Bernoulli **21**(2), 2015, 909–929
DOI: 10.3150/13-BEJ592

Two sample inference for the second-order property of temporally dependent functional data

XIANYANG ZHANG[1] and XIAOFENG SHAO[2]

[1]*Department of Statistics, University of Missouri-Columbia, Columbia, MO, 65211, USA.*
E-mail: zhangxiany@missouri.edu
[2]*Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL, 61820, USA.*
E-mail: xshao@illinois.edu

Motivated by the need to statistically quantify the difference between two spatio-temporal datasets that arise in climate downscaling studies, we propose new tests to detect the differences of the covariance operators and their associated characteristics of two functional time series. Our two sample tests are constructed on the basis of functional principal component analysis and self-normalization, the latter of which is a new studentization technique recently developed for the inference of a univariate time series. Compared to the existing tests, our SN-based tests allow for weak dependence within each sample and it is robust to the dependence between the two samples in the case of equal sample sizes. Asymptotic properties of the SN-based test statistics are derived under both the null and local alternatives. Through extensive simulations, our SN-based tests are shown to outperform existing alternatives in size and their powers are found to be respectable. The tests are then applied to the gridded climate model outputs and interpolated observations to detect the difference in their spatial dynamics.

Keywords: climate downscaling; functional data analysis; long run variance matrix; self-normalization; time series; two sample problem

1. Introduction

Functional data analysis (FDA) which deals with the analysis of curves and surfaces has received considerable attention in the statistical literature during the last decade (Ramsay and Silverman [19,20] and Ferraty and Vieu [6]). This paper falls into a sub-field of functional data analysis: inference for temporally dependent functional data. Specifically, we focus on testing the equality of the second-order structures (e.g., the covariance operators and their associated eigenvalues and eigenfunctions) of two temporally dependent functional sequences. Our work is partially motivated by our ongoing collaboration with atmospheric scientists on the development and assessment of high-resolution climate projections through statistical downscaling. Climate change is one of the most urgent problems facing the world this century. To study climate change, scientists have relied primarily on climate projections from global/regional climate models, which are numerical models that involve systems of differential equations and produce outputs at a prespecified grid. As numerical model outputs are widely used in situations where real observations are not available, it is an important but still open question whether the numerical model outputs

are able to mimic/capture the spatial and temporal dynamics of the real observations. To partly answer this question, we view the spatio-temporal model outputs and real observations as realizations from two temporally dependent functional time series defined on the two-dimensional space and test the equality of their second-order structures which reflects their spatial dynamics/dependence.

Two sample inference for functional data has been investigated by a few researchers. Fan and Lin [5], Cuevas *et al.* [4] and Horváth *et al.* [10] developed the tests for the equality of mean functions. Benko *et al.* [1], Panaretos *et al.* [16], Fremdt *et al.* [7], and Kraus and Panaretos [12] proposed tests for the equality of the second-order structures. All the above-mentioned works assumed the independence between the two samples and/or independence within each sample. However, the assumption of independence within the sample is often too strong to be realistic in many applications, especially if data are collected sequentially over time. For example, the independence assumption is questionable for the climate projection data considered in this paper, as the model outputs and real station observations are simulated or collected over time and temporal dependence is expected. Furthermore the dependence between numerical model outputs and station observations is likely because the numerical models are designed to mimic the dynamics of real observations. See Section 5 for empirical evidence of their dependence. In this paper, we develop new tests that are able to accommodate weak dependence between and within two samples. Our tests are constructed on the basis of functional principal component analysis (FPCA) and the recently developed self-normalization (SN) method (Shao [21]), the latter of which is a new studentization technique for the inference of a univariate time series.

FPCA attempts to find the dominant modes of variation around an overall trend function and has been proved a key technique in the context of FDA. The use of FPCA in the inference of temporally dependent functional data can be found in Gabrys and Kokoszka [8], Hörmann and Kokoszka [9], Horváth *et al.* [10] among others. To account for the dependence, the existing inference procedure requires a consistent estimator of the long run variance (LRV) matrix of the principal component scores or consistent estimator of the LRV operator. However, there is a bandwidth parameter involved in the LRV estimation and its selection has not been addressed in the functional setting. The same issue appears when one considers the block bootstrapping and subsampling schemes (Lahiri [13] and Politis *et al.* [18]), since these techniques also require the selection of a smoothing parameter, such as the block length in the moving block bootstrap, and the window width in the subsampling method (see, e.g., Politis and Romano [17] and McMurry and Politis [15]). Since the finite sample performance can be sensitive to the choice of these tuning parameters and the bandwidth choice can involve certain degree of arbitrariness, it is desirable to use inference methods that are free of bandwidth parameters. To this end, we build on the bandwidth-free SN method (Shao [21]) recently developed in the univariate time series setup, and propose SN-based tests in the functional setting by using recursive estimates obtained from functional data samples.

In time series analysis, the inference of a parameter using normal approximation typically requires consistent estimation of its asymptotic variance. The main difficulty with this approach (and other block-based resampling methods) is the sensitivity of the finite sample performance with respect to the bandwidth parameter, which is often difficult to choose in practice without any parametric assumptions. As a useful alternative, the self-normalized approach uses an inconsistent bandwidth-free estimate of asymptotic variance, which is proportional to asymptotic

variance, so the studentized quantity (statistic) is asymptotically pivotal. Extending the early idea of Lobato [14], Shao [21] proposed a very general kind of self-normalizers that are functions of recursive estimates and showed the theoretical validity for a wide class of parameters of interest. The settings in the latter two papers are however limited to univariate time series. The generalization of the SN method from univariate to functional time series was first done in Zhang *et al.* [23] where the focus was on testing the structure stability of temporally dependent functional data. Here we extend the SN method to test the equality of the second-order properties of two functional time series, which is rather different and new techniques and results are needed. To study the asymptotic properties of the proposed test statistics, we establish functional central limit theorems for the recursive estimates of quantities associated with the second-order properties of the functional time series which seems unexplored in the literature and are thus of independent interest. Based on the functional central limit theorem, we show that the SN-based test statistics have pivotal limiting distributions under the null and are consistent under the local alternatives. From a methodological viewpoint, this seems to be the first time that the SN method is extended to the two sample problem. Compared to most of the existing methods which assumed the independence between the two samples and/or independence within each sample, the SN method not only allows for unknown dependence within each sample but also allows for unknown dependence between the two samples when the sample sizes of the two sequences are equal.

2. Methodology

We shall consider temporally dependent functional processes $\{(X_i(t), Y_i(t)), t \in \mathcal{I}\}_{i=1}^{+\infty}$ defined on some compact set $\mathcal{I}$ of the Euclidian space, where $\mathcal{I}$ can be one-dimensional (e.g., a curve) or multidimensional (e.g., a surface or manifold). For simplicity, we consider the Hilbert space $\mathbb{H}$ of square integrable functions with $\mathcal{I} = [0, 1]$ (or $\mathcal{I} = [0, 1]^2$). For any functions $f, g \in \mathbb{H}$, the inner product between f and g is defined as $\int_{\mathcal{I}} f(t)g(t)\,\mathrm{d}t$ and $\|\cdot\|$ denotes the inner product induced norm. Assume the random elements all come from the same probability space $(\Omega, \mathcal{A}, \mathcal{P})$. Let L^p be the space of real valued random variables with finite L^p norm, that is, $(E|X|^p)^{1/p} < \infty$. Further, we denote $L^p_{\mathbb{H}}$ the space of $\mathbb{H}$ valued random variables X such that $(E\|X\|^p)^{1/p} < \infty$.

Given two sequences of temporally dependent functional observations, $\{X_i(t)\}_{i=1}^{N_1}$ and $\{Y_i(t)\}_{i=1}^{N_2}$ defined on a common region $\mathcal{I}$, we are interested in comparing their second-order properties. Suppose that the functional time series are second-order stationary. We assume that $E[X_i(t)] = E[Y_i(t)] = 0$. The result can be easily extended to the situation with nonzero mean functions. Define $C_X = E[\langle X_i, \cdot\rangle X_i]$ and $C_Y = E[\langle Y_i, \cdot\rangle Y_i]$ as the covariance operators of the two sequences respectively. For the convenience of presentation, we shall use the same notation for the covariance operator and the associated covariance function. Denote by $\{\phi_X^j\}_{j=1}^{\infty}$ and $\{\lambda_X^j\}_{j=1}^{\infty}$ the eigenfunctions and eigenvalues of C_X. Analogous quantities are $\{\phi_Y^j\}_{j=1}^{\infty}$ and $\{\lambda_Y^j\}_{j=1}^{\infty}$ for the second sample. Denote by $|\mathbf{v}|$ the Euclidean norm of a vector $\mathbf{v} \in \mathbb{R}^p$. Let vech$(\cdot)$ be the operator that stacks the columns below the diagonal of a symmetric $m \times m$ matrix as a vector with $m(m+1)/2$ components. Let $D[0, 1]$ be the space of functions on $[0, 1]$ which are right-continuous and have left limits, endowed with the Skorokhod topology (see Billingsley

[2]). Weak convergence in $D[0,1]$ or more generally in the $\mathbb{R}^m$-valued function space $D^m[0,1]$ is denoted by "$\Rightarrow$", where $m \in \mathbb{N}$ and convergence in distribution is denoted by "$\to^d$". Define $\lfloor a \rfloor$ the integer part of $a \in \mathbb{R}$, and $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$. In what follows, we shall discuss the tests for comparing the three quantities C_X, ϕ_X^j and λ_X^j with C_Y, ϕ_Y^j and λ_Y^j, respectively.

2.1. Covariance operator

Consider the problem of testing the hypothesis $H_{1,0}: C_X = C_Y$ versus the alternative $H_{1,a}: C_X \neq C_Y$ (in the operator norm sense) for two mean zero stationary functional time series $\{X_i(t)\}_{i=1}^{N_1}$ and $\{Y_i(t)\}_{i=1}^{N_2}$. Let $N = N_1 + N_2$. Throughout the paper, we assume that

$$N_1/N \to \gamma_1, \qquad N_2/N \to \gamma_2, \qquad \text{as } \min(N_1, N_2) \to +\infty,$$

where $\gamma_1, \gamma_2 \in (0,1)$ and $\gamma_1 + \gamma_2 = 1$. Define the one-dimensional operator $\mathcal{X}_i = \langle X_i, \cdot \rangle X_i = X_i \otimes X_i$ and $\mathcal{Y}_j = \langle Y_j, \cdot \rangle Y_j = Y_j \otimes Y_j$. Let $\hat{C}_{XY}$ be the empirical covariance operator based on the pooled samples, that is,

$$\hat{C}_{XY} = \frac{1}{N_1 + N_2} \left(\sum_{i=1}^{N_1} \mathcal{X}_i + \sum_{i=1}^{N_2} \mathcal{Y}_i \right). \tag{2.1}$$

Denote by $\{\hat{\lambda}_{XY}^j\}$ and $\{\hat{\phi}_{XY}^j\}$ the corresponding eigenvalues and eigenfunctions. The population counterpart of $\hat{C}_{XY}$ is then given by $\tilde{C}_{XY} = \gamma_1 C_X + \gamma_2 C_Y$ whose eigenvalues and eigenfunctions are denoted by $\{\tilde{\lambda}^j\}$ and $\{\tilde{\phi}^j\}$ respectively. Further let $\hat{C}_{X,m} = \frac{1}{m} \sum_{i=1}^m \mathcal{X}_i$ be the sample covariance operator based on the subsample $\{X_i(t)\}_{i=1}^m$ with $2 \leq m \leq N_1$. Define $\{\hat{\phi}_{X,m}^j\}_{j=1}^m$ and $\{\hat{\lambda}_{X,m}^j\}_{j=1}^m$ the eigenfunctions and eigenvalues of $\hat{C}_{X,m}$ respectively, that is,

$$\int_{\mathcal{I}} \hat{C}_{X,m}(t,s) \hat{\phi}_{X,m}^j(s)\, \mathrm{d}s = \hat{\lambda}_{X,m}^j \hat{\phi}_{X,m}^j(t), \tag{2.2}$$

and $\int_{\mathcal{I}} \hat{\phi}_{X,m}^i(t) \hat{\phi}_{X,m}^j(t)\, \mathrm{d}t = \delta_{ij}$. Similarly, quantities $\hat{C}_{Y,m'}$, $\{\hat{\phi}_{Y,m'}^j\}_{j=1}^{N_2}$ and $\{\hat{\lambda}_{Y,m'}^j\}_{j=1}^{N_2}$ are defined for the second sample with $2 \leq m' \leq N_2$. To introduce the SN-based test, we define the recursive estimates

$$c_k^{i,j} = \langle (\hat{C}_{X,\lfloor kN_1/N \rfloor} - \hat{C}_{Y,\lfloor kN_2/N \rfloor}) \hat{\phi}_{XY}^i, \hat{\phi}_{XY}^j \rangle, \qquad 2 \leq k \leq N, 1 \leq i, j \leq K,$$

which estimate the difference of the covariance operators on the space spanned by $\{\tilde{\phi}^j\}_{j=1}^K$. Here K is a user-chosen number, which is held fixed in the asymptotics. Denote by $\hat{\alpha}_k = \mathrm{vech}(\mathbf{C}_k)$ with $\mathbf{C}_k = (c_k^{i,j})_{i,j=1}^K$. In the independent and Gaussian case, Panaretos *et al.* [16] proposed the following test (hereafter, the PKM test),

$$T_{N_1,N_2} = \frac{N_1 N_2}{2N} \sum_{i=1}^K \sum_{j=1}^K \frac{(c_N^{i,j})^2}{\hat{\varrho}_i \hat{\varrho}_j}, \qquad \hat{\varrho}_j = \frac{1}{N} \left\{ \sum_{i=1}^{N_1} \big(\langle X_i, \hat{\phi}_{XY}^j \rangle \big)^2 + \sum_{i=1}^{N_2} \big(\langle Y_i, \hat{\phi}_{XY}^j \rangle \big)^2 \right\},$$

which converges to $\chi^2_{(K+1)K/2}$ under the null. To take the dependence into account, we introduce the SN matrix

$$V^{(1)}_{SN,N}(d) = \frac{1}{N^2}\sum_{k=1}^{N} k^2(\hat{\alpha}_k - \hat{\alpha}_N)(\hat{\alpha}_k - \hat{\alpha}_N)', \tag{2.3}$$

with $d = (K+1)K/2$. The SN-based test statistic is then defined as,

$$G^{(1)}_{SN,N}(d) = N\hat{\alpha}'_N\big(V^{(1)}_{SN,N}(d)\big)^{-1}\hat{\alpha}_N. \tag{2.4}$$

Notice that the PKM test statistic can also be written as a quadratic form of $\hat{\alpha}_N$ but with a different normalization matrix that is only applicable to the independent and Gaussian case. The special form of the SN-based test statistic makes it robust to the dependence within each sample and also the dependence between the two samples when their sample sizes are equal. We shall study the asymptotic behavior of $G^{(1)}_{SN,N}(d)$ under the weak dependence assumption in Section 3.

2.2. Eigenvalues and eigenfunctions

In practice, it is also interesting to infer how far the marginal distributions of two sequences of stationary functional time series coincide/differ and quantify the difference. By the Karhunen–Loève expansion (Bosq [3], page 26), we have

$$X_i(t) = \sum_{j=1}^{+\infty}\sqrt{\lambda^j_X}\beta_{X_i,j}\phi^j_X(t), \qquad Y_i(t) = \sum_{j=1}^{+\infty}\sqrt{\lambda^j_Y}\beta_{Y_i,j}\phi^j_Y(t),$$

where $\beta_{X_i,j} = \int_{\mathcal{I}} X_i(t)\phi^j_X(t)\,\mathrm{d}t$ and $\beta_{Y_i,j} = \int_{\mathcal{I}} Y_i(t)\phi^j_Y(t)\,\mathrm{d}t$ are the principal components (scores), which satisfy that $E[\beta_{X_i,j}\beta_{X_i,j'}] = \delta_{jj'}$ and $E[\beta_{Y_i,j}\beta_{Y_i,j'}] = \delta_{jj'}$. The problem is then translated into testing the equality of the functional principal components (FPC's) namely the eigenvalues and eigenfunctions. For a prespecified positive integer M, we denote the vector of the first M eigenvalues by $\lambda^{1:M}_X = (\lambda^1_X, \ldots, \lambda^M_X)$ and $\lambda^{1:M}_Y = (\lambda^1_Y, \ldots, \lambda^M_Y)$. Further define $\phi^{1:M}_X = (\phi^1_X, \ldots, \phi^M_X)$ and $\phi^{1:M}_Y = (\phi^1_Y, \ldots, \phi^M_Y)$ the first M eigenfunctions of the covariance operators C_X and C_Y, respectively. Since the eigenfunctions are determined up to a sign, we assume that $\langle\phi^j_X, \phi^j_Y\rangle \geq 0$ in order for the comparison to be meaningful. We aim to test the null hypothesis $H_{2,0}: \lambda^{1:M}_X = \lambda^{1:M}_Y$ and $H_{3,0}: \phi^{1:M}_X = \phi^{1:M}_Y$ versus the alternatives that $H_{2,a}: \lambda^{1:M}_X \neq \lambda^{1:M}_Y$ and $H_{3,a}: \phi^{1:M}_X \neq \phi^{1:M}_Y$ (in the L^2 norm sense). The problem of comparing the FPC's of two independent and identically distributed (i.i.d.) functional sequences has been considered in Benko *et al.* [1], where the authors proposed an i.i.d. bootstrap method which seems not applicable to the dependent case. The block bootstrap based method is expected to be valid in the weakly dependent case but the choice of the block size seems to be a difficult task in the current setting. To accommodate the dependence and avoid the bandwidth choice, we adopt the SN idea.

Recall the recursive estimates of the eigenvalues $\hat{\lambda}^j_{X,m}$ and $\hat{\lambda}^j_{Y,m'}$ which are calculated based on the subsamples $\{X_i(t)\}_{i=1}^m$ and $\{Y_i(t)\}_{i=1}^{m'}$. Let $\hat{\theta}^j_k = \hat{\lambda}^j_{X,\lfloor kN_1/N\rfloor} - \hat{\lambda}^j_{Y,\lfloor kN_2/N\rfloor}$ and $\hat{\theta}_k =$

$(\hat{\theta}_k^1, \ldots, \hat{\theta}_k^M)'$ with $\lfloor N\epsilon \rfloor \leq k \leq N$ for some $\epsilon \in (0, 1]$, which is held fixed in the asymptotics. We consider the trimmed SN-based test statistic

$$G_{SN,N}^{(2)}(M) = N^3 \hat{\theta}_N' \left\{ \sum_{k=\lfloor N\epsilon \rfloor}^{N} k^2 (\hat{\theta}_k - \hat{\theta}_N)(\hat{\theta}_k - \hat{\theta}_N)' \right\}^{-1} \hat{\theta}_N. \tag{2.5}$$

The trimmed version of the SN-based test statistic is proposed out of technical consideration when the functional observations lie on an infinite dimensional space. It can be seen from the proof in the supplemental material [22] that the trimming is not required when functional data lie on a finite-dimensional space; see Remark 0.1 in the supplemental material [22].

Remark 2.1. To compare the difference between the eigenvalues, one may also consider their ratios. Define $\hat{\zeta}_k = (\hat{\lambda}_{X,\lfloor kN_1/N \rfloor}^1 / \hat{\lambda}_{Y,\lfloor kN_2/N \rfloor}^1, \ldots, \hat{\lambda}_{X,\lfloor kN_1/N \rfloor}^M / \hat{\lambda}_{Y,\lfloor kN_2/N \rfloor}^M)'$ for $k = \lfloor N\epsilon \rfloor, \ldots, N$. An alternative SN-based test statistic is given by

$$\tilde{G}_{SN,N}^{(2)}(M) = N(\hat{\zeta}_N - \mathbf{l}_M)' \left\{ \frac{1}{N^2} \sum_{k=\lfloor N\epsilon \rfloor}^{N} k^2 (\hat{\zeta}_k - \hat{\zeta}_N)(\hat{\zeta}_k - \hat{\zeta}_N)' \right\}^{-1} (\hat{\zeta}_N - \mathbf{l}_M), \tag{2.6}$$

where $\mathbf{l}_M$ is a M-dimensional vector of all ones. Since the finite sample improvement by using $\tilde{G}_{SN,N}^{(2)}(M)$ is not apparent, we do not further investigate the properties of $\tilde{G}_{SN,N}^{(2)}(M)$.

We now turn to the problem of testing the equality of the eigenfunctions. To proceed, we let

$$\hat{v}_j = \left(\hat{\phi}_{XY}^{j+1}, \hat{\phi}_{XY}^{j+2}, \ldots, \hat{\phi}_{XY}^{p}\right) \tag{2.7}$$

be a vector of $p - j$ orthonormal basis functions for $j = 1, 2, \ldots, M$ with $M \leq p$ and p being a user chosen number. Recall that $\hat{\phi}_{X,m}^j(t)$ and $\hat{\phi}_{Y,m'}^j(t)$ are the jth eigenfunctions of the empirical covariance operators $\hat{C}_{X,m}$ and $\hat{C}_{Y,m'}$ which are computed based on the first m (and m') samples. Here we require that $\langle \hat{\phi}_{X,m}^j, \hat{\phi}_{X,N_1}^j \rangle \geq 0$ and $\langle \hat{\phi}^j{}_{Y,m'}, \hat{\phi}_{X,N_1}^j \rangle \geq 0$ for $2 \leq m \leq N_1$ and $2 \leq m' \leq N_2$. As the eigenfunctions are defined on an infinite-dimensional space, we project the difference between the jth eigenfunctions onto the space spanned by $\hat{v}_j$. Formally, we define the projection vectors

$$\hat{\eta}_k^j = \left(\langle \hat{\phi}_{X,\lfloor kN_1/N \rfloor}^j - \hat{\phi}_{Y,\lfloor kN_2/N \rfloor}^j, \hat{\phi}_{XY}^{j+1} \rangle, \ldots, \langle \hat{\phi}_{X,\lfloor kN_1/N \rfloor}^j - \hat{\phi}_{Y,\lfloor kN_2/N \rfloor}^j, \hat{\phi}_{XY}^{p} \rangle\right),$$

where $1 \leq j \leq M$ and $k = \lfloor N\epsilon \rfloor, \ldots, N$. Further let $\hat{\eta}_k = (\hat{\eta}_k^1, \hat{\eta}_k^2, \ldots, \hat{\eta}_k^M)' \in \mathbb{R}^{M_0}$ with $M_0 = \frac{M(2p-M-1)}{2}$. The trimmed SN-based test statistic is then defined as

$$G_{SN,N}^{(3)}(M_0) = N \hat{\eta}_N' \left\{ \frac{1}{N^2} \sum_{k=\lfloor N\epsilon \rfloor}^{N} k^2 (\hat{\eta}_k - \hat{\eta}_N)(\hat{\eta}_k - \hat{\eta}_N)' \right\}^{-1} \hat{\eta}_N, \tag{2.8}$$

for some $0 < \epsilon < 1$.

Remark 2.2. It is worth noting that $G^{(3)}_{SN,N}(M_0)$ is designed for testing the equality of the first M eigenfunctions. Suppose we are interested in testing the hypothesis for a particular eigenfunction, that is, the null $\phi^j_X = \phi^j_Y$ versus the alternative $\phi^j_X \neq \phi^j_Y$. We can consider the basis functions

$$\tilde{\nu}_j = \big(\hat{\phi}^1_{XY} \ldots, \hat{\phi}^{j-1}_{XY}, \hat{\phi}^{j+1}_{XY}, \ldots, \hat{\phi}^p_{XY}\big),$$

and the projection vector $\hat{\eta}^j_k = (\langle \hat{\phi}^j_{X,\lfloor kN_1/N \rfloor} - \hat{\phi}^j_{Y,\lfloor kN_2/N \rfloor}, \hat{\phi}^1_{XY} \rangle, \ldots, \langle \hat{\phi}^j_{X,\lfloor kN_1/N \rfloor} - \hat{\phi}^j_{Y,\lfloor kN_2/N \rfloor}, \hat{\phi}^{j-1}_{XY} \rangle, \langle \hat{\phi}^j_{X,\lfloor kN_1/N \rfloor} - \hat{\phi}^j_{Y,\lfloor kN_2/N \rfloor}, \hat{\phi}^{j+1}_{XY} \rangle, \ldots, \langle \hat{\phi}^j_{X,\lfloor kN_1/N \rfloor} - \hat{\phi}^j_{Y,\lfloor kN_2/N \rfloor}, \hat{\phi}^p_{XY} \rangle)'$. The SN-based test statistic can then be constructed in a similar manner. We also note that when $\phi^j_X \neq \phi^j_Y$ and $\phi^i_X = \phi^i_Y$ for $i \neq j$, the choice of $\tilde{\nu}_j$ may result in trivial power because $\langle \phi^j_X - \phi^j_Y, \tilde{\phi}^i \rangle$ for $i \neq j$ can be close to 0. In this case, one remedy is to consider alternative basis functions, for example, (4.5) and (4.6) as suggested in the simulation.

Remark 2.3. The choice of the basis functions $\hat{\nu}_j$ is motivated by the Bahadur representation of the recursive estimates in the supplemental material [22]. Under suitable assumptions as given in the next section, it can be shown that

$$\langle \hat{\phi}^a_{X,k}, \phi \rangle = \langle \phi^a_X, \phi \rangle - \frac{1}{k} \sum_{i=1}^{k} \left\{ \sum_{s \neq a} \frac{\beta_{X_i,s} \beta_{X_i,a}}{\lambda^s_X - \lambda^a_X} \langle \phi^s_X, \phi \rangle \right\} + R^a_{X,k}, \tag{2.9}$$

with $R^a_{X,k}$ being the remainder term and $\phi \in L^2(\mathcal{I})$. The second term on the RHS of (2.9) plays a key role in determining the limiting distribution of the SN-based test statistic. When $\phi = \phi^j_X$ with $j \neq a$, the linear term reduces to $-\frac{1}{k} \sum_{i=1}^{k} \frac{\beta_{X_i,j} \beta_{X_i,a}}{\lambda^j_X - \lambda^a_X}$, which satisfies the functional central limit theorem under suitable weak dependence assumption. Notice that the linear term vanishes when $\phi = \phi^a_X$ and the asymptotic distribution of the projection vector is degenerate. It is also worth noting that the linear terms in the Bahadur representations of $\langle \hat{\phi}^a_{X,k}, \phi^j_X \rangle$ and $\langle \hat{\phi}^j_{X,k}, \phi^a_X \rangle$ are opposite of each other which suggests that when testing the eigenfunctions jointly, the basis functions should be chosen in a proper way so that the asymptotic covariance matrix of the projection vector, that is, $\hat{\eta}_k$ is nondegenerate.

3. Theoretical results

To study the asymptotic properties of the proposed statistics, we adopt the dependence measure proposed in Hörmann and Kokoszka [9], which is applicable to the temporally dependent functional process. There are also other weak dependence measures (e.g., mixing) or specific processes (e.g., functional linear processes) suitable for the asymptotic analysis of functional time series (see Bosq [3]), we decide to use Hörmann and Kokoszka's L_p-m-approximating dependence measure for its broad applicability to linear and nonlinear functional processes as well as its theoretical convenience and elegance.

Definition 3.1. *Assume that $\{X_i\} \in L^p_{\mathbb{H}}$ with $p > 0$ admits the following representation*

$$X_i = f(\varepsilon_i, \varepsilon_{i-1}, \ldots), \qquad i = 1, 2, \ldots, \tag{3.1}$$

where the ε_i's are i.i.d. elements taking values in a measurable space S and f is a measurable function $f : S^\infty \to \mathbb{H}$. For each $i \in \mathbb{N}$, let $\{\varepsilon_j^{(i)}\}_{j\in\mathbb{Z}}$ be an independent copy of $\{\varepsilon_j\}_{j\in\mathbb{Z}}$. The sequence $\{X_i\}$ is said to be L^p-m-approximable if

$$\sum_{m=1}^{\infty} \big(E\big\|X_m - X_m^{(m)}\big\|^p\big)^{1/p} < \infty, \tag{3.2}$$

where $X_i^{(m)} = f(\varepsilon_i, \varepsilon_{i-1}, \ldots, \varepsilon_{i-m+1}, \varepsilon_{i-m}^{(i)}, \varepsilon_{i-m-1}^{(i)}, \ldots)$.

Define $B_q(r)$ as a q-dimensional vector of independent Brownian motions. For $\epsilon \in [0, 1)$, we let

$$W_q(\epsilon) = B_q(1)' J_q(\epsilon)^{-1} B_q(1), \qquad \text{where } J_q(\epsilon) = \int_\epsilon^1 \big(B_q(r) - r B_q(1)\big)\big(B_q(r) - r B_q(1)\big)' \,\mathrm{d}r.$$

The critical values of $W_q := W_q(0)$ have been tabulated by Lobato [14]. In general, the quantiles of $W_q(\epsilon)$ can be obtained via simulation. To derive the asymptotic properties of the proposed tests, we make the following assumptions.

Assumption 3.1. *Assume $\{X_i(t)\}_{i=1}^{+\infty} \subseteq L^2_{\mathbb{H}}$ and $\{Y_i(t)\}_{i=1}^{+\infty} \subseteq L^2_{\mathbb{H}}$ are both L^4-m-approximable and they are mutually independent.*

Assumption 3.2. *Assume $\{(X_i(t), Y_i(t))\}_{i=1}^{+\infty} \subseteq L^4_{\mathbb{H}\times\mathbb{H}}$ is an L^4-m-approximable sequence.*

Assumption 3.3. *Assume $\lambda_X^1 > \lambda_X^2 > \cdots > \lambda_X^{m_0+1}$ and $\lambda_Y^1 > \lambda_Y^2 > \cdots > \lambda_Y^{m_0+1}$, for some positive integer $m_0 \geq 2$.*

Note that Assumption 3.2 allows dependence between $\{X_i(t)\}$ and $\{Y_i(t)\}$, which is weaker than Assumption 3.1. To investigate the asymptotic properties of $G^{(1)}_{SN,N}(d)$ under the local alternatives, we consider the local alternative $H_{1,a} : C_X - C_Y = L\bar{C}/\sqrt{N}$ with $\bar{C}$ being a Hilbert–Schmidt operator, where L is a nonzero constant. Define $\Delta = (\langle \bar{C}\tilde{\phi}^i, \tilde{\phi}^j\rangle)_{i,j=1}^K \in \mathbb{R}^{K\times K}$ as the projection of $\bar{C}$ onto the space spanned by $\{\tilde{\phi}^1, \tilde{\phi}^2, \ldots, \tilde{\phi}^K\}$ and assume that $\mathrm{vech}(\Delta) \neq \mathbf{0} \in \mathbb{R}^d$. The following theorem states the asymptotic behaviors of $G^{(1)}_{SN,N}(d)$ under the null and the local alternatives.

Theorem 3.1. *Suppose Assumptions 3.1, 3.3 hold with $m_0 \geq K$. Further assume that the asymptotic covariance matrices $\Lambda_d^*(\Lambda_d^*)'$ given in Lemma 0.3 is positive definite. Then under $H_{1,0}$, $G^{(1)}_{SN,N}(d) \to^d W_d$ and under $H_{1,a}$, $\lim_{|L|\to+\infty} \lim_{N\to+\infty} G^{(1)}_{SN,N}(d) = +\infty$. Furthermore, if $\gamma_1 = \gamma_2$, then the conclusion also holds with Assumption 3.1 replaced by Assumption 3.2.*

It is seen from Theorem 3.1 that $G^{(1)}_{SN,N}(d)$ has pivotal limiting distributions under the null and they are consistent under the local alternatives as $L \to +\infty$. It is worth noting that in our asymptotic framework, d (or K) is assumed to be fixed as $n \to \infty$. Since K is usually chosen to make the first K principle components explain a certain percentage of variation (say 85%), the magnitude of K critically depends on the prespecified threshold and the decay rate of the eigenvalues. In some cases, $d = (K+1)K/2$ can be quite large relative to sample size so it may be more meaningful to use the asymptotic results established under the framework that $d \to \infty$ but $d/n \to 0$ as $n \to \infty$. This motivates the question that whether the following convergence result

$$\sup_{x\in\mathbb{R}} \left| P\left(G^{(1)}_{SN,N}(d) \le x\right) - P(W_d \le x) \right| \to 0 \qquad \text{as } n \to \infty$$

holds when d diverges to ∞ but at a slower rate than n. This would be an interesting future research topic but is beyond the scope of this paper.

To study the asymptotics of $G^{(2)}_{SN,N}(M)$ and $G^{(3)}_{SN,N}(M_0)$, we introduce some notation. Let $\omega^{jk}_{X_i} = \beta_{X_i,j}\beta_{X_i,k}$ and $r^{jk,j'k'}_X(h) = E[(\omega^{jk}_{X_i} - \delta_{jk}\lambda_j)(\omega^{j'k'}_{X_{i+h}} - \delta_{j'k'}\lambda_{j'})]$ be the cross-covariance function between $\omega^{jk}_{X_i}$ and $\omega^{j'k'}_{X_i}$ at lag h. Set $r^{jk}_X(h) := r^{jk,jk}_X(h)$. Define $v^{jk}_{X_i} = \omega^{jk}_{X_i} - E[\omega^{jk}_{X_i}] = \omega^{jk}_{X_i} - \delta_{jk}\lambda_j$. Analogous quantities $r^{jk,j'k'}_Y(h)$ and $v^{jk}_{Y_i}$ can be defined for the second sample. We make the following assumption to facilitate our derivation.

Assumption 3.4. *Suppose that*

$$\sum_{j,k}\sum_{j',k'}\left(\sum_{h=-\infty}^{+\infty}\left|r^{jk,j'k'}_X(h)\right|\right)^2 < +\infty, \qquad \sum_{j,k}\sum_{h=-\infty}^{+\infty}\left|r^{jk}_X(h)\right| < +\infty \tag{3.3}$$

and

$$\sum_{j,k}\sum_{j',k'}\sum_{i_1,i_2,i_3\in\mathbb{Z}}\left|\operatorname{cum}\left(v^{jk}_{X_0}, v^{jk}_{X_{i_1}}, v^{j'k'}_{X_{i_2}}, v^{j'k'}_{X_{i_3}}\right)\right| < \infty. \tag{3.4}$$

The summability conditions also hold for the second sample $\{Y_i(t)\}$.

Assumption 3.4 is parallel to the summability condition considered in Benko *et al.* [1] (see Assumption 1 therein) for i.i.d. functional data. It is not hard to verify the above assumption for Gaussian linear functional process (see, e.g., Bosq [3]), as demonstrated in the following proposition.

Proposition 3.1. *Consider the linear process* $X_i(t) = \sum_{j=0}^{\infty} b_j\varepsilon_{i-j}(t)$, *where* $\varepsilon_j(t) = \sum_{i=1}^{\infty}\sqrt{\lambda_i}z_{i,j}\phi_i(t)$ *with* $\{z_{i,j}\}$ *being a sequence of independent standard normal random variables across both index* i *and* j. *Let* $\pi(h) = \sum_i b_i b_{i+h}$. *Assume that* $\sum_{j=1}^{\infty}\lambda_j < \infty$ *and* $\sum_h |\pi(h)| < \infty$. *Then Assumption* 3.4 *holds for* $\{X_i(t)\}$.

Theorem 3.2. *Suppose Assumptions* 3.1, 3.3, 3.4 *hold with* $m_0 \ge M$ *and the asymptotic covariance matrix* $\tilde{\Lambda}_M\tilde{\Lambda}'_M$ *given in Lemma* 0.5 *is positive definite. Then under* $H_{2,0}$, *we have*

$G^{(2)}_{SN,N}(M) \to^d W_M(\epsilon)$. *Under the local alternative* $H_{2,a}: \lambda_X^{1:M} - \lambda_Y^{1:M} = \frac{L}{\sqrt{N}}\bar{\lambda}$ *with* $\bar{\lambda} \neq \mathbf{0} \in \mathbb{R}^M$, *we have* $\lim_{|L|\to\infty} \lim_{N\to+\infty} G^{(2)}_{SN,N}(M) = +\infty$.

In order to study the asymptotic properties of $G^{(3)}_{SN,N}(M_0)$ under the null and local alternative, we further make the following assumption.

Assumption 3.5. *Let* $\beta^{(m)}_{X_i,j} = \int X_i^{(m)}(t)\phi_X^j(t)\,\mathrm{d}t$, *where* $X_i^{(m)}$ *is the* m*-dependent approximation of* $X_i(t)$ (*see Definition* 3.1). *Suppose one of the following conditions holds*:

$$\sum_{m=1}^{\infty}\sum_{j=1}^{\infty}\left\{E\left(\beta_{X_1,j} - \beta^{(m)}_{X_1,j}\right)^4\right\}^{1/4} < \infty, \qquad \sum_{j=1}^{\infty}\left(E\beta^4_{X_1,j}\right)^{1/4} < \infty, \tag{3.5}$$

or

$$\sum_{s=1}^{+\infty}\left|\langle \phi_X^s, \tilde{\phi}^j\rangle\right| < +\infty, \qquad 2 \le j \le p. \tag{3.6}$$

The same condition holds for the second sample $\{Y_i(t)\}$.

Theorem 3.3. *Suppose Assumptions* 3.1, 3.3, 3.4 *and* 3.5 *hold with* $m_0 \ge M$ *and the asymptotic covariance matrix* $\bar{\Lambda}_{M_0}\bar{\Lambda}'_{M_0}$ *given in Lemma* 0.7 *is positive definite. Then under* $H_{3,0}$, *we have* $G^{(3)}_{SN,N}(M_0) \to^d W_{M_0}(\epsilon)$.

Proposition 3.2. *Define* $\tilde{\Delta}$ *by replacing* $\hat{\phi}^j_{X,N_1}$, $\hat{\phi}^j_{Y,N_2}$ *and* $\hat{\phi}^j_{XY}$ *with* ϕ^j_X, ϕ^j_Y *and* $\tilde{\phi}^j$ *in the definition of* $\hat{\eta}_N$. *Consider the local alternative* $H_{3,a}: \tilde{\Delta} = L\bar{\boldsymbol{\psi}}/\sqrt{N}$ *with* $\bar{\boldsymbol{\psi}} \neq \mathbf{0} \in \mathbb{R}^{M_0}$. *Suppose Assumptions* 3.1, 3.3, 3.4 *and* 3.5 *hold with* $m_0 \ge M$ *and the asymptotic covariance matrix* $\bar{\Lambda}_{M_0}\bar{\Lambda}'_{M_0}$ *given in Lemma* 0.7 *is positive definite. Then we have*

$$\lim_{|L|\to\infty}\lim_{N\to+\infty} G^{(3)}_{SN,N}(M_0) = +\infty$$

under $H_{3,a}$.

It is worth noting that the conclusions in Theorem 3.2, Theorem 3.3 and Proposition 3.2 also hold with Assumption 3.1 replaced by Assumption 3.2 and $\gamma_1 = \gamma_2$. Finally, we point out that condition (3.5) can be verified for Gaussian linear functional process as shown in the following proposition.

Proposition 3.3. *Consider the Gaussian linear process in Proposition* 3.1. *Assume that* $\sum_{j=1}^{\infty}\sqrt{\lambda_j} < \infty$ *and* $\sum_{m=1}^{\infty}(\sum_{j=m}^{\infty} b_j^2)^{1/2} < \infty$. *Then Assumption* 3.4 *and condition* (3.5) *are satisfied for* $\{X_i(t)\}$.

4. Numerical studies

We conduct a number of simulation experiments to assess the performance of the proposed SN-based tests in comparison with the alternative methods in the literature. We generate functional data on a grid of 10^3 equispaced points in [0, 1], and then convert discrete observations into functional objects by using B-splines with 20 basis functions. We also tried 40 and 100 basis functions and found that the number of basis functions does not affect our results much. Throughout the simulations, we set the number of Monte Carlo replications to be 1000 except for the i.i.d. bootstrap method in Benko *et al.* [1], where the number of replications is only 250 because of high computational cost.

4.1. Comparison of covariance operators

To investigate the finite sample properties of $G^{(1)}_{SN,N}(d)$ for dependent functional data, we modify the simulation setting considered in Panaretos *et al.* [16]. Formally, we consider the model,

$$\sum_{j=1}^{3}\bigl\{\xi^i_{j,1}\sqrt{2}\sin(2\pi jt)+\xi^i_{j,2}\sqrt{2}\cos(2\pi jt)\bigr\},\qquad i=1,2,\ldots,t\in[0,1], \tag{4.1}$$

where the coefficients $\xi_i=(\xi^i_{1,1},\xi^i_{2,1},\xi^i_{3,1},\xi^i_{1,2},\xi^i_{2,2},\xi^i_{3,2})'$ are generated from a VAR process,

$$\xi_i=\rho\xi_{i-1}+\sqrt{1-\rho^2}e_i, \tag{4.2}$$

with $e_i\in\mathbb{R}^6$ being a sequence of i.i.d. normal random variables with mean zero and covariance matrix $\Sigma_e=\frac{1}{1+\mu^2}\operatorname{diag}(\mathbf{v})+\frac{\mu^2}{1+\mu^2}\mathbf{1}_6\mathbf{1}_6'$. We generate two independent functional time series $\{X_i(t)\}$ and $\{Y_i(t)\}$ from (4.1) with $\rho=0.5$ and $\mu=1$. We compare the SN-based test with the PKM test which is designed for independent Gaussian process, and the traditional test which is constructed based on a consistent LRV estimator (denoted by CLRV), that is, $G_{CL,N}(d)=N\hat{\alpha}_N\hat{\Sigma}_\alpha^{-1}\hat{\alpha}_N$, where $\hat{\Sigma}_\alpha$ is a lag window LRV estimator with Bartlett kernel and data dependent bandwidth (see Andrews (1991)). We report the simulation results for $N_1=N_2=100,200$, $K=1,2,3,4,5$ ($d=1,3,6,10,15$) and various values of $\mathbf{v}$ in Table 1. Results in scenario A show that the size distortion of all the three tests increases as K gets larger. The SN-based test has the best size compared to the other two tests. The PKM test is severely oversized due to the fact that it does not take the dependence into account. It is seen from the table that the CLRV test also has severe size distortion especially for large K, which is presumably due to the poor estimation of the LRV matrix of $\hat{\alpha}_N$ when the dimension is high. Under the alternatives, we report the size-adjusted power which is computed using finite sample critical values based on the simulation under the null model where we assume that both $\{X_i(t)\}$ and $\{Y_i(t)\}$ are generated from (4.1) with $\rho=0.5$, $\mu=1$ and $\mathbf{v}=\mathbf{v}_X$. From scenarios B–D in Table 1, we observe that the PKM is most powerful which is largely due to its severe upward size distortion. The SN-based test is less powerful compared to the other two tests but the power loss is generally

Table 1. Empirical sizes and size-adjusted powers of (i) the SN-based test, (ii) the PKM test and (iii) the CLRV test for testing the equality of the covariance operators. The nominal level is 5%

				K						
	Parameter	$N_1 = N_2$		1	2	3	4	5	K_1^*	K_2^*
A	$\mathbf{v}_X = (12, 7, 0.5, 9, 5, 0.3)$	100	(i)	4.3	5.7	6.8	8.7	14.3	8.7	10.7
			(ii)	14.5	20.9	22.9	32.2	39.5	32.0	34.0
			(iii)	9.1	12.9	20.9	39.8	67.9	38.8	48.6
	$\mathbf{v}_Y = (12, 7, 0.5, 9, 5, 0.3)$	200	(i)	4.7	5.7	4.6	7.0	8.0	7.0	7.0
			(ii)	12.8	20.6	26.7	34.7	42.6	34.8	37.3
			(iii)	6.9	9.6	14.5	25.2	41.5	25.1	28.9
B	$\mathbf{v}_X = (14, 7, 0.5, 6, 5, 0.3)$	100	(i)	19.1	23.6	17.7	14.2	12.7	14.1	13.0
			(ii)	27.6	37.7	31.6	22.9	21.2	23.1	22.7
			(iii)	27.0	33.8	23.0	20.5	14.0	20.5	15.8
	$\mathbf{v}_Y = (8, 7, 0.5, 6, 5, 0.3)$	200	(i)	31.2	37.7	30.4	21.9	21.9	21.9	22.1
			(ii)	39.1	61.6	51.7	44.2	41.2	44.2	41.5
			(iii)	37.6	57.0	44.7	30.1	24.3	30.1	24.9
C	$\mathbf{v}_X = (12, 7, 0.5, 9, 3, 0.3)$	100	(i)	5.5	10.9	30.8	62.4	64.7	62.3	63.7
			(ii)	4.7	16.1	57.3	94.6	98.7	94.4	97.1
			(iii)	5.5	13.4	42.3	79.4	70.8	79.2	74.3
	$\mathbf{v}_Y = (12, 7, 0.5, 3, 9, 0.3)$	200	(i)	5.3	10.0	45.7	90.3	94.3	90.4	92.5
			(ii)	6.4	13.0	67.8	99.9	100.0	99.9	100.0
			(iii)	6.1	12.5	60.5	99.9	99.8	99.9	99.8
D	$\mathbf{v}_X = (12, 7, 0.5, 9, 5, 0.3)$	100	(i)	6.1	8.3	28.3	80.1	82.2	75.6	80.6
			(ii)	5.5	14.6	47.2	100.0	100.0	94.7	100.0
			(iii)	6.9	12.3	37.2	95.7	88.6	90.6	90.7
	$\mathbf{v}_Y = (12, 7, 0.5, 0, 5, 0.3)$	200	(i)	5.7	8.9	39.7	96.3	98.4	95.5	98.3
			(ii)	6.4	14.5	53.6	100.0	100.0	99.4	100.0
			(iii)	6.0	12.9	47.7	100.0	100.0	99.3	100.0

Note: Under the alternatives, we simulate the size-adjusted critical values by assuming that both $\{X_i\}$ and $\{Y_i\}$ are generated from (4.1) with $\rho = 0.5$, $\mu = 1$ and $\mathbf{v} = \mathbf{v}_X$.

moderate in most cases. Furthermore, we present the results when choosing K by

$$K_j^* = \arg\min\left\{1 \le J \le 20 : \frac{\sum_{i=1}^{J} \hat{\lambda}_{XY}^i}{\sum_{i=1}^{20} \hat{\lambda}_{XY}^i} > \alpha_j^*\right\}, \qquad j = 1, 2, \tag{4.3}$$

where $\alpha_1^* = 85\%$ and $\alpha_2^* = 95\%$. An alternative way of choosing K is to consider the penalized fit criteria (see Panaretos *et al.* [16] for the details). We notice that the performance of all the three tests based on automatic choice K_j^* is fairly close to the performance when $K = 4$ or 5 in most cases. To sum up, the SN-based test provides the best size under the null and has reasonable power under different alternatives considered here, which is consistent with the "better size but less power" phenomenon seen in the univariate setup (Lobato [14] and Shao [21]).

4.2. Comparison of eigenvalues and eigenfunctions

In this subsection, we study the finite sample performance of the SN-based test for testing the equality of the eigenvalues and eigenfunctions. We consider the data generating process,

$$\sum_{j=1}^{2}\bigl\{\xi_{j,1}^{i}\sqrt{2}\sin(2\pi jt+\delta_j)+\xi_{j,2}^{i}\sqrt{2}\cos(2\pi jt+\delta_j)\bigr\},\qquad i=1,2,\ldots,\ t\in[0,1], \tag{4.4}$$

where $\xi_i^* = (\xi_{1,1}^i, \xi_{2,1}^i, \xi_{1,2}^i, \xi_{2,2}^i)'$ is a 4-variate VAR process (4.2) with $e_i \in \mathbb{R}^4$ being a sequence of i.i.d. normal random variables with mean zero and covariance matrix $\Sigma_e = \frac{1}{1+\mu^2}\operatorname{diag}(\mathbf{v}) + \frac{\mu^2}{1+\mu^2}\mathbf{1}_4\mathbf{1}_4'$. We set $\rho = 0.5$ and $\mu = 0$. Under $H_{2,0}$ (or $H_{2,a}$), $\{X_i(t)\}$ and $\{Y_i(t)\}$ are generated independently from (4.4) with $\delta_1 = \delta_2 = 0$ and $\mathbf{v}_X = \mathbf{v}_Y$ (or $\mathbf{v}_X \neq \mathbf{v}_Y$). Notice that the eigenvalues of $\{X_i(t)\}$ and $\{Y_i(t)\}$ are given respectively, by $\mathbf{v}_X$ and $\mathbf{v}_Y$ when $\delta_1 = \delta_2 = 0$. Under $H_{3,0}$ and $H_{3,a}$, we generate $\{X_i(t)\}$ and $\{Y_i(t)\}$ independently from (4.4) with $\mathbf{v}_X = \mathbf{v}_Y$, $\delta_{X,1} - \delta_{Y,1} = \delta$, and $\delta_{X,2} = \delta_{Y,2} = 0$, where $\delta = 0$ under the null and $\delta \neq 0$ under the alternatives. We aim to test the equality of the first four eigenvalues and eigenfunctions separately and jointly. Because functional data are finite dimensional, we implement the untrimmed version of the SN-based tests, that is, $\epsilon = 0$. To further assess the performance of the SN-based test, we compare our method with the subsampling approach with several choices of subsampling widths and the i.i.d. bootstrap method in Benko *et al.* [1]. Suppose $N_1 = N_2 = N_0$. Let l be the subsampling width and $\lambda_{\mathrm{sub},i}^{j} = \lambda_{\mathrm{sub},X,i}^{j} - \lambda_{\mathrm{sub},Y,i}^{j}$, $i = 1, 2, \ldots, s_{N_0}(l) = \lceil N_0/l \rceil$, where $\lambda_{\mathrm{sub},X,i}^{j}$ and $\lambda_{\mathrm{sub},Y,i}^{j}$ are estimates of the jth eigenvalues based on the ith nonoverlapping subsamples $\{X_k(t)\}_{k=(i-1)l+1}^{il}$ and $\{Y_k(t)\}_{k=(i-1)l+1}^{il}$, respectively. The subsampling variance estimate is given by $\sigma_{\mathrm{sub},j}^2 = \frac{l}{s_{N_0}(l)}\sum_{i=1}^{s_{N_0}(l)}(\hat{\lambda}_{\mathrm{sub},i}^{j} - \frac{1}{s_{N_0}(l)}\sum_{i=1}^{s_{N_0}(l)}\hat{\lambda}_{\mathrm{sub},i}^{j})^2$, and the test statistic based on the subsampling variance estimate for testing the equality of the jth eigenvalue is defined as $G_{\mathrm{sub},N} = N_0(\hat{\lambda}_{X,N_0}^{j} - \hat{\lambda}_{Y,N_0}^{j})^2/\sigma_{\mathrm{sub},j}^2$. Since the data-dependent rule for choosing the subsampling width is not available in the current setting, we tried $l = 8, 12, 16$ for $N_0 = 48, 96$. For testing the equality of eigenvalues jointly and equality of the eigenfunctions, we shall consider a multivariate version of the subsampling-based test statistic which can be defined in a similar fashion. Table 2 summarizes some selective simulation results for testing the eigenvalues with various values of $\mathbf{v}$. From scenario A, we see that performance of the SN-based test under the null is satisfactory while the size distortion of the subsampling-based method is quite severe and is sensitive to the choice of block size l. It is also not surprising to see that the i.i.d. bootstrap method has obvious size distortion as it does not take the dependence into account. Under the alternatives (scenarios B–D), we report the size-adjusted power by using the simulated critical values as described in previous subsection. When the sample size is 48, the SN-based method delivers the highest power among the tests and it tends to have some moderate power loss when the sample size increases to 96. On the other hand, the subsampling method is sensitive to the choice of subsampling width and its power tends to decrease when a larger subsampling width is chosen.

Table 2. Empirical sizes and size-adjusted powers of (i) the SN-based test, the subsampling-based test with (ii) $l = 8$, (iii) $l = 12$ and (iv) $l = 16$, and (v) Benko et al.'s i.i.d. bootstrap based method for testing the equality of the first two eigenvalues separately (the columns with $M = 1, 2$) and jointly (the column with $M = (1, 2)$), and the equality of the first four eigenvalues jointly (the column with $M = (1, 2, 3, 4)$). The nominal level is 5% and the number of replications for i.i.d. bootstrap method is 250

				M			
	Parameter	$N_1 = N_2$		1	2	(1, 2)	(1, 2, 3, 4)
A	$\mathbf{v}_X = (10, 0.5, 5, 0.3)$	48	(i)	5.4	5.1	4.6	3.8
			(ii)	24.2	38.5	52.4	90.8
			(iii)	21.9	28.8	51.3	68.8
			(iv)	21.8	28.1	57.9	44.7
			(v)	11.2	9.2	11.6	11.6
	$\mathbf{v}_Y = (10, 0.5, 5, 0.3)$	96	(i)	5.2	5.6	4.8	5.1
			(ii)	19.0	40.4	46.4	84.8
			(iii)	16.3	29.6	38.2	77.0
			(iv)	16.0	25.3	36.5	78.5
			(v)	14.4	8.4	15.2	15.2
B	$\mathbf{v}_X = (20, 0.5, 5, 0.3)$	48	(i)	25.1	4.3	21.8	15.5
			(ii)	24.2	5.4	13.3	7.1
			(iii)	19.8	6.8	8.8	8.8
			(iv)	14.1	6.8	8.0	9.1
	$\mathbf{v}_Y = (10, 0.5, 5, 0.3)$	96	(i)	48.4	4.8	35.6	25.0
			(ii)	58.4	6.9	29.4	11.4
			(iii)	50.9	6.1	29.7	13.3
			(iv)	53.8	6.0	29.3	11.4
C	$\mathbf{v}_X = (10, 0.5, 5, 0.3)$	48	(i)	6.2	70.6	58.9	44.1
			(ii)	5.5	68.1	54.6	13.9
			(iii)	4.8	49.3	23.0	16.9
			(iv)	6.1	34.2	15.4	21.2
	$\mathbf{v}_Y = (10, 0.5, 1, 0.3)$	96	(i)	4.7	91.4	84.6	77.6
			(ii)	4.7	98.7	96.3	69.7
			(iii)	5.5	97.9	92.5	51.6
			(iv)	5.4	96.5	83.0	29.4
D	$\mathbf{v}_X = (20, 0.5, 5, 0.3)$	48	(i)	27.0	70.1	68.4	55.3
			(ii)	25.8	65.7	51.1	14.0
			(iii)	20.9	58.4	23.9	19.9
			(iv)	14.9	40.1	11.8	17.2
	$\mathbf{v}_Y = (10, 0.5, 1, 0.3)$	96	(i)	55.3	87.9	88.4	83.3
			(ii)	54.5	98.3	96.9	62.3
			(iii)	48.8	97.6	95.2	53.2
			(iv)	50.1	95.2	88.0	27.7

Note: Under the alternatives, we simulate the size-adjusted critical values by assuming that both $\{X_i\}$ and $\{Y_i\}$ are generated from (4.1) with $\rho = 0.5$, $\mu = 0$ and $\mathbf{v} = \mathbf{v}_X$.

To test the equality of the first four eigenfunctions, we implement the SN-based test and the subsampling-based test with the basis functions,

$$\begin{aligned}\hat{\nu}_j^* = (&\hat{\phi}_{XY}^1 + \hat{\phi}_{XY}^j, \ldots, \hat{\phi}_{XY}^{j-1} + \hat{\phi}_{XY}^j, \\ &\hat{\phi}_{XY}^{j+1} + \hat{\phi}_{XY}^j, \ldots, \hat{\phi}_{XY}^p + \hat{\phi}_{XY}^j), \qquad 1 \le j \le 4,\ p = 4,\end{aligned} \tag{4.5}$$

for testing individual eigenfunction and

$$\hat{\nu}_j^{**} = \big(\hat{\phi}_{XY}^{j+1} + \hat{\phi}_{XY}^j, \hat{\phi}_{XY}^{j+2} + \hat{\phi}_{XY}^j, \ldots, \hat{\phi}_{XY}^p + \hat{\phi}_{XY}^j\big), \qquad 1 \le j \le j^*,\ p = 4, \tag{4.6}$$

with $j^* = 2, 3, 4$, for testing the first j^* eigenfunctions jointly (correspondingly $M_0 = 3, 5, 6$). The tests with the above basis functions tend to provide similar sizes but higher powers as compared to the tests with the basis functions $\hat{\nu}_i$ in our simulation study. The basis functions $\hat{\nu}_j^*$ is constructed by adding the same estimated eigenfunction $\hat{\phi}_{XY}^j$ to each component of $\hat{\nu}_j$, and the associated SN-based test is expected to be asymptotically valid in view of the Bahadur representation (2.9). Selective simulation results are summarized in Table 3 and Figure 1 which present the sizes of the SN-based test, the subsampling-based test and the i.i.d. bootstrap method, and the size adjusted powers of the former two respectively. It is seen from Table 3 that the sizes of the SN-based test are accurate while the subsampling-based test is apparently size-distorted. It is somewhat surprising to see that the i.i.d. bootstrap provides better sizes compared to the subsampling-based approach which is designed for dependent data. Figure 1 plots the (size-adjusted) power functions of the SN-based test and the subsampling-based test which are monotonically increasing on δ. When $N_1 = N_2 = 48$, the SN-based test delivers the highest power in most cases. The subsampling-based test with a small subsampling width becomes most powerful when sample size increases to 96. Overall, the SN-based test is preferable as it provides quite accurate size under the null and has respectable power under the alternatives.

5. Climate projections analysis

We apply the SN-based test to a gridded spatio-temporal temperature dataset covering a subregion of North America. The dataset comes from two separate sources: gridded observations generated from interpolation of station records (HadCRU), and gridded simulations generated by an AOGCM (NOAA GFDL CM2.1). Both datasets provide monthly average temperature for the same 19-year period, 1980–1998. Each surface is viewed as a two-dimensional functional datum. The yearly average data have been recently analyzed in Zhang *et al.* [23], where the goal is to detect a possible change point of the bias between the station observations and model outputs. In this paper, we analyze the monthly data from 1980 to 1998, which includes 228 functional images for each sequence. We focus on the second-order properties and aim to test the equality of the eigenvalues and eigenfunctions of the station observations and model outputs. To perform the analysis, we first remove the seasonal mean functions from the two functional sequences. At each location, we have two time series from the demeaned functional sequences. We apply the SN-based test developed in Shao [21] to test whether their cross-correlation at lag zero is equal to zero. The p-values of these tests are plotted in Figure 2. The result tends to suggest that the

Table 3. Empirical sizes of (i) the SN-based test, the subsampling-based test with (ii) $l = 8$, (iii) $l = 12$ and (iv) $l = 16$, and (v) Benko et al.'s i.i.d. bootstrap based method for testing the equality of the first two eigenfunctions separately (the columns with $M = 1, 2$) and jointly (the column with $M = (1, 2)$), and the equality of the first four eigenfunctions jointly (the column with $M = (1, 2, 3, 4)$). The nominal level is 5% and the number of replications for i.i.d. bootstrap is 250

				M			
	Parameter	$N_1 = N_2$		1	2	(1, 2)	(1, 2, 3, 4)
A	$\mathbf{v}_X = (10, 0.5, 5, 0.3)$	48	(i)	6.4	3.2	4.9	4.8
			(ii)	39.5	36.0	78.7	67.0
			(iii)	62.9	62.3	24.8	26.4
			(iv)	32.6	25.9	9.6	7.7
			(v)	2.4	11.2	2.4	2.4
	$\mathbf{v}_Y = (8, 0.5, 4, 0.3)$	96	(i)	4.5	3.3	4.3	4.7
			(ii)	18.2	16.8	27.2	45.8
			(iii)	24.9	21.0	49.0	71.2
			(iv)	32.6	30.6	75.9	61.0
			(v)	3.2	12.4	4.8	6.8
B	$\mathbf{v}_X = (4, 0.5, 2, 0.3)$	48	(i)	8.2	3.8	7.0	6.1
			(ii)	43.3	45.8	83.4	71.4
			(iii)	66.6	65.2	23.8	18.1
			(iv)	26.5	22.5	5.8	3.7
			(v)	2.4	6.0	3.6	1.6
	$\mathbf{v}_Y = (2, 0.5, 1, 0.3)$	96	(i)	5.3	4.5	5.2	4.9
			(ii)	20.3	24.2	36.7	54.6
			(iii)	25.3	27.9	53.0	75.7
			(iv)	33.1	32.6	78.1	60.1
			(v)	2.8	8.4	6.8	3.6

dependence between the station observations and model outputs may not be negligible at certain regions as the corresponding p-values are extremely small. The two sample tests introduced in this paper are useful in this case because they are robust to such dependence.

We perform FPCA on the demeaned sequences. Figure 3 plots the first three PC's of the station observations and model outputs. We then apply the SN-based tests $G^{(2)}_{SN,N}(M)$ and $G^{(3)}_{SN,N}(M_0)$ (with $p = 3$) to the demeaned sequences, which yields the results summarized in Table 4. It is seen from the table that the first two eigenvalues of the station observations and model outputs may be the same, at least statistical significance is below the 10% level, while there is a significant difference between their third eigenvalue. The SN-based tests also suggest that there are significant differences of the first and second PCs between the station observations and model outputs as the corresponding p-values are less than 5% while the difference between the third PCs is not significant at the 10% level; compare Figure 3. We also tried the basis functions v^*_j and v^{**}_j for $G^{(3)}_{SN,N}(M_0)$ (see (4.5) and (4.6)), which leads to the same conclusion. To sum up,

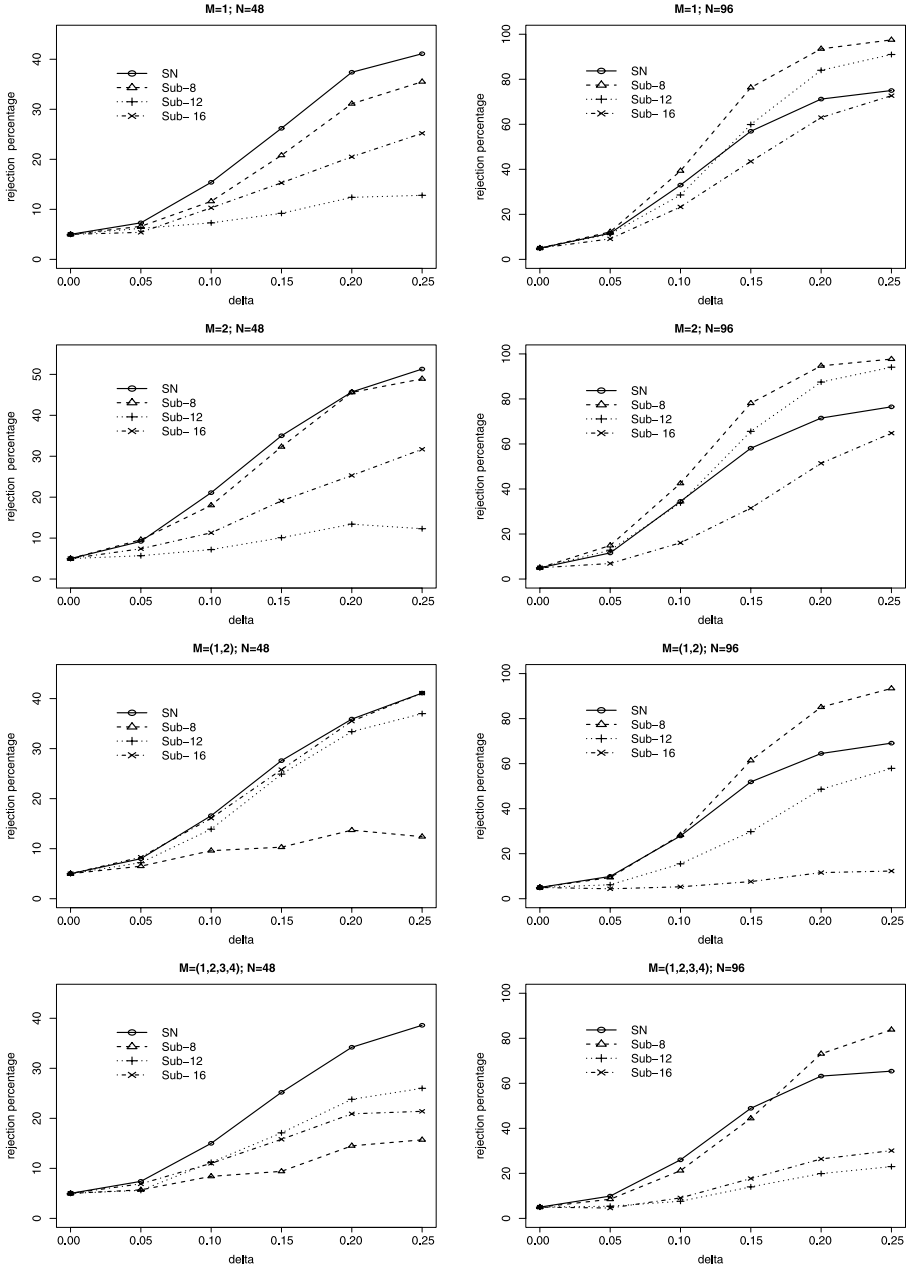

Figure 1. Size-adjusted powers of the SN-based test and the subsampling-based tests for testing the equality of the first two eigenfunctions separately and jointly, and the equality of the first four eigenfunctions jointly.

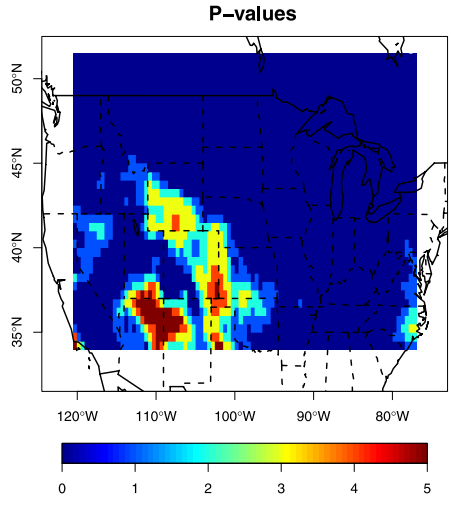

Figure 2. p-values for testing the nullity of lag zero cross-correlation between the station observations and model outputs at each location. The numbers 0–5 denote the ranges of the p-values, that is, 0 denotes [0.1, 1]; 1 denotes [0.05, 0.1]; 2 denotes [0.025, 0.05]; 3 denotes [0.01, 0.025]; 4 denotes [0.005, 0.01] and 5 denotes [0, 0.005].

our results suggest that the second-order properties of the station observations and model outputs may not be the same.

In climate projection studies, the use of numerical models outputs has become quite common nowadays because of advances in computing power and efficient numerical algorithms. As mentioned in Jun *et al.* [11], "Climate models are evaluated on how well they simulate the current mean climate state, how they can reproduce the observed climate change over the last century, how well they simulate specific processes, and how well they agree with proxy data for very different time periods in the past." Furthermore, different institutions produce different model outputs based on different choices of parametrizations, model components, as well as initial and boundary conditions. Thus there is a critical need to assess the discrepancy/similarity between numerical model outputs and real observations, as well as among various model outputs. The two sample tests proposed here can be used towards this assessment at a preliminary stage to get a quantitative idea of the difference, followed by a detailed statistical characterization using sophisticated spatio-temporal modeling techniques (see, e.g., Jun *et al.* [11]). In particular, the observed significance level for each test can be used as a similarity index that measures the similarity between numerical model outputs and real observations, and may be used to rank model outputs. A detailed study along this line would be interesting, but is beyond the scope of this article.

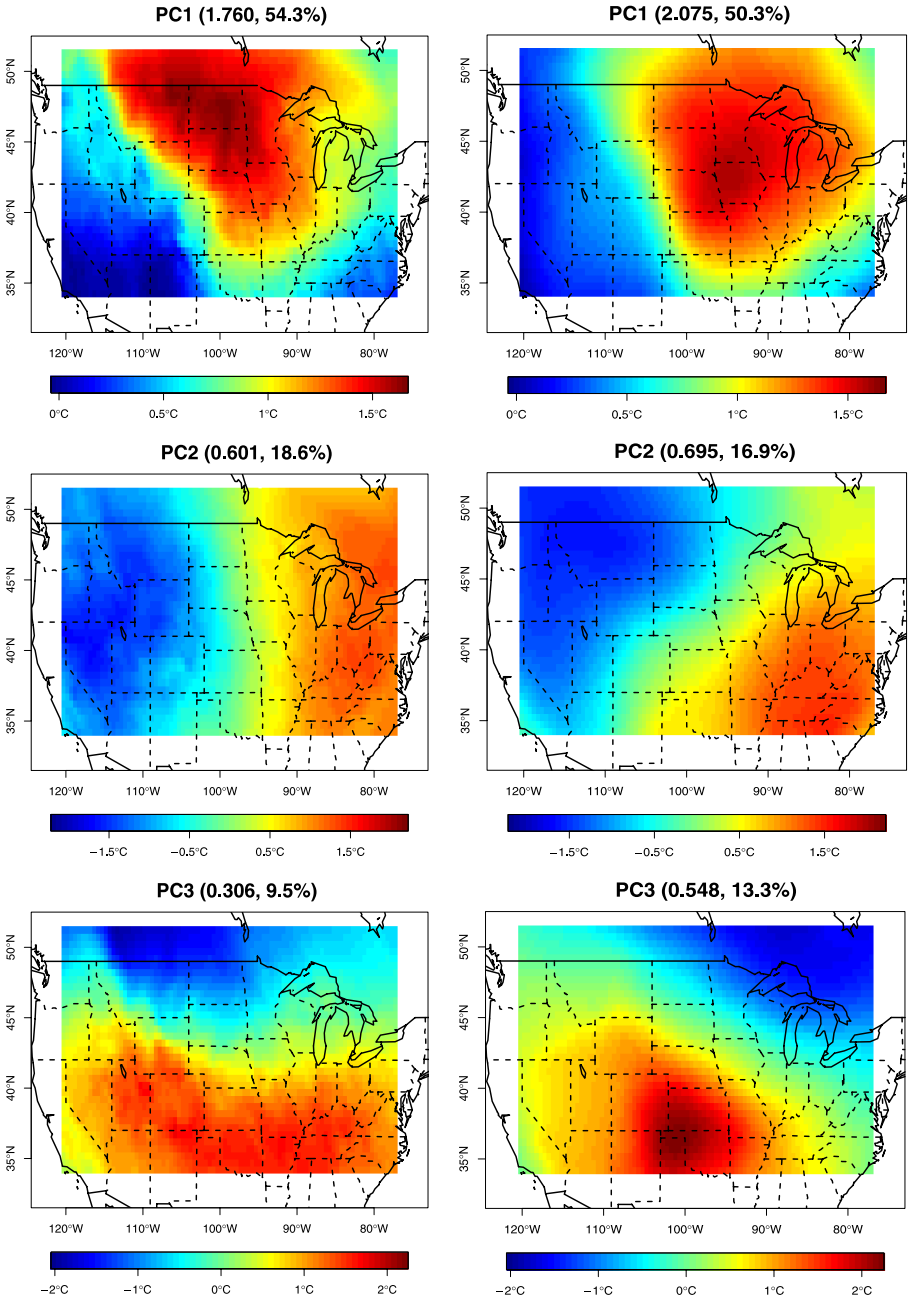

Figure 3. The first three PCs of the station observations (left panels) and model outputs (right panels), and the associated eigenvalues and percentage of variations explained.

Table 4. Comparison of the eigenvalues and eigenfunctions of the covariance operators of the station observations and model outputs

K	$G^{(2)}_{SN,N}(M)$	p-value	$G^{(3)}_{SN,N}(M_0)$	p-value
1	10.8	(0.1, 1)	126.4	(0.025, 0.05)
2	5.4	(0.1, 1)	295.4	(0, 0.005)
3	119.9	(0.005, 0.01)	34.2	(0.1, 1)
–	326.2	(0.005, 0.01)	318.0	(0.005, 0.01)

Note: The first three rows show the results for testing individual eigencomponent, and the last row shows the results for testing the first three eigenvalues and eigenfunctions jointly.

Acknowledgements

We are grateful to an associate editor and a referee for helpful comments, which led to substantial improvements. Shao's research is supported in part by National Science Foundation Grant DMS-11-04545.

Supplementary Material

Supplement to "Two sample inference for the second-order property of temporally dependent functional data" (DOI: 10.3150/13-BEJ592SUPP; .pdf). This supplement contains proofs of the results in Section 3.

References

[1] Benko, M., Härdle, W. and Kneip, A. (2009). Common functional principal components. *Ann. Statist.* **37** 1–34. MR2488343

[2] Billingsley, P. (1999). *Convergence of Probability Measures*, 2nd ed. *Wiley Series in Probability and Statistics*: *Probability and Statistics*. New York: Wiley. MR1700749

[3] Bosq, D. (2000). *Linear Processes in Function Spaces*. *Lecture Notes in Statistics* **149**. New York: Springer. Theory and applications. MR1783138

[4] Cuevas, A., Febrero, M. and Fraiman, R. (2004). An ANOVA test for functional data. *Comput. Statist. Data Anal.* **47** 111–122. MR2087932

[5] Fan, J. and Lin, S.-K. (1998). Test of significance when data are curves. *J. Amer. Statist. Assoc.* **93** 1007–1021. MR1649196

[6] Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis*. *Theory and Practice*. *Springer Series in Statistics*. New York: Springer. MR2229687

[7] Fremdt, S., Steinebach, J.G., Horváth, L. and Kokoszka, P. (2013). Testing the equality of covariance operators in functional samples. *Scand. J. Stat.* **40** 138–152. MR3024036

[8] Gabrys, R. and Kokoszka, P. (2007). Portmanteau test of independence for functional observations. *J. Amer. Statist. Assoc.* **102** 1338–1348. MR2412554

[9] Hörmann, S. and Kokoszka, P. (2010). Weakly dependent functional data. *Ann. Statist.* **38** 1845–1884. MR2662361

[10] Horváth, L., Kokoszka, P. and Reeder, R. (2013). Estimation of the mean of functional time series and a two-sample problem. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 103–122. MR3008273

[11] Jun, M., Knutti, R. and Nychka, D.W. (2008). Spatial analysis to quantify numerical model bias and dependence: How many climate models are there? *J. Amer. Statist. Assoc.* **103** 934–947. MR2528820

[12] Kraus, D. and Panaretos, V.M. (2012). Dispersion operators and resistant second-order functional data analysis. *Biometrika* **99** 813–832.

[13] Lahiri, S.N. (2003). *Resampling Methods for Dependent Data. Springer Series in Statistics*. New York: Springer. MR2001447

[14] Lobato, I.N. (2001). Testing that a dependent process is uncorrelated. *J. Amer. Statist. Assoc.* **96** 1066–1076. MR1947254

[15] McMurry, T. and Politis, D. (2011). Resampling methods for functional data. In *The Oxford Handbook of Functional Data Analysis* (F. Ferraty and Y. Romain, eds.) 189–209. Oxford Univ. Press, Oxford. MR2908023

[16] Panaretos, V.M., Kraus, D. and Maddocks, J.H. (2010). Second-order comparison of Gaussian random functions and the geometry of DNA minicircles. *J. Amer. Statist. Assoc.* **105** 670–682. MR2724851

[17] Politis, D.N. and Romano, J.P. (2010). K-sample subsampling in general spaces: The case of independent time series. *J. Multivariate Anal.* **101** 316–326. MR2564342

[18] Politis, D.N., Romano, J.P. and Wolf, M. (1999). *Subsampling. Springer Series in Statistics*. New York: Springer. MR1707286

[19] Ramsay, J.O. and Silverman, B.W. (2002). *Applied Functional Data Analysis. Springer Series in Statistics*. New York: Springer. Methods and case studies. MR1910407

[20] Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis*, 2nd ed. *Springer Series in Statistics*. New York: Springer. MR2168993

[21] Shao, X. (2010). A self-normalized approach to confidence interval construction in time series. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 343–366. MR2758116

[22] Zhang, X. and Shao, X. (2014). Supplement to "Two sample inference for the second-order property of temporally dependent functional data." DOI:10.3150/13-BEJ592SUPP.

[23] 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. *Electron. J. Stat.* **5** 1765–1796. MR2870150

Received December 2012 and revised December 2013