators C^X, C^Y, and their full complement of eigenvalues and eigenfunctions.

(2) Defining $\mathbb{K} = \big(\int_0^1 \lambda^2(\mathbb{B}(\lambda) - \lambda\mathbb{B}(1))^2\nu(d\lambda)\big)^{1/2}$, it follows from (2.18) that

$$P\big(\hat{\mathbb{W}}_{m,n}^{(j)} > q_{1-\alpha}\big) \approx P\bigg(\mathbb{W} > \frac{\sqrt{m+n}(\Delta_j - D_j)}{\zeta_j \cdot \mathbb{K}} + q_{1-\alpha}\bigg), \tag{2.20}$$

where the random variable $\mathbb{W}$ is defined in (2.14) and ζ_j is the long-run standard deviation appearing in Theorem 2, which is defined precisely in

equation (6.15) in the Supplementary Material. The probability on the right-hand side converges to zero, α, or 1, depending on $\Delta_j - D_j$ being negative, zero, or positive, respectively. From this one may also quite easily understand how the power of the test depends on ζ_j. Under the alternative, $\Delta_j - D_j < 0$ and the probability on the right-hand side of (2.20) increases if $(D_j - \Delta_j)/\zeta_j$ increases. Consequently, smaller long-run variances ζ_j^2 yield more powerful tests. Values of ζ_j are calculated via simulation for some of the examples in Section 3.

(3) Alongside the proposed test for relevant differences in the eigenfunctions, one might also consider the following test for relevant differences in the jth eigenvalues of the covariance operators C^X and C^Y:

$$H_{0,val}^{(j)}:\ D_{j,val} := (\tau_j^X - \tau_j^Y)^2 \leq \Delta_{j,val} \quad \text{versus} \quad H_{1,val}^{(j)} : (\tau_j^X - \tau_j^Y)^2 > \Delta_{j,val}. \tag{2.21}$$

Following the development of the above test for the eigenfunctions, a test of the hypothesis (2.21) can be constructed based on the partial sample estimates of the eigenvalues $\hat{\tau}_j^X(\lambda)$ and $\hat{\tau}_j^Y(\lambda)$ of the kernel integral operators with kernels $\hat{C}_m^X(\cdot,\cdot,\lambda)$ and $\hat{C}_n^Y(\cdot,\cdot,\lambda)$, respectively, in (2.7). In particular, let

$$\hat{T}_{m,n}^{(j)}(\lambda) = \lambda(\hat{\tau}_j^X(\lambda) - \hat{\tau}_j^Y(\lambda)), \text{ and}$$
$$\hat{M}_{m,n}^{(j)} = \left(\int_0^1 \{[\hat{T}_{m,n}^{(j)}(\lambda)]^2 - \lambda^2 [\hat{T}_{m,n}^{(j)}(1)]^2\}^2 \nu(d\lambda) \right)^{1/2}.$$

Then one can show, in fact somewhat more simply than in the case of the eigenfunctions, that the test procedure that rejects the null hypothesis whenever

$$\hat{\mathbb{Q}}_{m,n}^{(j)} = \frac{[\hat{T}_{m,n}^{(j)}(1)]^2 - \Delta_{j,val}}{\hat{M}_{m,n}^{(j)}} > q_{1-\alpha} \tag{2.22}$$

is a consistent and asymptotic level α test for the hypotheses (2.21). Moreover, the power of this test is approximately given by

$$P\big(\hat{\mathbb{Q}}_{m,n}^{(j)} > q_{1-\alpha}\big) \approx P\Big(\mathbb{W} > \frac{\sqrt{m+n}(\Delta_{j,val} - D_{j,val})}{\zeta_{j,val} \cdot \mathbb{K}} + q_{1-\alpha}\Big), \tag{2.23}$$

where $\zeta_{j,val}^2$ is a different long-run variance parameter. Although the tests (2.3) and (2.21) are constructed for completely different testing problems it might be of interest to compare their power properties. For this purpose note that the ratios $(D_j - \Delta_j)/\zeta_j$ and $(D_{j,val} - \Delta_{j,val})/\zeta_{j,val}$, for which the power of each test is an increasing function of, implicitly depend in a quite

complicated way on the dependence structure of the X and Y samples and on all eigenvalues and eigenfunctions of their corresponding covariance operators.

One might expect intuitively that relevant differences between the eigenvalues would be easier to detect than differences between the eigenfunctions (as the latter are more difficult to estimate). However, an empirical analysis shows that, in typical examples, the ratio $(D_{j,val} - \Delta_{j,val})/\zeta_{j,val}$ increases extremely slowly with increasing $D_{j,val}$ compared to the analogous ratio for the eigenfunction problem. Consequently, we expected and observed in numerical experiments (not presented for the sake of brevity) that the test (2.22) would be less powerful than the test (2.13) if in hypotheses (2.21) and (2.3) the thresholds $\Delta_{j,val}$ and Δ_j are similar. This observation also applies to the tests based on (intractable) long-run variance estimation. Here the power is approximately given by $1-\Phi\big(\sqrt{m+n}(\Delta_j - D)/z + u_{1-\alpha}\big)$, where Φ is the cdf of the standard normal distribution and z (and D) is either ζ_j (and $D^{(j)}$) for the test (2.19) or $\zeta_{j,val}$ (and $D_{j,val}$) for the corresponding test for the eigenvalues.

2.2. Justification of weak convergence

For the proof of (2.11) several technical assumptions are required. The first condition is standard in two-sample inference.

Assumption 1. *There exists a constant* $\theta \in (0,1)$ *such that* $\lim_{m,n\to\infty} m/(m+n) = \theta$.

Next, we specify the dependence structure of the time series $\{X_i\}_{i\in\mathbb{Z}}$ and $\{Y_i\}_{i\in\mathbb{Z}}$. Several mathematical concepts have been proposed for this purpose (see Bradley (2005); Bertail, Doukhan and Soulier (2006), among many others). Here, we use the general framework of L^p-m-approximability for weakly dependent functional data as put forward in Hörmann and Kokoszka (2010). Following these authors, a time series $\{X_i\}_{i\in\mathbb{Z}}$ in $L^2([0,1])$ is called *L^p-m-approximable* for some $p > 0$ if

(a) There exists a measurable function $g\colon S^\infty \to L^2([0,1])$, where S is a measurable space, and independent, identically distributed (iid) innovations $\{\epsilon_i\}_{i\in\mathbb{Z}}$ taking values in S such that $X_i = g(\epsilon_i, \epsilon_{i-1}, \ldots)$ for $i \in \mathbb{Z}$;

(b) Let $\{\epsilon_i'\}_{i\in\mathbb{Z}}$ be an independent copy of $\{\epsilon_i\}_{i\in\mathbb{Z}}$, and define $X_{i,m} = g(\epsilon_i, \ldots, \epsilon_{i-m+1}, \epsilon_{i-m}', \epsilon_{i-m-1}', \ldots)$. Then,

$$\sum_{m=0}^{\infty} \big(\mathbb{E}[\|X_i - X_{i,m}\|^p]\big)^{1/p} < \infty.$$

Assumption 2. *The sequences $\{X_i\}_{i\in\mathbb{Z}}$ and $\{Y_i\}_{i\in\mathbb{Z}}$ are independent, each centered and L^p-m-approximable for some $p > 4$.*

Under Assumption 2, there exist covariance operators C^X and C^Y of X_i and Y_i. For the corresponding eigenvalues $\tau_1^X \geq \tau_2^X \geq \cdots$ and $\tau_1^Y \geq \tau_2^Y \geq \cdots$, we assume the following, which guarantees that the eigenspaces corresponding to the leading eigenvalues of each covariance operator are one dimensional.

Assumption 3. *There exists a positive integer d such that $\tau_1^X > \cdots > \tau_d^X > \tau_{d+1}^X > 0$ and $\tau_1^Y > \cdots > \tau_d^Y > \tau_{d+1}^Y > 0$.*

The final assumption needed is a positivity condition on the long-run variance parameter ζ_j^2 appearing in (2.11). The formal definition of ζ_j is quite cumbersome, since it depends in a complicated way on expansions for the differences $\hat{v}_j^X(\cdot,\lambda) - v_j^X$ and $\hat{v}_j^Y(\cdot,\lambda) - v_j^Y$, but is provided in the Supplementary Material; see equations (6.11)–(6.15) therein.

Assumption 4. *The scalar ζ_j defined in* (6.15) *of the Supplementary Material is strictly positive, whenever $D^{(j)} > 0$.*

Recall the definition of the sequential processes $\hat{C}^X(\cdot,\cdot,\lambda)$ and $\hat{C}^Y(\cdot,\cdot,\lambda)$ in (2.7) and their corresponding eigenfunctions $\hat{v}_j^X(\cdot,\lambda)$ and $\hat{v}_j^Y(\cdot,\lambda)$. The first step in the proof of the weak convergence (2.11) is a stochastic expansion of the difference between the sample eigenfunctions $\hat{v}_j^X(\cdot,\lambda)$ and $\hat{v}_j^Y(\cdot,\lambda)$ and their respective population versions v_j^X and v_j^Y. Similar expansions that do not take into account uniformity in the partial sample parameter λ have been derived by Kokoszka and Reimherr (2013) and Hall and Hosseini-Nasab (2006), among others; see also Dauxois, Pousse and Romain (1982) for a general statement in this context. The proof of this result is contained in the Supplementary Material.

Proposition 1. *Suppose Assumptions* 2 *and* 3 *hold. Then, for any $j \leq d$,*

$$\begin{aligned}
&\sup_{\lambda\in[0,1]} \Bigg\| \lambda[\hat{v}_j^X(t,\lambda) - v_j^X(t)] \\
&- \frac{1}{\sqrt{m}} \sum_{k\neq j} \frac{v_k^X(t)}{\tau_j^X - \tau_k^X} \int_0^1\int_0^1 \hat{Z}_m^X(s_1,s_2,\lambda) v_k^X(s_2) v_j^X(s_1) ds_1 ds_2 \Bigg\| \\
&= O_{\mathbb{P}}\left(\frac{\log^{\kappa}(m)}{m}\right), \qquad (2.24)
\end{aligned}$$

and

$$\sup_{\lambda\in[0,1]} \Big\| \lambda[\hat{v}_j^Y(t,\lambda) - v_j^Y(t)]$$

$$- \frac{1}{\sqrt{m}} \sum_{k \neq j} \frac{v_k^Y(t)}{\tau_j^Y - \tau_k^Y} \int_0^1 \int_0^1 \hat{Z}_n^Y(s_1, s_2, \lambda) v_k^Y(s_2) v_j^Y(s_1) ds_1 ds_2 \Bigg\|$$
$$= O_{\mathbb{P}}\left(\frac{\log^{\kappa}(n)}{n} \right), \tag{2.25}$$

for some $\kappa > 0$, where the processes $\hat{Z}_m^X$ and $\hat{Z}_n^Y$ are defined by

$$\hat{Z}_m^X(s_1, s_2, \lambda) = \frac{1}{\sqrt{m}} \sum_{i=1}^{\lfloor m\lambda \rfloor} \big(X_i(s_1) X_i(s_2) - C^X(s_1, s_2) \big), \tag{2.26}$$

$$\hat{Z}_n^Y(s_1, s_2, \lambda) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor n\lambda \rfloor} \big(Y_i(s_1) Y_i(s_2) - C^Y(s_1, s_2) \big). \tag{2.27}$$

Moreover,

$$\sup_{\lambda \in [0,1]} \sqrt{\lambda} \big\| \hat{v}_j^X(\cdot, \lambda) - v_j^X \big\| = O_{\mathbb{P}}\left(\frac{\log^{(1/\kappa)}(m)}{\sqrt{m}} \right), \tag{2.28}$$

$$\sup_{\lambda \in [0,1]} \sqrt{\lambda} \big\| \hat{v}_j^Y(\cdot, \lambda) - v_j^Y \big\| = O_{\mathbb{P}}\left(\frac{\log^{(1/\kappa)}(n)}{\sqrt{n}} \right). \tag{2.29}$$

Recalling notation (2.8), Proposition 1 motivates the approximation

$$\hat{D}_{m,n}^{(j)}(t, \lambda) - \lambda D_j(t) = \lambda(\hat{v}_j^X(t) - \hat{v}_j^Y(t)) - \lambda(v_j^X(t) - v_j^Y(t)) \approx \tilde{D}_{m,n}^{(j)}(t, \lambda), \tag{2.30}$$

where the process $\tilde{D}_{m,n}^{(j)}$ is defined by

$$\tilde{D}_{m,n}^{(j)}(t, \lambda) = \frac{1}{\sqrt{m}} \sum_{k \neq j} \frac{v_k^X(t)}{\tau_j^X - \tau_k^X} \int_0^1 \int_0^1 \hat{Z}_m^X(s_1, s_2, \lambda) v_k^X(s_2) v_j^X(s_1) ds_1 ds$$
$$- \frac{1}{\sqrt{n}} \sum_{k \neq j} \frac{v_k^Y(t)}{\tau_j^Y - \tau_k^Y} \int_0^1 \int_0^1 \hat{Z}_n^Y(s_1, s_2, \lambda) v_k^Y(s_2) v_j^Y(s_1) ds_1, ds_2. \tag{2.31}$$

The next result makes the foregoing heuristic arguments rigorous and shows that the approximation holds in fact uniformly with respect to $\lambda \in [0, 1]$.

Proposition 2. *Suppose Assumptions 1–4 hold. Then, for any $j \leq d$,*

$$\sup_{\lambda \in [0,1]} \Big\| \hat{D}_{m,n}^{(j)}(\cdot, \lambda) - \lambda D_j(\cdot) - \tilde{D}_{m,n}^{(j)}(\cdot, \lambda) \Big\| = o_{\mathbb{P}}\left(\frac{1}{\sqrt{m+n}} \right),$$
$$\sup_{\lambda \in [0,1]} \Big| \Big\| \hat{D}_{m,n}^{(j)}(\cdot, \lambda) - \lambda D_j(\cdot) \Big\|^2 - \Big\| \tilde{D}_{m,n}^{(j)}(\cdot, \lambda) \Big\|^2 \Big| = o_{\mathbb{P}}\left(\frac{1}{\sqrt{m+n}} \right),$$

and

$$\sqrt{m+n}\sup_{\lambda\in[0,1]}\int_0^1(\tilde{D}^{(j)}_{m,n}(t,\lambda))^2dt=o_{\mathbb{P}}(1). \tag{2.32}$$

Proof. According to their definitions,

$$\begin{aligned}
&\hat{D}^{(j)}_{m,n}(t,\lambda)-\lambda D_j(t)-\tilde{D}^{(j)}_{m,n}(t,\lambda)\\
&=\lambda[\hat{v}^X_j(t,\lambda)-v^X_j(t)]\\
&\quad-\frac{1}{\sqrt{m}}\sum_{k\neq j}\frac{v^X_k(t)}{\tau^X_j-\tau^X_k}\int_0^1\int_0^1\hat{Z}^X_m(s_1,s_2,\lambda)v^X_k(s_2)v^X_j(s_1)ds_1ds_2\\
&\quad+\lambda[\hat{v}^Y_j(t,\lambda)-v^Y_j(t)]\\
&\quad-\frac{1}{\sqrt{m}}\sum_{k\neq j}\frac{v^Y_k(t)}{\tau^Y_j-\tau^Y_k}\int_0^1\int_0^1\hat{Z}^Y_n(s_1,s_2,\lambda)v^Y_k(s_2)v^Y_j(s_1)ds_1ds_2.
\end{aligned}$$

Therefore, by the triangle inequality, Proposition 1, and Assumption 1,

$$\begin{aligned}
&\sup_{\lambda\in[0,1]}\left\|\hat{D}^{(j)}_{m,n}(\cdot,\lambda)-\lambda D_j(\cdot)-\tilde{D}^{(j)}_{m,n}(\cdot,\lambda)\right\|\\
&\leq\sup_{\lambda\in[0,1]}\Big\|\lambda[\hat{v}^X_j(t,\lambda)-v^X_j(t)]\\
&\quad-\frac{1}{\sqrt{m}}\sum_{k\neq j}\frac{v^X_k(t)}{\tau^X_j-\tau^X_k}\int_0^1\int_0^1\hat{Z}^X_m(s_1,s_2,\lambda)v^X_k(s_2)v^X_j(s_1)ds_1ds_2\Big\|\\
&\quad+\sup_{\lambda\in[0,1]}\Big\|\lambda[\hat{v}^Y_j(t,\lambda)-v^Y_j(t)]\\
&\quad-\frac{1}{\sqrt{m}}\sum_{k\neq j}\frac{v^Y_k(t)}{\tau^Y_j-\tau^Y_k}\int_0^1\int_0^1\hat{Z}^Y_n(s_1,s_2,\lambda)v^Y_k(s_2)v^Y_j(s_1)ds_1ds_2\Big\|\\
&=O_{\mathbb{P}}\left(\frac{\log^{\kappa}(m)}{m}\right)=o_{\mathbb{P}}\left(\frac{1}{\sqrt{m+n}}\right).
\end{aligned}$$

The second assertion follows immediately from the first and the reverse triangle inequality. With the second assertion in place, we have, using (2.28) and (2.29), that

$$\begin{aligned}
&\sqrt{m+n}\sup_{\lambda\in[0,1]}\int_0^1(\tilde{D}^{(j)}_{m,n}(t,\lambda))^2dt\\
&=\sqrt{m+n}\sup_{\lambda\in[0,1]}\int_0^1(\hat{D}^{(j)}_{m,n}(t,\lambda)-\lambda D_j(t))^2dt+o_{\mathbb{P}}(1)\\
&\leq 4\sqrt{m+n}\left[\sup_{\lambda\in[0,1]}\lambda^2\|\hat{v}^X_j(\cdot,\lambda)-v^X_j\|^2+\sup_{\lambda\in[0,1]}\lambda^2\|\hat{v}^Y_j(\cdot,\lambda)-v^Y_j\|^2\right]
\end{aligned}$$

$$= O_{\mathbb{P}}\Big(\frac{\log^{(2/\kappa)}(m)}{\sqrt{m}}\Big) = o_{\mathbb{P}}(1)$$

which completes the proof.

Introduce the process

$$\hat{Z}^{(j)}_{m,n}(\lambda) = \sqrt{m+n}\int_0^1 ((\hat{D}^{(j)}_{m,n}(t,\lambda))^2 - \lambda^2 D_j^2(t))dt \tag{2.33}$$

to obtain the following result. The proof is somewhat complicated and therefore deferred to Supplementary Material.

Proposition 3. *Let $\hat{Z}^{(j)}_{m,n}$ be defined by* (2.33)*, then, under Assumptions* 1-4 *we have for any $j \le d$,*

$$\{\hat{Z}^{(j)}_{m,n}(\lambda)\}_{\lambda\in[0,1]} \rightsquigarrow \{\lambda\zeta_j\mathbb{B}(\lambda)\}_{\lambda\in[0,1]},$$

where ζ_j is a positive constant, $\{\mathbb{B}(\lambda)\}_{\lambda\in[0,1]}$ is a Brownian motion and $\rightsquigarrow$ denotes weak convergence in Skorokhod topology on $D[0,1]$.

Theorem 2. *If Assumptions* 1, 2 *and* 3 *are satisfied, then for any $j \le d$*

$$\sqrt{m+n}\big(\hat{D}^{(j)}_{m,n} - D^{(j)}, \hat{V}^{(j)}_{m,n}\big) \rightsquigarrow \left(\zeta_j\mathbb{B}(1), \left\{\zeta_j^2\int_0^1 \lambda^2(\mathbb{B}(\lambda) - \lambda\mathbb{B}(1))^2\nu(d\lambda)\right\}^{1/2}\right),$$

where $\hat{D}^{(j)}_{m,n}$ and $\hat{V}^{(j)}_{m,n}$ are defined by (2.9) *and* (2.10)*, respectively, and $\{\mathbb{B}(\lambda)\}_{\lambda\in[0,1]}$ is a Brownian motion.*

Proof. Observing the definition of $\hat{D}^{(j)}_{m,n}$, $D^{(j)}$, $\hat{Z}^{(j)}_{m,n}$ and $\hat{V}^{(j)}_{m,n}$ in (2.9), (2.6) and (2.33) and (2.10), we have

$$\begin{aligned}
\hat{D}^{(j)}_{m,n} - D^{(j)} &= \int_0^1 (\hat{D}_{m,n}(t,1))^2 dt - \int_0^1 D_j^2(t)dt = \frac{\hat{Z}^{(j)}_{m,n}(1)}{\sqrt{m+n}},\\
\hat{V}^{(j)}_{m,n} &= \left\{\int_0^1 \left(\int_0^1 \big[(\hat{D}^{(j)}_{m,n}(t,\lambda))^2 - \lambda^2 D_j^2(t)\big]dt\right.\right.\\
&\qquad \left.\left. -\lambda^2\int_0^1 \big[(\hat{D}^{(j)}_{m,n}(t,1))^2 - D_j^2(t)\big]dt\right)^2 \nu(d\lambda)\right\}^{1/2}\\
&= \frac{1}{\sqrt{m+n}}\left\{\int_0^1 \big(\hat{Z}^{(j)}_{m,n}(\lambda) - \lambda^2\hat{Z}^{(j)}_{m,n}(1)\big)^2\nu(d\lambda)\right\}^{1/2}.
\end{aligned}$$

The assertion now follows directly from Proposition 3 and the continuous mapping theorem.

Remark 2 (Relaxing the assumption of independence between the samples)**.** Often in applications with real functional time series the assumption that the samples of the X_i and Y_i processes are independent is either in question, or is overly restrictive. It can be shown though that if the two samples are obtained by observing a jointly stationary "bivariate" functional time series (X_i, Y_i) satisfying a version of Assumption 2, then Theorem 2 in fact still holds, but with a different constant ζ_j. Since this constant is not estimated in applying the self-normalized test, one still obtains a consistent test satisfying the properties of Theorem 1 in this case. This observation is relevant to interpreting the results of our application to Australian temperature profiles below. Another important instance in which this assumption can be relaxed is when the two samples are constructed by partitioning a single weakly dependent time series. For example, suppose W_i, $i \in \{1, \ldots, n\}$ is an observed stretch of a series satisfying Assumption 2. If $X_i = W_i$, $1 \leq i \leq \lfloor n\theta \rfloor$, and $Y_i = W_{i+\lfloor n\theta \rfloor}$, $1 \leq i \leq n - \lfloor n\theta \rfloor$, then Theorem 2 still holds. This setting might arise if one is interested in evaluating whether the eigenfunctions of the covariance operator of the W_i series exhibit relevant differences before or after the point $\lfloor n\theta \rfloor$. If interested, the assumption of independence between two concurrently observed functional time series can be evaluated using the tests developed in Horváth and Rice (2015). The stationarity of such series can be investigated by applying the methods in Aue and van Delft (2020).

2.3. Testing for relevant differences in multiple eigenfunctions

In this subsection, we are interested in testing if there is no relevant difference between several eigenfunctions of the covariance operators C^X and C^Y. To be precise, let $j_1 < \cdots < j_p$ denote positive indices defining the orders of the eigenfunctions to be compared. This leads to testing the hypotheses

$$H_{0,p}\colon D^{(j_\ell)} = \|v_{j_\ell}^X - v_{j_\ell}^Y\|_2^2 \leq \Delta_\ell \quad \text{for all } \ell \in \{1, \ldots, p\}, \tag{2.34}$$

versus

$$H_{1,p}\colon D^{(j_\ell)} = \|v_{j_\ell}^X - v_{j_\ell}^Y\|_2^2 > \Delta_\ell \quad \text{for at least one } \ell \in \{1, \ldots, p\}, \tag{2.35}$$

where $\Delta_1, \ldots, \Delta_p > 0$ are pre-specified constants.

After trying a number of methods to perform such a test, including deriving joint asymptotic results for the vector of pairwise distances $\hat{D}_{m,n} = \big(\hat{D}_{m,n}^{(j_1)}, \ldots,$

$\hat{D}_{m,n}^{(j_p)}\big)^\top$, and using these to perform confidence region-type tests as described in Aitchison (1964), we ultimately found that the best approach for relatively small p was to simply apply the marginal tests as proposed above to each eigenfunction, and then control the family-wise error rate using a Bonferroni correction. Specifically, suppose $P_{j_1},\ldots,P_{j_p}$ are P-values of the marginal relevant difference in eigenfunction tests calculated from (2.15). Then, under the null hypothesis $H_{0,p}$ in (2.34) is rejected at level α if $P_{j_k} < \alpha/p$ for some k between 1 and p. This asymptotically controls the overall type one error to be less than α. A similar approach is the Bonferroni method with Holm correction; see Holm (1979). These methods are investigated by simulation in Section 3.1 below.

3. Simulation Study

A simulation study was conducted to evaluate the finite-sample performance of the tests described in (2.3), as well as a comparison with the self-normalized two-sample test introduced in Zhang and Shao (2015), hereafter referred to as the ZS test. However, it should be emphasized that their test is for the classical hypothesis

$$H_{0,class}\colon \|v_j^X - v_j^Y\|^2 = 0 \qquad \text{versus} \qquad H_{1,class}^{(j)}\colon \|v_j^X - v_j^Y\|^2 > 0, \tag{3.1}$$

and not for the relevant hypotheses (2.3) studied here. Such a comparison is nevertheless useful to demonstrate that both procedures behave similarly in the different testing problems. All simulations below were performed using the `R` programming language (R Core Team (2016)). Data were generated according to the basic model proposed and studied in Panaretos, Kraus and Maddocks (2010) and Zhang and Shao (2015), which is of the form

$$X_i(t) = \sum_{j=1}^{2} \left\{ \xi_{X,j,1}^{(i)} \sqrt{2} \sin\left(2\pi j t + \delta_j\right) + \xi_{X,j,2}^{(i)} \sqrt{2} \cos\left(2\pi j t + \delta_j\right) \right\}, \quad t \in [0,1], \tag{3.2}$$

for $i = 1,\ldots,m$, where the coefficients $\xi_{X,i} = (\xi_{X,1,1}^{(i)}, \xi_{X,2,1}^{(i)}, \xi_{X,1,2}^{(i)}, \xi_{X,2,2}^{(i)})'$ follow a vector autoregressive model

$$\xi_{X,i} = \rho \xi_{X,i-1} + \sqrt{1-\rho^2} e_{X,i},$$

with $\rho = 0.5$ and $e_{X,i} \in \mathbb{R}^4$ a sequence of iid normal random variables with mean zero and covariance matrix

$$\Sigma_e = \operatorname{diag}(\mathbf{v_X}),$$

with $\mathbf{v_X} = (\tau_1^X, \ldots, \tau_4^X)$. Note that with this specification, the population level eigenvalues of the covariance operator of X_i are $\tau_1^X, \ldots, \tau_4^X$. If $\delta_1 = \delta_2 = 0$, the corresponding eigenfunctions are $v_1^X = \sqrt{2}\sin(2\pi\cdot)$, $v_2^X = \sqrt{2}\cos(2\pi\cdot)$, $v_3^X = \sqrt{2}\sin(4\pi\cdot)$, and $v_4^X = \sqrt{2}\cos(4\pi\cdot)$. Each process X_i was produced by evaluating the right-hand side of (3.2) at 1,000 equally spaced points in the unit interval, and then smoothing over a cubic B-spline basis with 20 equally spaced knots using the `fda` package; see Ramsay, Hooker and Graves (2009). A burn-in sample of length 30 was generated and discarded to produce the autoregressive processes. The sample Y_i, $i = 1, \ldots, n$, was generated independently in the same way, always choosing $\delta_j = 0$, $j = 1, 2$, in (3.2). With this setup, one can produce data satisfying either $H_0^{(j)}$ or $H_1^{(j)}$ by changing the constants δ_j.

In order to measure the finite sample properties of the proposed test for the hypotheses $H_0^{(j)}$ versus $H_1^{(j)}$ in (2.3) , data was generated as described above from two scenarios:

- Scenario 1: $\tau_{\mathbf{X}} = \tau_{\mathbf{Y}} = (8, 4, 0.5, 0.3)$, $\delta_2 = 0$, and δ_1 varying from 0 to 0.25.
- Scenario 2: $\tau_{\mathbf{X}} = \tau_{\mathbf{Y}} = (8, 4, 0.5, 0.3)$, $\delta_1 = 0$, and δ_2 varying from 0 to 2.

In both cases, we tested the hypotheses (2.3) with $\Delta_j = 0.1$, for $j = 1, 2, 3$. We took the measure ν, used to define the self-normalizing sequence in (2.10), to be the uniform probability measure on the interval $(0.1, 1)$. We also tried other values between 0 and 0.2 for the lower bound of this uniform measure and found that selecting values above 0.05 tended to yield similar performance. In general we recommend evaluating the sensitivity of the conclusions of the test to the choice of the measure ν, but our simulation evidence to date suggests these tests are generally robust to this choice. When $\delta_1 \approx 0.05$, $\|v_j^X - v_j^Y\|_2^2 \approx 0.1$, and taking $\delta_1 = 0.25$ causes v_j^X and v_j^Y to be orthogonal, $j = 1, 2$. Hence the null hypothesis $H_0^{(j)}$ holds for $\delta_1 < 0.05$, and $H_1^{(j)}$ holds for $\delta_1 > 0.05$ for $j = 1, 2$. Similarly, in Scenario 2, one has that $\|v_j^X - v_j^Y\|_2^2 \approx 0.1$ when $\delta_2 = 0.3155$, $j = 3, 4$. For this reason, we let δ_2 vary from 0 to 2. In reference to Remark 1, we obtained via simulation that the parameter ζ_j for the largest eigenvalue process is approximately 4 when $\delta_1 = 0$ and approximately 10.5 when $\delta_1 = 0.25$.

The percentage of rejections from 1,000 independent simulations when the size of the test is fixed at 0.05 are reported in Figures 1 and 2 as power curves that are functions of δ_1 and δ_2 when $n = m = 50$ and 100. These figures also display the number of rejections of the ZS test for the classical hypothesis (3.1) (which corresponds to $H_0^{(j)}$ with $\Delta_j = 0$). From this, the following conclusions can be drawn.

1. The tests of $H_0^{(j)}$ based on $\hat{\mathbb{W}}_{m,n}^{(1)}$ exhibited the behaviour predicted by (2.16), even for relatively small values of n and m. Focusing on the tests of $H_0^{(j)}$ with results displayed in Figure 1, we observed that the empirical rejection rate was less than nominal for $\|v_1^X - v_1^Y\|^2 < \Delta_1 = 0.1$, approximately nominal when $\|v_1^X - v_1^Y\|^2 = \Delta_1 = 0.1$, and the power increased as $\|v_1^X - v_1^Y\|^2$ began to exceed 0.1. In additional simulations not reported here, these results were improved further by taking larger values of n and m.

2. Observe that with data generated according to Scenario 2, $H_0^{(2)}$ is satisfied while $H_0^{(3)}$ is not satisfied for values of $\delta_2 > 0.3155$. This scenario is seen in Figure 2 where the tests for $H_0^{(2)}$ exhibited less than nominal size, as predicated by (2.16), even in the presence of differences in higher-order eigenfunctions. The tests of $H_0^{(3)}$ performed similarly to the tests of $H_0^{(1)}$.

3. The self-normalized ZS test for the classical hypothesis (3.1), which is based on the bootstrap, performed well in our simulations, exhibiting empirical size approximately equal to the nominal size when $\|v_1^X - v_1^Y\|^2 = 0$, and increasing power as $\|v_1^X - v_1^Y\|^2$ increased. For the sample size $m = n = 50$ it overestimated the nominal level of 5%. Interestingly, the proposed tests tended to exhibit higher power than the ZS test for large values of $\|v_1^X - v_1^Y\|^2$, even while only testing for relevant differences. Additionally, the computational time required to perform the proposed test is substantially less than what is required to perform the ZS test, since it does not need to employ the bootstrap.

3.1. Multiple comparisons

In order to investigate the multiple testing procedure of Section 2.3, X and Y samples were generated according to model (3.2) with $n = m = 100$ in two situations: one with $\delta_1 = 0.0504915$ and $\delta_2 = 0.3155$, and another with $\delta_1 = 0.25$ and $\delta_2 = 2$. In the former case, $\|v_j^X - v_j^Y\|^2 \approx 0.1$ for $j = 1, \ldots, 4$, while in the latter case $\|v_j^X - v_j^Y\|^2 > 0.1$, $j = 1, \ldots, 4$. We then applied tests of $H_{0,p}$ in (2.34) with $\Delta_j = 0.1$ for $j = 1, \ldots, 4$ and varied $p = 1, \ldots, 4$. These tests were carried out by combining the marginal tests for relevant differences of the respective eigenfunctions using the standard Bonferroni correction as well as the Holm–Bonferroni correction. Empirical size and power calculated from 1,000 simulations with nominal size 0.05 for each value of p and correction are reported in Table 1, as well as for the case when neither of these corrections are applied, and we simply reject $H_{0,p}$ if any of the marginal p-values are less

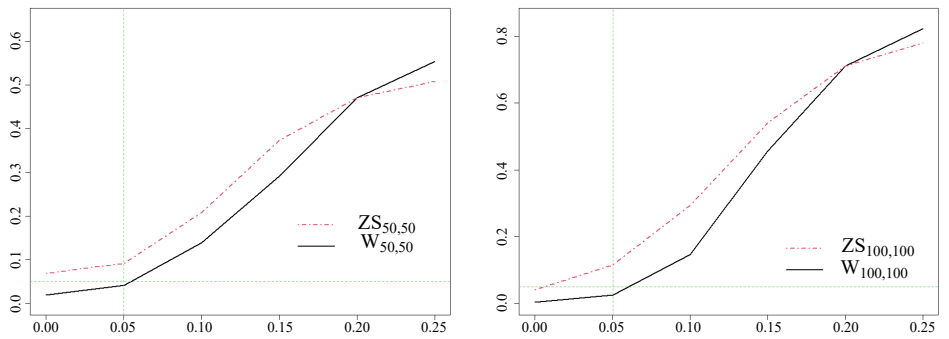

Figure 1. Percentage of rejections (out of 1,000 simulations) of the self-normalized statistic of Zhang and Shao (2015) for the classical hypotheses (3.1) (denoted $ZS_{n,m}^{(1)}$) and the new test (2.13) for the relevant hypotheses (2.3) (denoted by $W_{m,n}^{(1)}$) as a function of δ_1 in Scenario 1. In the left hand panel $n = m = 50$, and in the right hand panel $n = m = 100$. The horizontal green line is at the nominal level 0.05, and the vertical green line at $\delta_1 = 0.05$ indicates the case when $\|v_1^X - v_1^Y\|^2 = \Delta_1 = 0.1$.

.

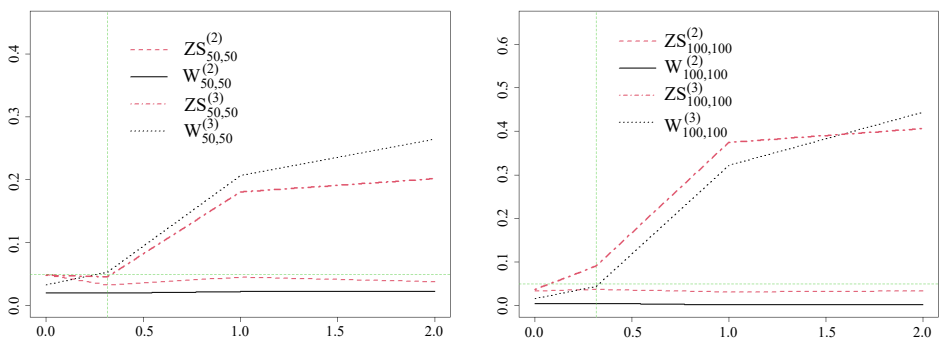

Figure 2. Percentage of rejections (out of 1,000 simulations) of the self-normalized statistic of Zhang and Shao (2015) for the classical hypotheses (3.1) (denoted $ZS_{n,m}^{(j)}$, $j = 2, 3$) and the new test (2.13) for the relevant hypohteses (2.3) (denoted by $W_{m,n}^{(j)}$, $j = 2, 3$) as a function of δ_2 in Scenario 2. In the left hand panel $n = m = 50$, and in the right hand panel $n = m = 100$. The horizontal green line is at the nominal level 0.05, and the vertical green line at $\delta_2 = 0.3155$ indicates the case when $\|v_3^X - v_3^Y\|^2 = \Delta_1 = 0.1$.

than the nominal level 0.05. It can be seen that these corrections controlled the family-wise error rate well. The tests still retain similar power when comparing up to four eigenfunctions, although one may notice the expected declining power as the number of tests increases.

Table 1. Rejection rates from 1,000 simulations of tests $H_{0,j}$ with nominal level 0.05 for $j = 1, \ldots, 4$ using the Bonferroni correction (B), Holm–Bonferroni correction (HB), and no correction (NC).

δ_1	δ_2		$j=1$	2	3	4
0.0504915	0.3155	B	0.039	0.019	0.032	0.029
		HB	0.039	0.037	0.035	0.032
		NC	0.039	0.043	0.075	0.075
0.25	2	B	0.817	0.715	0.708	0.660
		HB	0.817	0.817	0.794	0.734
		NC	0.817	0.817	0.890	0.891

Table 2. Locations and names of six measuring stations at which annual temperature data was recorded, and respective observation periods. The numbers of available annual temperature profiles are shown in parentheses. The 1932 and 1970 curves were removed from the Boulia series due to missing values.

Location	Years	
Sydney, Observatory Hill	1860–2011	(151)
Melbourne, Regional Office	1856–2011	(155)
Boulia, Airport*	1900–2009	(107)
Gayndah, Post Office	1905–2008	(103)
Hobart, Ellerslie Road	1896–2011	(115)
Robe	1885–2011	(126)

4. Data Illustration: Australian Annual Temperature Profiles

To illustrate the practical use of the proposed tests, we apply them to annual minimum temperature profiles. These functions were constructed from data collected at various measuring stations across Australia. The raw data consisted of approximately daily minimum temperature measurements recorded in degrees Celsius over approximately the last 150 years at six stations, and is available in the supplementary material to this article, as well as from `www.bom.gov.au`. The exact station locations and time periods considered are summarized in Table 2. In addition, Figure 3 provides a map of eastern Australia showing the relative locations of these stations.

In each year and for each station, 365 (366 in leap years) raw data points were converted into functional data objects using cubic B-splines at 20 equally spaced knots using the `fda` package in `R`; see Ramsay, Hooker and Graves (2009) for details. We also tried using cubic B-splines with between 20 and 40 equally spaced knots, as well as using the same numbers of standard Fourier basis elements to smooth the raw data into functional data objects, and the test results reported below were essentially unchanged. The resulting curves from the sta-

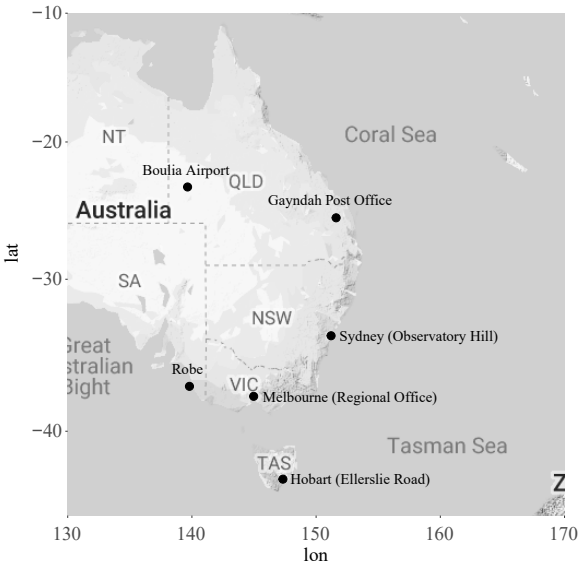

Figure 3. Map of eastern Australia showing the locations of the six measuring stations whose data were used in the data analysis. This map was produced using the `ggmap` package in `R`; see Kahle and Wickham (2013).

tions located in Sydney and Gayndah are displayed respectively in the left and right hand panels of Figure 4 as rainbow plots, with earlier curves drawn in red and progressing through the color spectrum to later curves drawn in violet; see Shang and Hyndman (2016). One may notice that the curves appear to generally increase in level over the years. In order to remove this trend, a linear time trend was estimated for the series of average yearly minimum temperatures, and then this linear trend was subtracted pointwise from the time series of curves. The detrended Sydney and Gayndah curves are displayed again as rainbow plots in the left and right-hand panels of Figure 5, which appear to be stationary; see Remark 2.

We took as the goal of the analysis to evaluate whether or not there are relevant differences in the primary modes of variability of these curves between locations, as measured by differences in the leading eigenfunctions of the sample covariance operators. We applied tests of $H_0^{(1)}$ and $H_0^{(2)}$ with thresholds $\Delta_1 = \Delta_2 = 0.1$ based on the statistics $\hat{\mathbb{W}}_{m,n}^{(j)}$, $j = 1, 2$, to each pair of functional time series from the six stations. We took the measure ν in the definition of $\hat{\mathbb{W}}_{m,n}^{(j)}$ to be the uniform probability measure on the interval $(0.1, 1)$. We also recomputed each test when this measure was uniform on $(0.05, 1)$, and on $(0.2, 1)$, and the

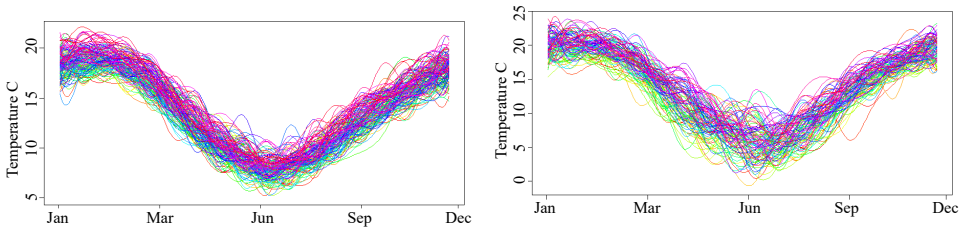

Figure 4. Rainbow plots of minimum temperature profiles based on data collected at the Sydney (left panel) and Gayndah (right panel) stations constructed using cubic B-splines.

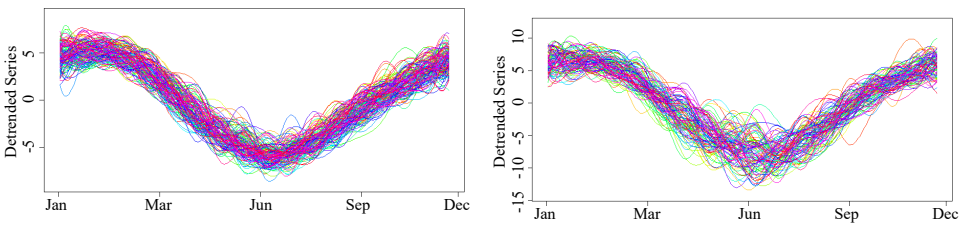

Figure 5. Rainbow plots of detrended minimum temperature profiles from Sydney (left panel) and Ganydah (right panel). Detrending was carried out by fitting a linear time trend to the series of average yearly minimum temperatures, and then removing this trend pointwise from the time series of curves.

difference between the results was negligible, so we only report results in the first case. The results of these tests are reported in terms of P-values in Table 3. Plots of the estimated leading eigenfunctions from each sample are displayed in Figure 6.

One observes in five out of six stations, excluding the Gayndah station, that the leading eigenfunction of the sample covariances operators is approximately constant, suggesting that the primary mode of variability of those temperature profiles is essentially level fluctuations around the increasing trend. Pairwise comparisons based on tests of $H_0^{(1)}$ suggest that these functions in general do not exhibit relevant differences to any reasonable significance. In contrast, the leading eigenfunction calculated from the Gayndah station curves evidently puts more mass in the winter months than the summer months. This is almost expected given the comparison of the detrended curves in Figure 5, in which the Gayndah curves evidently exhibit more variability in the winter months relative to the Sydney curves. Pairwise comparisons of the Gayndah data with the other stations suggest that this difference is significant, and even that the change is relevant to the level $\Delta_1 = 0.1$. The analysis of the second eigenfunction leads to a similar

Table 3. Approximate P-values of tests of $H_0^{(1)}$ and $H_0^{(2)}$ with $\Delta_1 = \Delta_2 = 0.1$ for all pairwise comparisons of the series of curves from each of the six monitoring stations. Values that are less than 0.05 are **bolded**.

$H_0^{(1)},\ \Delta_1 = 0.1$					
	Melbourne	Boulia	Gayndah	Hobart	Robe
Sydney	0.2075	0.4545	**0.0327**	0.2211	0.5614
Melbourne		0.1450	**0.0046**	0.5007	0.2203
Boulia			**0.0466**	**0.0321**	0.5419
Gayndah				**0.0002**	**0.0011**
Hobart					0.0885
$H_0^{(2)},\ \Delta_2 = 0.1$					
	Melbourne	Boulia	Gayndah	Hobart	Robe
Sydney	0.1712	0.0708	0.0865	0.1201	0.0785
Melbourne		0.0862	**0.0082**	0.1502	0.1778
Boulia			0.0542	0.0553	0.1438
Gayndah				**0.0371**	**0.0037**
Hobart					0.4430

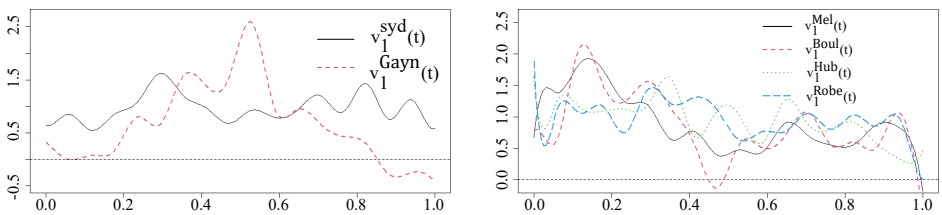

Figure 6. Left panel: Plot of sample eigenfunctions corresponding to the largest eigenvalue of the sample covariance operators of the Sydney and Gayndah detrended minimum temperature profiles, $\hat{v}_1^{\text{Syd}}$ and $\hat{v}_1^{\text{Gayn}}$. A test of $H_0^{(1)}$ suggests that the squared norm of the difference between these curves is significantly larger than 0.1 (P-value ≈ 0.0327). Right panel: Plots of sample eigenfunctions corresponding to the largest eigenvalues of the sample covariance operators from the remaining four stations.

conclusion here: the stations other than Gayndah have similar eigenfunction structure, and the curves calculated from the Gayndah station have different eigenfunction structure. However, for the second eigenfunction conclusions about the uniqueness of the Gayndah station cannot be made with the same level of confidence as for the first eigenfuction.

5. Conclusion

New two-sample tests were introduced to detect relevant differences in the eigenfunctions and eigenvectors of covariance operators of two independent func-

tional time series. These tests can be applied both marginally and, with Bonferroni-type corrections, jointly. The tests are constructed using a self-normalizing strategy, leading to an intricate theoretical analysis to derive the large-sample behavior of the proposed tests. Finite-sample evaluations, done through a simulation study and an application to annual minimum temperature data from Australia, indicate that the tests have very good finite-sample properties and exhibit the features predicted by the theory.

Supplementary Material

The supplement contains the technical details required for the arguments given in Section 2.2 of the main paper.

Acknowledgments

We would like to thank the editor-in-chief, associate editor, and two anonymous referees for a number of insightful comments and suggestions that led to significant improvements to this work. This work was supported in part by the Collaborative Research Center "Statistical modeling of nonlinear dynamic processes" (SFB 823, Teilprojekt A1,C1) of the German Research Foundation (DFG), and the Natural Sciences and Engineering Research Council of Canada, Discovery Grant. We gratefully acknowledge Professors Xiaofeng Shao and Xianyang Zhang for sharing code to reproduce their numerical examples with us.

References

Aitchison, J. (1964). Confidence-region tests. *Journal of the Royal Statistical Society, Series B (Methodological)* **26**, 462–476.

Aston, J. and Kirch, C. (2012). Estimation of the distribution of change-points with application to fMRI data. *The Annals of Applied Statistics* **6**, 1906–1948.

Aue, A., Dubart Norinho, D. and Hörmann, S. (2015). On the prediction of stationary functional time series. *Journal of the American Statistical Association* **110**, 378–392.

Aue, A., Rice, G. and Sönmez, O. (2018). Detecting and dating structural breaks in functional data without dimension reduction. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **80**, 509–529.

Aue, A. and van Delft, A. (2020). Testing for stationarity of functional time series in the frequency domain. *The Annals of Statistics* **48**, 2505–2547.

Benko, M., Härdle, W. and Kneip, A. (2009). Common functional principal components. *The Annals of Statistics* **37**, 1–34.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association* **33**, 526–536.

Bertail, P., Doukhan, P. and Soulier, P. (2006). *Dependence in Probability and Statistics (Lecture Notes in Statistics 187)*. Springer, New York.

Bradley, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys* **2**, 107–144.

Bücher, A., Dette, H. and Heinrichs, F. (2020). Detecting deviations from second-order stationarity in locally stationary functional time series. *Annals of the Institute of Statistical Mathematics* **72**, 1055–1094.

Dauxois, J., Pousse, A. and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of Multivariate Analysis* **12**, 136–154.

Dette, H., Kokot, K. and Aue, A. (2020). Functional data analysis in the Banach space of continuous functions. *The Annals of Statistics* **48**, 1168–1192.

Dette, H., Kokot, K. and Volgushev, S. (2020). Testing relevant hypotheses in functional time series via self-normalization. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82**, 629–660.

Fremdt, S., Steinebach, J. G., Horváth, L. and Kokoskza, P. (2013). Testing the equality of covariance operators in functional samples. *Scandinavian Journal of Statistics* **40**, 138–152.

Guo, J., Zhou, B. and Zhang, J.-T. (2018). Testing the equality of several covariance functions for functional data: A supremum-norm based test. *Computational Statistics & Data Analysis* **124**, 15–26.

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

Hall, P. and Van Keilegom, I. (2007). Two-sample tests in functional data analysis starting from discrete data. *Statistica Sinica* **17**, 1511–1531.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65–70.

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

Horváth, L., Kokoskza, P. and Reeder, R. (2013). Estimation of the mean of functional time series and a two sample problem. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **75**, 103–122.

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

Horváth, L. and Rice, G. (2015). Testing for independence between functional time series. *Journal of Econometrics* **189**, 371–382.

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

Kahle, D. and Wickham, H. (2013). ggmap: Spatial visualization with ggplot2. *The R Journal* **5**, 144–161.

Kokoszka, P. and Reimherr, M. (2013). Asymptotic normality of the principal components of functional time series. *Stochastic Processes and their Applications* **123**, 1546–1562.

Kowal, D. R., Matteson, D. S. and Ruppert, D. (2019). Functional autoregression for sparsely sampled data. *Journal of Business and Economic Statistics* **37**, 97–109.

Panaretos, V. M., Kraus, D. and Maddocks, J. H. (2010). Second-order comparison of Gaussian random functions and the geometry of DNA minicircles. *Journal of the American Statistical Association* **105**, 670–682.

Panaretos, V. M. and Tavakoli, S. (2013). Fourier analysis of stationary time series in function

space. *The Annals of Statistics* **41**, 568–603.

Paparoditis, E. and Sapatinas, T. (2016). Bootstrap-based testing of equality of mean functions or equality of covariance operators for functional data. *Biometrika* **103**, 727–733.

Pigoli, D., Aston, J., Dryden, I. and Secchi, P. (2014). Distances and inference for covariance operators. *Biometrika* **101**, 409–422.

Pomann, G.-M., Staicu, A.-M. and Ghosh, S. (2016). A two-sample distribution-free test for functional data with application to a diffusion tensor imaging study of multiple sclerosis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **65**, 395–414.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Ramsay, J., Hooker, G. and Graves, S. (2009). *Functional Data Analysis with R and MATLAB*. Springer, New York.

Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. Springer, New York.

Shang, H. L. and Hyndman, R. J. (2016). *rainbow: Rainbow Plots, Bagplots and Boxplots for Functional Data*. r package version 3.4.

Shao, X. (2010). A self-normalized approach to confidence interval construction in time series. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **72**, 343–366.

Shao, X. (2015). Self-normalization for time series: A review of recent developments. *Journal of the American Statistical Association* **110**, 1797–1817.

Shao, X. and Zhang, X. (2010). Testing for change points in time series. *Journal of the American Statistical Association* **105**, 1228–1240.

Tavakoli, S. and Panaretos, V. M. (2016). Detecting and localizing differences in functional time series dynamics: A case study in molecular biophysics. *Journal of the American Statistical Association* **111**, 1020–1035.

Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority*. CRC Press, Boca Raton.

Zhang, X. and Shao, X. (2015). Two sample inference for the second-order property of temporally dependent functional data. *Bernoulli* **21**, 909–929.

Zhang, X., Shao, X., Hayhoe, K. and Wuebbles, D. J. (2011). Testing the structural stability of temporally dependent functional observations and application to climate projections. *Electronic Journal of Statistics* **5**, 1765–1796.

Alexander Aue

Department of Statistics, University of California, One Shields Avenue, Davis, CA 95616, USA.

E-mail: aaue@ucdavis.edu

Holger Dette

Fakultät für Mathematik, Ruhr-Universität Bochum, 44780 Bochum, Germany.

E-mail: holger.dette@rub.de

Gregory Rice

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada.

E-mail: grice@uwaterloo.ca

(Received September 2020; accepted June 2021)