A Wavelet-Based Independence Test for Functional Data with an Application to MEG Functional Connectivity

Rui Miao, Xiaoke Zhang *

Department of Statistics, The George Washington University
and
Raymond K. W. Wong †
Department of Statistics, Texas A&M University

Abstract

Measuring and testing the dependency between multiple random functions is often an important task in functional data analysis. In the literature, a model-based method relies on a model which is subject to the risk of model misspecification, while a model-free method only provides a correlation measure which is inadequate to test independence. In this paper, we adopt the Hilbert-Schmidt Independence Criterion (HSIC) to measure the dependency between two random functions. We develop a two-step procedure by first pre-smoothing each function based on its discrete and noisy measurements and then applying the HSIC to recovered functions. To ensure the compatibility between the two steps such that the effect of the pre-smoothing error on the subsequent HSIC is asymptotically negligible when the data are densely measured, we propose a new wavelet thresholding method for pre-smoothing and to use Besov-norm-induced kernels for HSIC. We also provide the corresponding asymptotic analysis. The superior numerical performance of the proposed method over existing ones is demonstrated in a simulation study. Moreover, in an magnetoencephalography (MEG) data application, the functional connectivity patterns identified by the proposed method are more anatomically interpretable than those by existing methods.

Keywords: Reproducing kernel Hilbert space; Besov spaces; Permutation test; Human connectome project; Dense functional data.

The authors would like to thank the editor, an associate editor and two referees for their constructive comments and suggestions.

*The research of Xiaoke Zhang is partially supported by National Science Foundation grant DMS-1832046.

†The research of Raymond K. W. Wong is partially supported by National Science Foundation grants DMS-1806063, DMS-1711952 and CCF-1934904.

1 Introduction

In recent decades, functional data analysis (FDA) has developed rapidly due to a huge and increasing number of datasets collected in the form of curves, surfaces and volumes. General introductions to the subject may be found in a few monographs (e.g., Ramsay and Silverman, 2005; Ferraty and Vieu, 2006). In many scientific fields, measurements are taken from multiple random functions per subject and the dependency between these functions is of interest. For instance, neuroscientists are interested in functional connectivity patterns between signals at multiple brain regions, which are measured over time in functional magnetic resonance imaging data. It is thus an important task in FDA to measure their dependency and to further test the significance of the dependency. Among extensive relevant research endeavors, most dependency test methods can be categorized as either model-based or model-free.

A model-based method typically infers the dependency between multiple functions by first assuming a functional regression model (see, e.g., Morris, 2015, for a survey) which characterizes their structural relationship, and then testing the significance of the assumed model. See examples of model-based methods by Guo (2002); Huang et al. (2002); Shen and Faraway (2004); Antoniadis and Sapatinas (2007) for concurrent/varying-coefficient models and by Kokoszka et al. (2008); Chen et al. (2020) for function-on-function regression models. The main disadvantage of a model-based method is its reliance on correct model specification. If the model is misspecified, the inference is not well grounded and might be inaccurate.

A model-free method can avoid the misspecification issue associated with model-based methods since it typically quantifies the dependency between random functions by a correlation measure, without assuming any particular model. As a natural extension of the canonical correlation for multivariate data, the functional canonical correlation is a popular correlation measure for functional data (e.g., Leurgans et al., 1993; He et al., 2003; Eubank and Hsing, 2008; Shin and Lee, 2015). However, it is plagued by the involvement of inverting a covariance operator, which is an ill-posed problem and often requires proper regularizations. The

dynamical correlation (Dubin and Müller, 2005; Sang et al., 2019) and temporal correlation (Zhou et al., 2018) are two functional correlation measures without the aforementioned inverse problem. The former measures the angle between two random functions in the L^2 space. The latter essentially computes the Pearson correlation between two random functions at each time point and then averages all pointwise Pearson correlations over the time domain. However, since uncorrelatedness does not imply independence, these functional correlations are insufficient to test independence. Recently a few model-free approaches have been developed to test mean independence for functional data (e.g., Patilea et al., 2016; Lee et al., 2020), but they can only test a weaker notion of independence.

In this paper we develop a model-free independence test for functional data. Under the reproducing kernel Hilbert space (RKHS) framework, we propose to use the Hilbert-Schmidt Independence Criterion (HSIC, e.g., Gretton et al., 2005, 2008) to measure the dependency between two random functions. An appealing property is that HSIC endowed with characteristic kernels is zero if and only if the two random functions are independent. However, the application of HSIC requires fully observed and noiseless functional data, while in practice functional data are always discretely measured and contaminated by noise. To tackle this problem, one may perform a two-step procedure: first pre-smooth the data, and then apply HSIC to the resulting functions. Clearly, pre-smoothing will affect the performance of HSIC. Indeed, the functional distance with respect to which the asymptotic convergence of the pre-smoothing procedure is measured is crucial, as HSIC is fundamentally based on a functional distance. Some common pre-smoothing procedures do not have existing convergence results on the required functional distance, and hence may not be compatible; namely, the pre-smoothing error may have a profound effect on the subsequent HSIC. See Section 3 for more discussion. In this work, we carefully design our procedure to ensure that the two steps are compatible. For the first step, we propose a new wavelet thresholding method while we use Besov-norm-induced kernels for HSIC in the second step. We can show that these choices in the two steps are theoretically compatible if the functional data are sufficiently densely mea-

sured. See Section 4 for details. Our work is motivated by the Human Connectome Project (HCP, https://www.humanconnectome.org) from which various brain imaging datasets are publicly accessible. In Section 7, the application of our method to a magnetoencephalography (MEG) dataset from HCP is capable of identifying anatomically interpretable functional connectivity patterns, suggesting a great potential of the proposed method in the study of functional connectivity between brain regions.

The main contribution of this paper is three-fold. First, we design some suitable kernels such that the corresponding HSIC can identify the independence of a pair of random functions of which sample paths belong to Besov spaces, a larger class of functions than Sobolev spaces which are popular in RKHS modeling. We propose to use the Besov sequence norm for the wavelet coefficients of these random functions induce such kernel, which is shown to be characteristic. Second, for dense functional data, we develop the asymptotic distribution of the empirical HSIC based on pre-smoothed functions by wavelet thresholding. To theoretically guarantee the compatibility between the pre-smoothing and empirical HSIC, we propose a new wavelet thresholding method that can efficiently reduce the pre-smoothing error measured by the Besov sequence norm used in the empirical HSIC when the noise is nearly independent. Since the asymptotic distribution involves many unknown quantities, we suggest a permutation test in practice and prove that not only can the test control the Type I error probability but also it is consistent. The theoretical results show that the two steps in our proposed procedure are compatible. Finally, we propose a data-adaptive approach to tuning the smoothness parameter for the Besov norm needed to induce the kernel for HSIC. It is numerically shown that this approach is able to enhance the sensitivity of HSIC to detecting dependencies at high frequencies.

The rest of the paper proceeds as follows. Section 2 provides a brief introduction to HSIC. The two-step procedure for the proposed wavelet-based HSIC test is given in Section 3. Its asymptotic properties are presented in Section 4. Section 5 discusses tuning parameter selection. The numerical performance of the proposed method is illustrated in a simulation

study in Section 6 and an MEG functional connectivity study in Section 7 where it is also compared with representative existing methods. Section 8 concludes the paper. The code to implement the proposed method is publicly available on GitHub (`https://github.com/rui-miao/wavHSIC`).

2 Hilbert-Schmidt Independence Criterion

In this section we give a brief introduction to HSIC. Let X and Y be two random functions of which sample paths belong to function spaces $\mathcal{X}$ and $\mathcal{Y}$ respectively, and $\mathcal{H}(\kappa_{\mathcal{X}})$ and $\mathcal{H}(\kappa_{\mathcal{Y}})$ be the RKHS equipped with kernels $\kappa_{\mathcal{X}}$ and $\kappa_{\mathcal{Y}}$ defined on $\mathcal{X} \times \mathcal{X}$ and $\mathcal{Y} \times \mathcal{Y}$ respectively.

HSIC requires that both $\kappa_{\mathcal{X}}$ and $\kappa_{\mathcal{Y}}$ are characteristic, in the sense that two probability measures $P = P'$ if and only if $\mathbf{P}^{\kappa_{\mathcal{Z}}}(P) = \mathbf{P}^{\kappa_{\mathcal{Z}}}(P')$ where $\mathbf{P}^{\kappa_{\mathcal{Z}}}(P) = E_P\{\kappa_{\mathcal{Z}}(Z, \cdot)\}$ for a random function $Z \in \mathcal{Z}$ which follows P and $(Z, \mathcal{Z}) = (X, \mathcal{X})$ or $(Y, \mathcal{Y})$. A characteristic kernel may be induced by a strong negative type semi-metric (see Definition S1 and Proposition S1 in the supplementary material). Denote the joint probability measure of X and Y by P_{XY} and their marginal probability measures by P_X and P_Y respectively. Since $\kappa_{\mathcal{X}}$ and $\kappa_{\mathcal{Y}}$ are characteristic, P_X and P_Y are fully characterized by $\mathbf{P}^{\kappa_{\mathcal{X}}}(P_X) = E_{P_X}\{\kappa_{\mathcal{X}}(X, \cdot)\}$ and $\mathbf{P}^{\kappa_{\mathcal{Y}}}(P_Y) = E_{P_Y}\{\kappa_{\mathcal{Y}}(Y, \cdot)\}$ respectively. Let $\mathbf{P}^{\kappa_{\mathcal{X}} \otimes \kappa_{\mathcal{Y}}}(P_{XY}) = E_{P_{XY}}\{(\kappa_{\mathcal{X}} \otimes \kappa_{\mathcal{Y}})((X, Y), (*, \cdot))\}$, where the tensor product kernel $\kappa_{\mathcal{X}} \otimes \kappa_{\mathcal{Y}}$ is defined by $(\kappa_{\mathcal{X}} \otimes \kappa_{\mathcal{Y}})((x, y), (x', y')) = \kappa_{\mathcal{X}}(x, x')\kappa_{\mathcal{Y}}(y, y'), x, x' \in \mathcal{X}, y, y' \in \mathcal{Y}$.

Sejdinovic et al. (2013) showed that X and Y are independent, i.e., $P_{XY} = P_X P_Y$, if and only if $\mathbf{P}^{\kappa_{\mathcal{X}} \otimes \kappa_{\mathcal{Y}}}(P_{XY}) = \mathbf{P}^{\kappa_{\mathcal{X}}}(P_X)\mathbf{P}^{\kappa_{\mathcal{Y}}}(P_Y)$, although $\kappa_{\mathcal{X}} \otimes \kappa_{\mathcal{Y}}$ is not characteristic for all probability measures on $\mathcal{H}(\kappa_{\mathcal{Y}}) \times \mathcal{H}(\kappa_{\mathcal{Y}})$. Therefore, to test the independence between X and Y, it suffices to study the difference between $\mathbf{P}^{\kappa_{\mathcal{X}} \otimes \kappa_{\mathcal{Y}}}(P_{XY})$ and $\mathbf{P}^{\kappa_{\mathcal{X}}}(P_X)\mathbf{P}^{\kappa_{\mathcal{Y}}}(P_Y)$. Since $\mathbf{P}^{\kappa_{\mathcal{X}}}(P_X) \in \mathcal{H}(\kappa_{\mathcal{X}})$, $\mathbf{P}^{\kappa_{\mathcal{Y}}}(P_Y) \in \mathcal{H}(\kappa_{\mathcal{Y}})$ and $\mathbf{P}^{\kappa_{\mathcal{X}} \otimes \kappa_{\mathcal{Y}}}(P_{XY}) \in \mathcal{H}(\kappa_{\mathcal{X}} \otimes \kappa_{\mathcal{Y}})$ where $\mathcal{H}(\kappa_{\mathcal{X}} \otimes \kappa_{\mathcal{Y}})$ is the RKHS equipped with $\kappa_{\mathcal{X}} \otimes \kappa_{\mathcal{Y}}$, HSIC may be used to measure this difference under the norm of $\mathcal{H}(\kappa_{\mathcal{X}} \otimes \kappa_{\mathcal{Y}})$.

Definition 1 (HSIC). *Suppose that* $\int_{\mathcal{X}} \kappa_{\mathcal{X}}(x,x)\mathrm{d}P_X(x) < \infty$ *and* $\int_{\mathcal{Y}} \kappa_{\mathcal{Y}}(y,y)\mathrm{d}P_Y(y) < \infty$. *The HSIC of* P_{XY} *is defined by*

$$\begin{aligned}\gamma(P_{XY}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}}) &= \|\mathbf{P}^{\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}}}(P_{XY}) - \mathbf{P}^{\kappa_{\mathcal{X}}}(P_X)\mathbf{P}^{\kappa_{\mathcal{Y}}}(P_Y)\|^2_{\mathcal{H}(\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}})} \\ &= 4\int_{\mathcal{X}\times\mathcal{Y}}\int_{\mathcal{X}\times\mathcal{Y}} \kappa_{\mathcal{X}}(x,x')\kappa_{\mathcal{Y}}(y,y')\mathrm{d}(P_{XY}-P_XP_Y)(x,y)\mathrm{d}(P_{XY}-P_XP_Y)(x',y').\end{aligned}$$

In practice with $\{(X_i, Y_i) : i = 1, \ldots, n\}$ which are independently and identically distributed (i.i.d.) copies of (X, Y), the sample versions of $\mathbf{P}^{\kappa_{\mathcal{X}}}(P_X)$, $\mathbf{P}^{\kappa_{\mathcal{Y}}}(P_Y)$ and $\mathbf{P}^{\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}}}(P_{XY})$ are defined by $\mathbf{P}^{\kappa_{\mathcal{X}}}(P_{n,X}) = n^{-1}\sum_{i=1}^n \kappa_{\mathcal{X}}(X_i, \cdot)$, $\mathbf{P}^{\kappa_{\mathcal{Y}}}(P_{n,Y}) = n^{-1}\sum_{i=1}^n \kappa_{\mathcal{Y}}(Y_i, \cdot)$, and $\mathbf{P}^{\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}}}(P_{n,XY}) = n^{-1}\sum_{i=1}^n\{(\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}})((X_i,Y_i),(*,\cdot))\}$. Obviously $\mathbf{P}^{\kappa_{\mathcal{X}}}(P_{n,X}) \in \mathcal{H}(\kappa_{\mathcal{X}})$, $\mathbf{P}^{\kappa_{\mathcal{Y}}}(P_{n,Y}) \in \mathcal{H}(\kappa_{\mathcal{Y}})$ and $\mathbf{P}^{\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}}}(P_{n,XY}) \in \mathcal{H}(\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}})$, so we can obtain a sample version of HSIC as follows.

Definition 2 (Empirical HSIC). *Under the same setting in Definition 1, the empirical HSIC, which is an estimator of HSIC, is defined by*

$$\begin{aligned}&\gamma(P_{n,XY}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}}) = \|\mathbf{P}^{\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}}}(P_{n,XY}) - \mathbf{P}^{\kappa_{\mathcal{X}}}(P_{n,X})\mathbf{P}^{\kappa_{\mathcal{Y}}}(P_{n,Y})\|^2_{\mathcal{H}(\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}})} \\ &= 4\int_{\mathcal{X}\times\mathcal{Y}}\int_{\mathcal{X}\times\mathcal{Y}} \kappa_{\mathcal{X}}(x,x')\kappa_{\mathcal{Y}}(y,y')\mathrm{d}(P_{n,XY}-P_{n,X}P_{n,Y})(x,y)\mathrm{d}(P_{XY}-P_{n,X}P_{n,Y})(x',y').\end{aligned}$$

By Sejdinovic et al. (2013), the empirical HSIC can be rewritten as

$$\gamma(P_{n,XY}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}}) = n^{-2}\mathrm{tr}(\mathbf{\Gamma}^X\mathbf{H}\mathbf{\Gamma}^Y\mathbf{H}),$$

where $\mathbf{\Gamma}^X = (\kappa_{\mathcal{X}}(X_i, X_j))_{1\le i,j\le n}$ and $\mathbf{\Gamma}^Y = (\kappa_{\mathcal{Y}}(Y_i, Y_j))_{1\le i,j\le n}$ are Gram matrices, and $\mathbf{H} = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top$ is the centering matrix with the $n \times n$ identity matrix $\mathbf{I}_n$ and $\mathbf{1}_n = (1, \ldots, 1)^\top$ of dimension n.

3 Methodology

Suppose that bivariate functional data $\{(X_i, Y_i) : i = 1, \ldots, n\}$ collected from n subjects are i.i.d. copies of a pair of random functions (X, Y), which, without loss of generality, is defined on the domain $[0, 1] \times [0, 1]$. Let the sample paths of X and Y belong to function spaces $\mathcal{X}$ and

$\mathcal{Y}$ respectively. In many applications such as brain imaging analysis, the measurements of each function are sampled at a discrete and regular grid and subject to noise contamination. Hence we assume that the observations are $\{(\tilde{X}_{il}, \tilde{Y}_{il}) := (X_i(T_l) + e_{il}^X, Y_i(T_l) + e_{il}^Y) : i = 1, \ldots, n; l = 1, \ldots, m\}$, where $\{T_l = (l-1)/m : l = 1, \ldots, m\}$ is a regular grid with $m = 2^{J+1}$ for some integer $J > 0$ and the two sets of mean-zero random noise, $\{e_{il}^X : i = 1, \ldots, n; l = 1, \ldots, m\}$ and $\{e_{il}^Y : i = 1, \ldots, n; l = 1, \ldots, m\}$, are independent of each other and of $\{(X_i, Y_i) : i = 1, \ldots, n\}$. The error terms in each set are further assumed to be identically distributed, independent across subjects, but *possibly* dependent within each subject. We defer the discussion on the error dependence structures to Section 4. Our goal is to formulate an HSIC-based test for the independence between X and Y via $\{(\tilde{X}_{il}, \tilde{Y}_{il}) : i = 1, \ldots, n; l = 1, \ldots, m\}$. For simplicity we assume that all functions share the same measurement grid and $m = 2^{J+1}$, but the proposed method is applicable with minor modifications if the grid is irregular, the functions are measured at different grids, or $m \neq 2^{J+1}$ (see Remark 1).

Due to the success of existing HSIC-based independence tests for multivariate data, it is tempted to treat the discretized observations as multivariate data and directly apply existing methods. However, there are two issues with this approach. First, in order to capture enough information, m should be large enough, which naturally leads to high-dimensional data. Without reasonable structure across these m dimensions, HSIC does not perform well. In the FDA literature, modeling the sample paths with certain form of smoothness has been shown an empirically successful strategy in many applications. It is beneficial to incorporate smoothness structure during the design of a tailor-made HSIC method. Second, the discretized observations are contaminated by noise. Hence these raw observations are indeed not "smooth" but the noiseless ones are.

The proposed method is directly based on the definition of HSIC (Definition 1) when applied to random functions. Clearly, the application of such HSIC requires the trajectories of all random functions to be fully observed and noiseless. Thus, with discrete and noisy measurements in practice, a natural idea is to perform pre-smoothing to recover these trajectories

followed by applying HSIC to random functions. However, the compatibility of these two steps is generally unclear. Namely, it is non-trivial to know whether the pre-smoothing error (measured in a certain norm) would have a profound effect on the subsequent HSIC-based test. For instance, if the sample paths of all random functions are assumed to belong to a Sobolev space, it is seemingly reasonable to pre-smooth each trajectory by a smoothing spline followed by the HSIC based on Sobolev-norm-induced kernels. However, the compatibility of the two steps is unknown since there is no theoretical result to guarantee that the pre-smoothing error under a Sobolev norm converges to zero, although the corresponding results with respect to the L^2 or empirical norm exist.

To address this compatibility issue, we propose to use HSIC based on Besov-norm-induced kernels for testing independence under the assumption that the sample paths of all random functions belong to Besov spaces, a larger class of functions than Sobolev spaces. To recover each trajectory, we develop a new wavelet thresholding method for pre-smoothing. Its theoretical compatibility with the proposed HSIC is given in Section 4. In the rest of this section, we first introduce wavelets (e.g., Ogden, 1997; Vidakovic, 2009; Morettin et al., 2017) together with other related results and then the details of the proposed two-step procedure.

3.1 Wavelets and Besov Sequence Norms

Following the Cohen-Daubechies-Jawerth-Vial (CDJV) construction (Cohen et al., 1993), let father and mother wavelets be $\phi, \psi \in C^R[0,1]$ respectively with D vanishing moments (e.g., Daubechies, 1992) where $C^R[0,1]$ is the space of all functions on $[0,1]$ with R-th order continuous derivatives. We consider a Besov space $B^{\alpha}_{p,q}[0,1]$ with norm $\|\cdot\|_{B^{\alpha}_{p,q}[0,1]}$ of which smoothness parameter α satisfies $1/p < \alpha < \min\{R, D\}$ such that $B^{\alpha}_{p,q}[0,1]$ can be embedded continuously in $C[0,1]$. Formal definitions of $B^{\alpha}_{p,q}[0,1]$ and its norm $\|\cdot\|_{B^{\alpha}_{p,q}[0,1]}$ are given in Section S1.2 in the supplementary material. Then for any function $f \in B^{\alpha}_{p,q}[0,1] \cap L^2[0,1]$ and a fixed coarse scale L, we have the following decomposition

$$f(t) = \sum_{k=0}^{2^L-1} \xi_k \{2^{L/2}\phi(2^L t - k)\} + \sum_{j \geq L} \sum_{k=0}^{2^j-1} \theta_{j,k} \{2^{j/2}\psi(2^j t - k)\}, \quad t \in [0,1]. \tag{1}$$

Denote $\theta_{j,k} = \xi_{2^j+k}, 0 \leq j < L, 0 \leq k < 2^j$ and $\theta_{-1,0} = \xi_0$. Based on the wavelet coefficients of f, $\boldsymbol{\theta}^f = ((\boldsymbol{\theta}^f_{-1})^\top, (\boldsymbol{\theta}^f_0)^\top, \ldots, (\boldsymbol{\theta}^f_L)^\top, (\boldsymbol{\theta}^f_{L+1})^\top, \ldots)^\top$ where $\boldsymbol{\theta}^f_j = (\theta_{j,0}, \theta_{j,1}, \ldots, \theta_{j,2^j-1})^\top$ and $\boldsymbol{\theta}^f_{-1} = \theta_{-1,0}$, the Besov sequence norm $\|\cdot\|_{b^\alpha_{p,q}}$ (e.g., Donoho et al., 1995; Johnstone and Silverman, 2005) is defined by

$$\|\boldsymbol{\theta}^f\|_{b^\alpha_{p,q}} = \left(\sum_{j \geq -1} 2^{jsq} \|\boldsymbol{\theta}^f_j\|_p^q \right)^{1/q}, \quad s = \alpha + 1/2 - 1/p, \tag{2}$$

where $\|\cdot\|_p$ refers to the ℓ_p-norm for vectors. Denote the corresponding space by $b^\alpha_{p,q} = \{\mathbf{a} : \|\mathbf{a}\|_{b^\alpha_{p,q}} < \infty\}$. Note that the two norms $\|\cdot\|_{B^\alpha_{p,q}}$ and $\|\cdot\|_{b^\alpha_{p,q}}$ are equivalent (e.g., DeVore and Lorentz, 1993; Donoho et al., 1995) and obviously $b^\alpha_{p,q} \subset b^\beta_{p,q}$ if $\beta \leq \alpha$. In practice, if f is observed at $m = 2^{J+1}$ dyadic time points $\{0/m, 1/m, \ldots, (m-1)/m\}$, the discrete wavelet transformation can be used to calculate the wavelet coefficients $\boldsymbol{\theta}^f$ with $\boldsymbol{\theta}^f_j = \mathbf{0}$ when $j > J$. Then we can denote $\boldsymbol{\theta}^f = ((\boldsymbol{\theta}^f_{-1})^\top, (\boldsymbol{\theta}^f_0)^\top, \ldots, (\boldsymbol{\theta}^f_J)^\top)^\top$.

We can show that some Besov sequence norm can induce a characteristic kernel, which is required by HSIC.

Theorem 1. *For $0 < q' < q \leq p \leq 2$, $0 \leq \alpha \leq \alpha'$ and $\alpha' > 1/p$, let the semi-metric $\rho_{b^\alpha_{p,q}}(f,g) = \|\boldsymbol{\theta}^f - \boldsymbol{\theta}^g\|^{q'}_{b^\alpha_{p,q}}$ for $f, g \in B^{\alpha'}_{p,q}[0,1]$, where $\boldsymbol{\theta}^f$ and $\boldsymbol{\theta}^g$ are the wavelet coefficients of f and g respectively. The function induced by $\rho_{b^\alpha_{p,q}}$, which is $\kappa(z,z') = \rho_{b^\alpha_{p,q}}(z,0) + \rho_{b^\alpha_{p,q}}(z',0) - \rho_{b^\alpha_{p,q}}(z,z')$, $z, z' \in B^{\alpha'}_{p,q}[0,1]$, is a characteristic kernel.*

The proof of Theorem 1 is given in Section S2.1 in the supplementary material. By Theorem 1, we can define HSIC properly based on kernels induced by Besov sequence norms. For simplicity, hereafter we focus on popular choices of $p = q = 2$ and $q' = 1$. Accordingly we abbreviate $B^\alpha_{2,2}[0,1]$ and $b^\alpha_{2,2}$ to B^α and b^α respectively, and the kernel functions are

$$\kappa_{\mathcal{Z}}(z_1, z_2) = \|\boldsymbol{\theta}^{z_1}\|_{b^{\beta_Z}} + \|\boldsymbol{\theta}^{z_2}\|_{b^{\beta_Z}} - \|\boldsymbol{\theta}^{z_1} - \boldsymbol{\theta}^{z_2}\|_{b^{\beta_Z}}, \quad z_1, z_2 \in \mathcal{Z}, 0 \leq \beta_Z \leq \alpha_Z,$$

for $(Z, \mathcal{Z}) = (X, \mathcal{X})$ and $(Y, \mathcal{Y})$.

3.2 Two-Step Procedure

Let $Z = X$ or Y. Under the setting in Section 3.1, we assume $Z \in B^{\alpha_Z}$ where $1/2 < \alpha_Z < \min\{R, D\}$. Note that $B^{\alpha_Z} \subset B^{\beta_Z}$ for $0 < \beta_Z < \alpha_Z$ so $Z \in B^{\beta_Z}$ as well. To test the independence between X and Y based on their discretely measured and noisy observations, we propose to first denoise each function and then apply HSIC to the recovered functions. The two-step procedure is explicitly stated as follows:

Step 1 By the decomposition (1) and the resolution limitation due to a finite number of measurements $m = 2^{J+1}$ taken for each subject, we obtain the initial wavelet coefficient estimates for each Z_i, denoted by $\boldsymbol{\theta}^{\tilde{Z}_i} = ((\boldsymbol{\theta}_{-1}^{\tilde{Z}_i})^\top, (\boldsymbol{\theta}_0^{\tilde{Z}_i})^\top, \ldots, (\boldsymbol{\theta}_J^{\tilde{Z}_i})^\top)^\top$, via the discrete wavelet transformation with the coarse scale L_Z. The coarse scale L_Z may be selected by cross-validation or domain knowledge. We propose to denoise $\boldsymbol{\theta}^{\tilde{Z}_i}$ and accordingly obtain $\boldsymbol{\theta}^{\hat{Z}_i} = ((\boldsymbol{\theta}_{-1}^{\hat{Z}_i})^\top, (\boldsymbol{\theta}_0^{\hat{Z}_i})^\top, \ldots, (\boldsymbol{\theta}_J^{\hat{Z}_i})^\top)^\top$ as follows. First, we let $\boldsymbol{\theta}_j^{\hat{Z}_i} = \boldsymbol{\theta}_j^{\tilde{Z}_i}$ for $j = -1, \ldots, L_Z - 1$. Moreover, we apply the following penalized least squares to obtain $\boldsymbol{\theta}_j^{\hat{Z}_i}, j = L_Z, \ldots, J$:

$$\boldsymbol{\theta}_j^{\hat{Z}_i} = \arg\min_{\boldsymbol{\theta}_j} \left\{ \|\boldsymbol{\theta}_j^{\tilde{Z}_i} - \boldsymbol{\theta}_j\|_2^2 + \delta_{Z,j}^2 \mathrm{pen}_j(\|\boldsymbol{\theta}_j\|_0) \right\}, \quad j = L_Z, \ldots, J, \tag{3}$$

where $\|\cdot\|_2$ denotes the Euclidean norm, $\|\cdot\|_0$ denotes the number of non-zero elements, $\delta_{Z,j} = 2^{\varsigma_Z j}\delta_Z$ is the noise standard deviation at the resolution level j with $\delta_Z > 0$, and the penalty $\mathrm{pen}_j(k) = k\zeta_Z\{1 + \sqrt{2(1+2\varsigma_Z)\log(\tau_j m_j/k)}\}^2$ that depends on $\zeta_Z > 1$, $\varsigma_Z > -1/2$, $m_j = 2^j$, and $\tau_j = \tau_Z 2^{2\alpha_Z(j - j_\#^Z)_+}$ with $\tau_Z > e$ and $j_\#^Z = (1 + (\varsigma_Z + 1/2)/\alpha_Z)/(\alpha_Z + \varsigma_Z + 1/2) \cdot \log_2 \delta_Z^{-1}$.

The proposed procedure in (3) is capable of denoising a certain type of correlated noise (see technical assumptions in Theorem 2 in Section 4). Compared to the penalty (12.34) in Johnstone (2019), we employ a different τ_j in the penalty in (3) such that the pre-smoothing error measured by the Besov sequence norm used in the empirical HSIC in Step 2 below converges to zero if m diverges to infinity (see Theorem 2 in Section 4). This can guarantee the compatibility between this and the next steps.

Similar to Johnstone and Paul (2014), to obtain the estimate $\boldsymbol{\theta}_j^{\hat{Z}_i}$, $L_Z \le j \le J$, defined in

(3), one may apply the level-wise hard thresholding as follows: For each level $L_Z \le j \le J$, let $|\theta^{\tilde{Z}_i}_{j,(k)}|$ be the k-th term after the elements of $\boldsymbol{\theta}^{\tilde{Z}_i}_j$ are sorted in a decreasing order of their absolute values, namely $|\theta^{\tilde{Z}_i}_{j,(0)}| \ge |\theta^{\tilde{Z}_i}_{j,(1)}| \ge \cdots \ge |\theta^{\tilde{Z}_i}_{j,(2^j-1)}|$. Then the hard threshold at level j is $\delta_{Z,j}\sqrt{\mathrm{pen}_j(\hat{k}_j) - \mathrm{pen}_j(\hat{k}_j - 1)}$, where $\hat{k}_j = \arg\min_{k \ge 0}\left\{\sum_{k' \ge k} |\theta^{\tilde{Z}_i}_{j,(k')}|^2 + \delta^2_{Z,j}\mathrm{pen}_j(k)\right\}$. Detailed steps of solving (3) are summarized in Algorithm 1. The discussion of tuning parameter selection is deferred to Section 5.

Step 2 Since the wavelet coefficient estimates $\boldsymbol{\theta}^{\hat{X}_i} \in b^{\alpha_X} \subset b^{\beta_X}$ and $\boldsymbol{\theta}^{\hat{Y}_i} \in b^{\alpha_Y} \subset b^{\beta_Y}$, $i = 1, \ldots, n$, for $\beta_X < \alpha_X$ and $\beta_Y < \alpha_Y$, we may apply HSIC to the denoised functions where the kernels $\kappa_{\mathcal{X}}$ and $\kappa_{\mathcal{Y}}$ are induced by $\rho_{b^{\beta_X}}$ and $\rho_{b^{\beta_Y}}$ respectively as defined in Theorem 1. Explicitly, we have $\gamma(P_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}}) = n^{-2}\mathrm{tr}(\boldsymbol{\Gamma}^{\hat{X}}\mathbf{H}\boldsymbol{\Gamma}^{\hat{Y}}\mathbf{H})$, where

$$\boldsymbol{\Gamma}^{\hat{X}} = \left(\|\boldsymbol{\theta}^{\hat{X}_i}\|_{b^{\beta_X}} + \|\boldsymbol{\theta}^{\hat{X}_j}\|_{b^{\beta_X}} - \|\boldsymbol{\theta}^{\hat{X}_i} - \boldsymbol{\theta}^{\hat{X}_j}\|_{b^{\beta_X}}\right)_{1 \le i,j \le n}, \quad \text{and}$$
$$\boldsymbol{\Gamma}^{\hat{Y}} = \left(\|\boldsymbol{\theta}^{\hat{Y}_i}\|_{b^{\beta_Y}} + \|\boldsymbol{\theta}^{\hat{Y}_j}\|_{b^{\beta_Y}} - \|\boldsymbol{\theta}^{\hat{Y}_i} - \boldsymbol{\theta}^{\hat{Y}_j}\|_{b^{\beta_Y}}\right)_{1 \le i,j \le n}.$$

By adopting $\rho_{b^{\beta_X}}$ and $\rho_{b^{\beta_Y}}$ where $\beta_X < \alpha_X$ and $\beta_Y < \alpha_Y$ to construct kernels, we are able to make the pre-smoothing step theoretically compatible with the HSIC. As revealed in Theorems 2 and 3 in Section 4 below, if the observations of all functions are sufficiently dense, the denoising error is asymptotically negligible in the asymptotic distribution of the HSIC. This is a key benefit of using wavelets and Besov norms for pre-smoothing.

In Section 4, the asymptotic distribution of $\gamma(P_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}})$ is developed in Theorem 3 under the independence hypothesis. Despite its theoretical appeal, the asymptotic distribution unfortunately involves many unknown quantities. Therefore, we suggest using permutations to perform the independence test which, as shown in Theorem 4, can control the Type I error probability and is also consistent.

Remark 1. *Since denoising is performed separately for each function and subject, the proposed method is applicable when the functions of different subjects are not measured at the same grid. For $m \ne 2^{J+1}$ at possibly irregular and uncommon designs, linear interpolation may be applied*

if the original measurement resolution is sufficiently high (e.g., Kovac and Silverman, 2000). We demonstrate the satisfactory performance of this strategy via a simulation study, and the corresponding results are given in Section S3 of the supplementary material.

Algorithm 1: Solving (3) by wavelet thresholding.

Input : $\{\boldsymbol{\theta}_j^{\tilde{Z}_i} : j = L_Z, \ldots, J\}$;
fixed tuning parameters $\zeta_Z > 1, \tau_Z > e, \varsigma_Z > -1/2, \delta_Z > 0$.

Output: $\{\boldsymbol{\theta}_j^{\hat{Z}_i} : j = L_Z, \ldots, J\}$.

for $j \leftarrow L_Z$ **to** J **do**
  $|\theta_{j,(0)}^{\tilde{Z}_i}| \geq |\theta_{j,(1)}^{\tilde{Z}_i}| \geq \cdots \geq |\theta_{j,(2^j-1)}^{\tilde{Z}_i}| \leftarrow$ sort $\{|\theta_{j,k}^{\tilde{Z}_i}|\}_{k=0}^{2^j-1}$ in descending order;
  $\theta_j^{\text{Thresh}} \leftarrow \delta_{Z,j}\sqrt{\text{pen}_j(\hat{k}_j) - \text{pen}_j(\hat{k}_j - 1)}$, where
  $\hat{k}_j = \arg\min_{k \geq 0}\left\{\sum_{k' \geq k} |\theta_{j,(k')}^{\tilde{Z}_i}|^2 + \delta_{Z,j}^2 \text{pen}_j(k)\right\}$;
  for $k \leftarrow 0$ **to** $2^j - 1$ **do**
    $\theta_{j,k}^{\hat{Z}_i} \leftarrow \theta_{j,k}^{\tilde{Z}_i} \mathbb{I}(|\theta_{j,k}^{\tilde{Z}_i}| \geq \theta_j^{\text{Thresh}})$
  end
end

Remark 2. *In Step 1, the time complexity for the discrete wavelet transformation is $O(m)$ for each subject (Cohen et al., 1993) and so is that for denoising. In Step 2, the time complexity for calculating Gram matrices is $O(mn^2)$ and so is that for calculating the empirical HSIC. Therefore, the permutation test based on B permutations requires $O(Bmn^2)$ of time. In addition to the time complexity analysis, we report the computing time for the proposed method when applied to the MEG data in Section 7.*

4 Asymptotic Theory

In this section we show that the proposed two-step procedure can lead to an asymptotically valid test, which addresses the compatibility issue raised in Section 3. Explicitly, we first

provide the rate of convergence for the denoising error involved in Step 1 in Theorem 2, then the asymptotic distribution of HSIC $\gamma(P_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}})$ in Step 2 in Theorem 3, and finally the asymptotic properties of the permutation test in Theorem 4. Hereafter, the kernels $\kappa_{\mathcal{X}}$ and $\kappa_{\mathcal{Y}}$ are induced by $\rho_{b^{\beta_X}}$ and $\rho_{b^{\beta_Y}}$ respectively. For the noise terms $\{e_{il}^Z : i = 1, \ldots, n; l = 1, \ldots, m\}$ where $Z = X$ or Y, we assume that $e_{il}^Z = e_i^Z(T_l), l = 1, \ldots, m$ where $\{e_i^Z : i = 1, \ldots, n\}$ are i.i.d. copies of a stationary stochastic process e^Z.

Theorem 2. *Assume that* $\beta_Z < \alpha_Z$, $\|\boldsymbol{\theta}^Z\|_{b^{\alpha_Z}} \leq C_Z$ *for a constant* $C_Z > 0$, *and* $\boldsymbol{\theta}^{e^Z} = (\theta_{j,k}^{e^Z})_{-1 \leq j \leq J_Z; k=0,\ldots,2^j-1}$, *the discrete wavelet coefficients of* e^Z, *satisfy* $\theta_{j,k}^{e^Z} = 2^{\varsigma_Z j}\delta_Z z_{j,k}$ *where* $\varsigma_Z > -1/2$ *and* $\mathbf{z} = (z_{j,k})_{-1 \leq j \leq J_Z; k=0,\ldots,2^j-1}$ *is a zero mean Gaussian random vector that is weakly correlated, i.e., its covariance matrix* $\boldsymbol{\Sigma}$ *satisfies* $\xi_0^Z \mathbf{I} \preceq \boldsymbol{\Sigma} \preceq \xi_1^Z \mathbf{I}$ *where* $\mathbf{I}$ *is the identity matrix,* $0 < \xi_0^Z \leq 1 \leq \xi_1^Z < \infty$ *are constants, and* $\mathbf{A} \preceq \mathbf{B}$ *means that* $\mathbf{B} - \mathbf{A}$ *is positive semidefinite. Then for* $\boldsymbol{\theta}^{Z_i}$ *obtained by (3), we have*

$$\sup_{\|\boldsymbol{\theta}^{Z_i}\|_{b^{\alpha_Z}} \leq C_Z} \mathrm{E}\left(\|\boldsymbol{\theta}^{\hat{Z}_i} - \boldsymbol{\theta}^{Z_i}\|_{b^{\beta_Z}}^2 \mid \boldsymbol{\theta}^{Z_i}\right) = O(\delta_Z^{2r}) = O(m^{-r}), \quad \textit{uniformly for } i = 1, \ldots, n,$$

as $m \to \infty$, *where* $r = (\alpha_Z - \beta_Z)/(\alpha_Z + \varsigma_Z + 1/2)$. *This implies that* $\|\boldsymbol{\theta}^{\hat{Z}_i} - \boldsymbol{\theta}^{Z_i}\|_{b^{\beta_Z}}^2 = O_p(m^{-r})$ *uniformly for* $i = 1, \ldots, n$ *as* $m \to \infty$.

The proof of Theorem 2 is given in Section S2.2 in the supplementary material. Theorem 2 shows that the pre-smoothing error under the Besov sequence norm b^{β_Z} converges to zero uniformly for all subjects if m diverges to infinity. This theoretical guarantee is achieved due to the new penalty in the proposed wavelet thresholding method (3). The assumption on the noise $\theta_{j,k}^{e^Z} = 2^{\varsigma_Z j}\delta_Z z_{j,k}$ where $\mathbf{z} = (z_{j,k})_{-1 \leq j \leq J_Z; k=0,\ldots,2^j-1}$ is weakly correlated Gaussian is a generalization of the Gaussian white noise model by allowing correlation among noise terms to some extent. First, the assumption encompasses both short- and long-range dependences of the noise process when it is a stationary and Gaussian (Johnstone and Silverman, 1997). For the short-range dependence case where $\varsigma_Z = 0$, there is no variance inflation with the increase of level j. For the long-range dependence case where $-1/2 < \varsigma_Z < 0$, the process

$e_i^m(t) = m^{-1}\sum_{l=1}^{\lfloor mt \rfloor} e_{il}^Z$ can be approximated by a fractional Brownian motion $\delta_Z^{2-2H} B_H(t)$, $H = 1/2 - \varsigma_Z$ (Taqqu, 1975), which is widely used for modeling long-range dependence. Then the convergence rate (with δ_Z replaced by $\delta_Z^{2-2H} = \delta_Z^{1+2\varsigma_Z}$ in (2)) is asymptotically minimax up to a constant. When $\beta_Z = 0$ in particular, this rate coincides with those of Wang (1996) and Johnstone and Silverman (1997). Second, when $\varsigma_Z > 0$, ς_Z captures noise amplification as reflected in the noise level $\delta_{Z,j} = 2^{\varsigma_Z j}\delta_Z$, which is common in the linear inverse problem (Abramovich and Silverman, 1998; Johnstone and Paul, 2014), e.g., $\varsigma_Z = 1/2$ for the two-dimensional Radon transformation (Donoho, 1995).

Since the HSIC is constructed based on the kernels induced by $\rho_{b^{\beta_X}}$ and $\rho_{b^{\beta_Y}}$, the same norms used to evaluate the denoising error as in Theorem 2, the compatibility between the pre-smoothing by wavelet soft-thresholding and HSIC is theoretically guaranteed. As shown in Theorem 3, the effect of the denoising error on the distribution of the HSIC is asymptotically negligible for dense functional data.

To develop the asymptotic distribution of $\gamma(P_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}})$, we further define the centered kernel for $\kappa_{\mathcal{X}}$ by $\check{\kappa}_{\mathcal{X}}(X, X') = \langle \kappa_{\mathcal{X}}(X,\cdot) - \mathbf{P}^{\kappa_{\mathcal{X}}}(P_X), \kappa_{\mathcal{X}}(X',\cdot) - \mathbf{P}^{\kappa_{\mathcal{X}}}(P_X)\rangle_{\mathcal{H}(\kappa_{\mathcal{X}})}$. Furthermore define an integral kernel operator $S_{\check{\kappa}_{\mathcal{X}}} : \mathcal{H}(\kappa_{\mathcal{X}}) \to \mathcal{H}(\kappa_{\mathcal{X}})$ by $S_{\check{\kappa}_{\mathcal{X}}}(g) = \int_{\mathcal{X}} \check{\kappa}_{\mathcal{X}}(x,\cdot) g(x) \mathrm{d}P_X(x)$ for any $g \in \mathcal{H}(\kappa_{\mathcal{X}})$. An integral kernel operator $S_{\check{\kappa}_{\mathcal{Y}}}$ for Y can be similarly defined.

Theorem 3. *Under the same assumptions of Theorem 2, if m satisfies*

$$m^{-(\alpha_Z - \beta_Z)/(2\alpha_Z + 2\varsigma_Z + 1)} = o(n^{-1}), \tag{4}$$

for both $Z = X$ and $Z = Y$, then

$$n\gamma(P_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}}) \rightsquigarrow \begin{cases} \sum_{u=1}^{\infty}\sum_{v=1}^{\infty} \mu_u \nu_v N_{uv}^2, & \text{if } X \text{ and } Y \text{ are independent,} \\ \infty, & \text{otherwise,} \end{cases}$$

where "$\rightsquigarrow$" represents weak convergence, $N_{uv} \sim N(0,1), u, v \geq 1$ are i.i.d. and $\{\mu_u : u \geq 1\}$ and $\{\nu_v : v \geq 1\}$ are eigenvalues of $S_{\check{\kappa}_{\mathcal{X}}}$ and $S_{\check{\kappa}_{\mathcal{Y}}}$ respectively.

The proof of Theorem 3 is given in Section S2.3 in the supplementary material. The

asymptotic distribution of $\gamma(P_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}})$ in Theorem 3 is the same as that for fully observed $\{(X_i, Y_i) : X_i \in B^{\beta_X}, Y_i \in B^{\beta_Y}, i = 1, \ldots, n\}$ (Sejdinovic et al., 2013). The requirement (4) ensures that the error due to the denoising procedure is asymptotically negligible under b^{β_Z} norm if the measurements are sufficiently dense. In general, for fixed α_Z, β_Z and ς_Z, the order of m should be higher than $n^{1/r}$ where $r = (\alpha_Z - \beta_Z)/(2\alpha_Z + 2\varsigma_Z + 1)$ which, for example, is $n^{10/3}$ if $(\alpha_Z, \beta_Z, \varsigma_Z) = (2, 1/2, 0)$ and n^4 if $(\alpha_Z, \beta_Z, \varsigma_Z) = (3, 1, 1/2)$.

Since the asymptotic reference distribution of $\gamma(P_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}})$ when X and Y are assumed independent involves many unknown quantities, in practice we perform the test by permutation. As shown in Theorem 4, the permutation test can control the Type I error probability and is also consistent.

Theorem 4 (Permutation Test)**.** *Let the level of significance be $\alpha \in (0, 1)$. If the null hypothesis that X and Y are independent is true, the permutation test of $\gamma(P_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}})$ based on a finite number of permutations rejects the null hypothesis with probability at most α. If the alternative hypothesis that X and Y are dependent is true and the assumptions of Theorem 2 and (4) hold, the permutation test of $\gamma(P_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}})$ based on $B \geq 1/\alpha - 1$ permutations is consistent, i.e., $P(\hat{p}_{\hat{X}\hat{Y}} \leq \alpha) \to 1$ as $n \to \infty$, where $\hat{p}_{\hat{X}\hat{Y}}$ is the p-value.*

The proof of Theorem 4 is given in Section S2.4 in the supplementary material. Theorem 4 shows that the proposed permutation test is also theoretically compatible with the proposed wavelet thresholding method in Step 1.

5 Tuning Parameter Selection

In this section, we discuss the selection of tuning parameters involved in the two-step procedure proposed in Section 3. They include ζ_Z, τ_Z, ς_Z and δ_Z in Step 1 and β_Z in Step 2, where $Z = X$ or Y.

First, to guarantee $\zeta_Z > 1$ and $\tau_Z > e$, we suggest $\zeta_Z = 1.0001$ and $\tau_Z = 1.0001e$ which

are slightly larger than their respective lower bounds, unless domain knowledge is available.

Second, for ς_Z which captures noise amplification and δ_Z which reflects the noise level, we adopt crude estimates for them based on the top two levels of the wavelet coefficients (Johnstone and Silverman, 1997). Explicitly, we obtain $\hat{\varsigma}_Z = \log_2(\hat{\delta}_{Z,J}/\hat{\delta}_{Z,J-1})$ and $\hat{\delta}_Z = \hat{\delta}_{Z,J}/2^{\hat{\varsigma}_Z J}$, where $\hat{\delta}_{Z,j} = \text{median}\left\{\sqrt{m}\theta_{j,k}^{\tilde{Z}_i} : k = 0, \ldots, 2^j - 1\right\}/\text{median}(|W|)$ for $j = J-1, J$, and W is a standard normal random variable.

Finally, for the smoothness parameter β_Z, we will first discuss its role in dependency detection and then propose a data-adaptive selection method for it.

In Section 4, Theorem 2 seems to imply that given α_X and α_Y, the best choice is $\beta_X = \beta_Y = 0$ because the corresponding denoising error attains the best rate of convergence. However, this choice of β_X and β_Y may result in a poor dependency detection especially when the dependency of X and Y originates from their high frequency bands.

For illustration, by Definition 1 and (2), we consider the first-order approximation (Chakraborty and Zhang, 2019, Theorem 5.1)

$$\gamma(P_{XY}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}}) \approx c_{XY}^{-1} \sum_{j_X \geq -1} \sum_{j_Y \geq -1} \gamma\Big(P_{XY}, 2^{2\beta_X j_X} \kappa_{\mathcal{X}}^{(j_X)}, 2^{2\beta_Y j_Y} \kappa_{\mathcal{Y}}^{(j_Y)}\Big), \tag{5}$$

where $\kappa_{\mathcal{Z}}^{(j_Z)}(z, z') = \|\boldsymbol{\theta}_j^z\|_2^2 + \|\boldsymbol{\theta}_j^{z'}\|_2^2 - \|\boldsymbol{\theta}_j^z - \boldsymbol{\theta}_j^{z'}\|_2^2$ for $j_Z \geq -1$, with $(z, Z, \mathcal{Z}) = (x, X, \mathcal{X})$ or $(y, Y, \mathcal{Y})$ and Euclidean norm $\|\cdot\|_2$, and $c_{XY} = 4\sqrt{\mathrm{E}\|X - X'\|_{b^{\alpha_X}}^2 \mathrm{E}\|Y - Y'\|_{b^{\alpha_Y}}^2}$ with X' and Y' being the independent copies of X and Y, respectively. Apparently $\gamma\Big(P_{XY}, 2^{2\beta_X j_X} \kappa_{\mathcal{X}}^{(j_X)}, 2^{2\beta_Y j_Y} \kappa_{\mathcal{Y}}^{(j_Y)}\Big)$ measures the dependency contribution to the HSIC at j_X and j_Y of X and Y respectively, which is zero if and only if X and Y are independent at j_X and j_Y. If $\beta_X = \beta_Y = 0$, the scaling factors $2^{\beta_X j_X} = 2^{\beta_Y j_Y} = 1$ for all $j_X \geq -1, j_Y \geq -1$ and it will be very difficult to detect the dependency between X and Y at high frequencies since the dependency contributions contained at high frequencies are very likely to be overwhelmed by the independent signals at low frequencies. Therefore, we aim to select β_X and β_Y such that the dependency contributions at high frequencies, if any, are detectable.

The idea of the proposed tuning method is to balance the dependency contributions to

HSIC at all frequency scales such that they are approximately the same. To lessen the computational burden, a marginal selection algorithm is proposed in the sense that the optimal β_X is selected only based on X without reliance on Y. Note that, by Appendix A in Sejdinovic et al. (2013) and the properties of distance covariance, the dependency contribution at each $j_X, j_Y \geq -1$ satisfies

$$\gamma\Big(P_{XY}, 2^{2\beta_X j_X}\kappa_{\mathcal{X}}^{(j_X)}, 2^{2\beta_Y j_Y}\kappa_{\mathcal{Y}}^{(j_Y)}\Big) \leq 2^{2(\beta_X j_X+\beta_Y j_Y)}\sqrt{\gamma(P_X, \kappa_{\mathcal{X}}^{(j_X)})\gamma(P_Y, \kappa_{\mathcal{Y}}^{(j_Y)})},$$

where $\gamma(P_Z, \kappa_{\mathcal{Z}}^{(j_Z)}) = \|\mathbf{P}^{\kappa_{\mathcal{Z}}^{(j_Z)}\otimes\kappa_{\mathcal{Z}}^{(j_Z)}}(P_{ZZ})(*,\cdot) - \mathbf{P}^{\kappa_{\mathcal{Z}}^{(j_Z)}}(P_Z)(*)\mathbf{P}^{\kappa_{\mathcal{Z}}^{(j_Z)}}(P_Z)(\cdot)\|^2_{\mathcal{H}(\kappa_{\mathcal{Z}}^{(j_Z)}\otimes\kappa_{\mathcal{Z}}^{(j_Z)})}$, $j_Z \geq -1$, is essentially a distance variance (Székely et al., 2007) with $(z, Z, \mathcal{Z}) = (x, X, \mathcal{X})$ or $(y, Y, \mathcal{Y})$ (Sejdinovic et al., 2013). Thus we propose to select β_X by balancing $2^{2\beta_X j_X}\sqrt{\gamma(P_X, \kappa_{\mathcal{X}}^{(j_X)})}$ at all $j_X \geq -1$. If $2^{2\beta_X j_X}\sqrt{\gamma(P_X, \kappa_{\mathcal{X}}^{(j_X)})} \approx C$ where $C > 0$ is a constant, then $2\beta_X j_X + \frac{1}{2}\log_2\gamma(P_X, \kappa_{\mathcal{X}}^{(j_X)}) \approx \log_2 C$, so β_X may be selected as the estimated slope of the linear regression on $(-2j_X, \log_2\gamma(P_X, \kappa_{\mathcal{X}}^{(j_X)})/2)$.

In practice, we could estimate $\gamma(P_X, \kappa_{\mathcal{X}}^{(j_X)})$ by $\gamma(P_{n,\hat{X}}, \kappa_{\mathcal{X}}^{(j_X)})$ for each $j_X \geq -1$, but its accuracy is poor for very high frequencies due to noise contamination. Thus we only consider j_X up to $\bar{j}_X = \max_{j_X \geq L_X}\{\gamma(P_{n,\hat{X}}, \kappa_{\mathcal{X}}^{(j_X)}) \geq \gamma(P_{n,\hat{e}^X}, \kappa_{\mathcal{X}}^{(j_X)})\}$ where $\hat{e}^X = \tilde{X} - \hat{X}$ is the residual, such that the distance variances of all $j_X \leq \bar{j}_X$ are not smaller than that of the residual. If a known frequency band is of interest in the context of a study, e.g., the alpha band of brain signals, one may alternatively select β_X by balancing $2^{2\beta_X j_X}\sqrt{\gamma(P_X, \kappa_{\mathcal{X}}^{(j_X)})}$ over that frequency band. Last, we remark that the computational benefit of the proposed marginal approach for tuning parameter selection is substantial when many tests have to be performed, such as in the functional connectivity analysis (Section 7).

6 Simulation

In this section we evaluate the numerical performance of our proposed wavelet-based HSIC method wavHSIC in both controlling the Type I error probability and statistical power. We

also compare it with a few representative existing methods, including

(a) Pearson Correlation (Pearson). It is a one-sample t-test based on Fisher-Z transformed correlation coefficients of all subjects. The correlation coefficient for each subject is obtained by applying the Pearson correlation formula to the bivariate time series of the subject, without adjusting for any possible dependence within the time series. It is a popular functional connectivity measure in neuroscience (e.g., He et al., 2012).

(b) Dynamical Correlation (dnm, Dubin and Müller, 2005). It is defined as the expectation of the cosine of the L^2 angle between the standardized versions of two random functions.

(c) Global Temporal Correlation (gtemp, Zhou et al., 2018). It is the integral of the Pearson correlation obtained at each time point.

(d) Bias-Corrected Distance Covariance (dCov-c, Székely and Rizzo, 2013). It is a t-test designed to correct the bias of distance covariance for high-dimensional multivariate data. We apply it by treating the discrete measurements of two random functions as multivariate data. If the bias is not corrected, it is equivalent to wavHSIC with $\beta_X = \beta_Y = 0$.

(e) Functional Principle Component Analysis (FPCA) Based Distance Covariance (FPCA, Kosorok, 2009). The distance covariance (Székely et al., 2007) is applied to top Functional Principle Component (FPC) scores which cumulatively account for 95% of the variation of each random function. When all FPC scores are used, it is equivalent to wavHSIC when $\beta_X = \beta_Y = 0$.

(f) Functional Linearity Test (KMSZ, Kokoszka et al., 2008). It is an approximate chi-squared test for the nullity of the coefficient function by assuming a functional linear model between the two random functions. The model fitting requires a satisfactory approximation of each random function by its top FPC scores and we select those which cumulatively account for 95% of variation of each random function.

(g) Permutation-Based Functional Linearity Test (KMSZ-p). It is the same as KMSZ except that the p-value is obtained by permutation. Such a modification can be regarded as a finite-sample correction of KMSZ.

(h) Projection-based Mean Independence Test (PSS, Patilea et al., 2016). For a functional response Y and a functional predictor X, PSS aims to test the conditional mean independence of Y given X, i.e., $E(Y \mid X) = E(Y), a.s.$ PSS is a model-free test that does not specify a model for $E(Y \mid X)$. It requires a finite-dimensional projection of X and uses wild bootstrap to find critical values. To implement PSS, we used the R package `fdapss` [1], which is publicly available at `http://webspersoais.usc.es/persoais/cesar.sanchez/`.

(i) Functional Martingale Difference Divergence Based Mean Independence Test (FMDD, Lee et al., 2020). FMDD is also a model-free mean independence test. It measures the conditional mean independence using the metric of functional martingale difference divergence and uses wild bootstrap to find critical values. To implement FMDD, we used the R code publicly available at `https://publish.illinois.edu/xshao/files/2019/06/CodeCMDexample1.txt`.

The first five (a-e) in comparison are model-free methods. KMSZ is one of the most popular model-based methods in the FDA literature, but it can only test for linearity. PSS and FMDD can handle nonlinear effects of the functional predictor, but only on the mean of the functional response, so they can only test a weaker notion of independence. Hereafter, for bivariate random functions (X, Y), PSS$(Y \sim X)$ denotes testing $E(Y \mid X) = E(Y), a.s.$ using PSS. Moreover, PSS(Omnibus) denotes the omnibus test which takes the smaller p-value between those obtained by PSS$(Y \sim X)$ and PSS$(X \sim Y)$ respectively. FMDD$(Y \sim$

[1] The package is only for Windows platform. For the user-chosen parameters required by this package, we followed the recommendation in Section 4.1 of Patilea et al. (2016) and set the bandwidth $h = n^{-2/9}$, penalty coefficient $\alpha = 2$, grid size $n_q = 50$ and number of FPCs which cumulatively account for 95% of the variation of the functional predictor.

X), FMDD($X \sim Y$) and FMDD(Omnibus) are similarly defined. To obtain p-values, 1,999 permutations were used for wavHSIC, dnm, gtemp, FPCA and KMSZ-p while 1,999 bootstrap samples were used for PSS and FMDD. We declare statistical significance in each simulated data based on the level of significance 0.05.

We generated 199 simulated datasets, where the number 199 is chosen to prevent empirical Type I and Type II error probabilities from coinciding with the level of significance 0.05. In each simulated dataset $n = 50$ or 200 independent subjects with bivariate functions $\{(X_i(t), Y_i(t)) : t \in [0,1], i = 1, \ldots, n\}$ were generated where for the i-th subject, $X_i(t) = \sum_{k=1}^{16} \eta_{ik}\phi_k(t)$ and $Y_i(t) = \sum_{k=1}^{16} \zeta_{ik}\phi_k(t+0.2)$ with $\phi_{2k-1}(t) = \sqrt{2}\cos(2\pi kt), \phi_{2k}(t) = \sqrt{2}\sin(2\pi kt)$ for $k = 1, \ldots, 8$. We considered three settings with different dependency structures of the bivariate functional data which are controlled by the FPC scores $\{(\eta_{ik}, \zeta_{ik}) : k = 1, \ldots, 16; i = 1, \ldots, n\}$.

- Setting 1. We generated $\eta_{ik} \sim N(0, k^{-1.05}), k = 1, \ldots, 16$ and $\zeta_{ik} \sim N(0, k^{-1.2}), k = 1, \ldots, 16$ independently.
- Setting 2. With $\rho = 0$ for $k = 1, \ldots, 8$ and $\rho = 0.6$ for $k = 9, \ldots, 16$, we generated
$$\begin{bmatrix} \eta_{ik} \\ \zeta_{ik} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} k^{-1.05} & \rho k^{-1.125} \\ \rho k^{-1.125} & k^{-1.2} \end{bmatrix} \right).$$
- Setting 3. For $k = 1, \ldots, 8$, $\eta_{ik} \sim N(0, k^{-1.05})$ was generated independently of $\zeta_{ik} \sim N(0, k^{-1.2})$. For $k = 9, \ldots, 16$, $\eta_{ik} \sim N(0, k^{-1.05})$ and $\zeta_{ik} = \eta_{ik}^2 - \mathrm{E}\eta_{ik}^2$.

Apparently X and Y are independent in Setting 1 and dependent in Settings 2 and 3. In Setting 2, the FPC scores of X and Y are linearly correlated but only at high spectral frequencies, while in Setting 3 they are linearly uncorrelated but dependent only at high spectral frequencies, so it is more difficult to detect dependency for all methods in Setting 3 than Setting 2.

Both functions are measured at $m = 64$ or 256 equidistant points on the time domain $[0, 1]$. We added Gaussian noise to all measurements with signal-to-noise ratio SNR = 4

or 8, which is the variance of all measurements over the noise variance. The noise terms were generated independently across subjects. Within each subject, we experimented with both independent (white noise) and dependent (correlated noise) settings. For the dependent setting, the Gaussian noise was generated by differencing the fractional Brownian motion with Hurst exponent 0.7.

Since all methods in comparison require noiseless functions, we used the same denoising procedure in Step 1 for all of them for fairness. We chose the CDJV wavelet basis functions with vanishing moment $D = 10$ for both X and Y, which leads to $\alpha_X = \alpha_Y \approx 2.902$ (Daubechies, 1992). The tuning parameters β_X and β_Y were selected by the method in Section 5. The results are given in Tables 1–6.

Tables 1 and 4 show that all methods are almost always able to control type I error probabilities except for PSS($X \sim Y$) and the two omnibus tests when the two random functions are truly independent. Relatively, KMSZ is very conservative in many cases and KMSZ-p corrects its p-values to some extent. However, KMSZ-p seems more likely to detect spurious dependency when $(n, m) = (50, 64)$, so does dCov-c when $(n, m) = (200, 256)$.

Tables 2, 3, 5 and 6 show that the statistical powers of all methods typically improve when one of n, m and SNR increases under Setting 2, but unnecessarily under Setting 3 except for KMSZ, KMSZ-p and wavHSIC. This demonstrates the difficulty of Setting 3 in detecting dependency to some extent. Except wavHSIC, all model-free methods have very low powers in all scenarios under either Setting 2 or 3, which indicates their poor performances in detecting linear dependency in high frequencies or nonlinear dependency. The performance of KMSZ is satisfactory for $n = 200$ under Setting 2 when the relationship between X and Y is truly linear. KMSZ-p improves the statistical power of KMSZ further for $n = 50$ under Setting 2 by permutation. However, both KMSZ and KMSZ-p are poor at testing nonlinear dependency in Setting 3. The performances of PSS and FMDD, which can detect nonlinear mean dependency, are comparable with those of dCov-c and FPCA in Settings 2 and 3, but worse than those of KMSZ and KMSZ-p in Setting 2 where X and Y are linearly dependent.

Tables 2, 3, 5 and 6 also demonstrate the appealing performance of wavHSIC. It is always the most powerful method, and substantially better than the other methods. Only the powers of KMSZ and KMSZ-p are comparable with those of wavHSIC when the sample size $n = 200$ is large and the linearity assumption is valid under Setting 2. For fixed (n, m, SNR), the medians of the selected parameters β_X and β_Y for wavHSIC are always similar between Settings 2 and 3 since they were tuned marginally regardless of the dependency structure. On average, both β_X and β_Y were considerably away from zero, which confirms the need and benefit of choosing them properly to enhance the detection sensitivity of wavHSIC.

We also performed an additional simulation study described in Section S3.2 of the supplementary material, which follows the same settings in Section 1.2 of the supplementary material of Lee et al. (2020). The results also demonstrate the superiority of wavHSIC.

Remark 3. *It is worth noting that the development of the asymptotic distribution of wavHSIC as in Theorem 3 requires the number of measurements per curve m to be large compared to the sample size n (see (4)), but the simulation results here show that the finite sample performance of wavHSIC is still satisfactory, even when m is small relatively to n. However, this is not entirely surprising. First, under the null hypothesis that X and Y are independent, a poor pre-smoothing due to a relatively small m does not inflate the empirical Type I error probability since the remaining noise does not enhance the dependency between X and Y and the critical value is obtained by permutation. Second, under the alternative hypothesis that X and Y are dependent, as long as m is sufficiently large such that the dependency signals can captured by the wavelet coefficients, wavHSIC can still detect dependency, but its power may be worse if (4) is not satisfied.*

Table 1: Empirical Type I error probabilities for all methods under Setting 1 with white noise. The last two rows provide the medians of the selected β_X and β_Y for wavHSIC.

Setting 1 with	$n = 50$				$n = 200$			
white noise	$m = 64$		$m = 256$		$m = 64$		$m = 256$	
Type I error rate	SNR=4	SNR=8	SNR=4	SNR=8	SNR=4	SNR=8	SNR=4	SNR=8
Pearson	0.0452	0.0352	0.0503	0.0452	0.0704	0.0704	0.0553	0.0553
dnm	0.0452	0.0452	0.0653	0.0503	0.0553	0.0553	0.0653	0.0603
gtemp	0.0503	0.0603	0.0653	0.0603	0.0402	0.0352	0.0352	0.0352
dCov-c	0.0653	0.0603	0.0603	0.0603	0.0503	0.0603	0.0704	0.0754
FPCA	0.0503	0.0452	0.0503	0.0452	0.0452	0.0503	0.0452	0.0452
KMSZ	0.0201	0.0101	0.0101	0.0151	0.0201	0.0151	0.0402	0.0251
KMSZ-p	0.0804	0.0905	0.0553	0.0402	0.0302	0.0352	0.0402	0.0352
PSS($X \sim Y$)	0.0804	0.0754	0.1005	0.0804	0.0653	0.0905	0.0402	0.0955
PSS($Y \sim X$)	0.0402	0.0754	0.0704	0.0452	0.0302	0.0754	0.0352	0.0503
PSS(Omnibus)	0.1156	0.1307	0.1558	0.1106	0.0955	0.1407	0.0754	0.1407
FMDD($X \sim Y$)	0.0553	0.0552	0.0704	0.0653	0.0503	0.0603	0.0553	0.0603
FMDD($Y \sim X$)	0.0553	0.0552	0.0553	0.0553	0.0503	0.0603	0.0503	0.0503
FMDD(Omnibus)	0.0653	0.0603	0.0704	0.0704	0.0603	0.0704	0.0704	0.0704
wavHSIC	0.0452	0.0352	0.0503	0.0653	0.0302	0.0402	0.0251	0.0251
median$\{\beta_X\}$	0.948	0.989	0.983	0.991	0.959	1.000	0.990	1.001
median$\{\beta_Y\}$	0.671	0.724	0.733	0.741	0.696	0.745	0.747	0.761

Table 2: Empirical powers for all methods under Setting 2 with white noise. The last rows provide the medians of the selected β_X and β_Y for wavHSIC.

Setting 2 with	$n = 50$				$n = 200$			
white noise	$m = 64$		$m = 256$		$m = 64$		$m = 256$	
Power	SNR=4	SNR=8	SNR=4	SNR=8	SNR=4	SNR=8	SNR=4	SNR=8
Pearson	0.0854	0.0804	0.0804	0.0804	0.1357	0.1357	0.1357	0.1206
dnm	0.0704	0.0653	0.0704	0.0704	0.1608	0.1558	0.1508	0.1457
gtemp	0.0653	0.0653	0.0804	0.0754	0.0905	0.0854	0.0754	0.0754
dCov-c	0.1106	0.1055	0.0905	0.0804	0.2362	0.2462	0.2714	0.2663
FPCA	0.0854	0.0804	0.0804	0.0804	0.1709	0.1709	0.1859	0.1809
KMSZ	0.4221	0.4925	0.5025	0.5075	1.0000	1.0000	1.0000	1.0000
KMSZ-p	0.7035	0.7889	0.7688	0.7990	1.0000	1.0000	1.0000	1.0000
PSS($X \sim Y$)	0.0955	0.1055	0.0804	0.0905	0.1256	0.0955	0.1106	0.0804
PSS($Y \sim X$)	0.0653	0.0653	0.0553	0.0653	0.0804	0.0653	0.0503	0.0503
PSS(Omnibus)	0.1558	0.1658	0.1307	0.1508	0.2060	0.1608	0.1558	0.1206
FMDD($X \sim Y$)	0.0854	0.0905	0.0905	0.0854	0.1859	0.1960	0.2915	0.2814
FMDD($Y \sim X$)	0.0754	0.0804	0.0704	0.0653	0.1407	0.1759	0.2161	0.2211
FMDD(Omnibus)	0.0955	0.1005	0.0905	0.0854	0.1960	0.2211	0.3116	0.3116
wavHSIC	0.9548	0.9849	0.9849	0.9899	1.0000	1.0000	1.0000	1.0000
median$\{\beta_X\}$	0.942	0.987	0.975	0.983	0.955	0.996	0.994	1.001
median$\{\beta_Y\}$	0.674	0.720	0.741	0.752	0.693	0.739	0.742	0.762

Table 3: Empirical powers for all methods under Setting 3 with white noise. The last two rows provide the medians of the selected β_X and β_Y for wavHSIC.

Setting 3 with	$n = 50$				$n = 200$			
white noise	$m = 64$		$m = 256$		$m = 64$		$m = 256$	
Power	SNR=4	SNR=8	SNR=4	SNR=8	SNR=4	SNR=8	SNR=4	SNR=8
Pearson	0.0452	0.0402	0.0603	0.0603	0.0503	0.0503	0.0402	0.0503
dnm	0.0804	0.0804	0.0754	0.0704	0.0704	0.0603	0.0754	0.0754
gtemp	0.0754	0.0804	0.0754	0.0704	0.0704	0.0653	0.0754	0.0704
dCov-c	0.0955	0.1055	0.1005	0.1005	0.0854	0.0905	0.0854	0.0854
FPCA	0.0704	0.0854	0.0955	0.1005	0.0704	0.0704	0.0653	0.0704
KMSZ	0.0101	0.0101	0.0201	0.0251	0.1206	0.1307	0.1206	0.1357
KMSZ-p	0.1106	0.0854	0.1307	0.1357	0.1558	0.1608	0.1407	0.1709
PSS($X \sim Y$)	0.0754	0.0854	0.0905	0.1055	0.1005	0.0452	0.0553	0.0653
PSS($Y \sim X$)	0.0653	0.0553	0.0754	0.0804	0.0603	0.0503	0.0603	0.0704
PSS(Omnibus)	0.1357	0.1307	0.1508	0.1658	0.1558	0.0905	0.1106	0.1256
FMDD($X \sim Y$)	0.0804	0.0804	0.0804	0.0804	0.0704	0.0704	0.0754	0.0704
FMDD($Y \sim X$)	0.0955	0.1005	0.0955	0.1005	0.0804	0.0754	0.0754	0.0754
FMDD(Omnibus)	0.1005	0.1005	0.0955	0.1005	0.0854	0.0804	0.0804	0.0804
wavHSIC	0.2613	0.3618	0.3367	0.407	0.804	0.9347	0.9347	0.9749
median$\{\beta_X\}$	0.948	0.993	0.968	0.979	0.949	0.993	0.981	0.989
median$\{\beta_Y\}$	0.724	0.771	0.773	0.790	0.723	0.769	0.775	0.785

Table 4: Empirical Type I error probabilities for all methods under Setting 1 with correlated noise. The last two rows provide the medians of the selected β_X and β_Y for wavHSIC.

Setting 1 with	$n = 50$				$n = 200$			
correlated noise	$m = 64$		$m = 256$		$m = 64$		$m = 256$	
Type I error rate	SNR=4	SNR=8	SNR=4	SNR=8	SNR=4	SNR=8	SNR=4	SNR=8
Pearson	0.0352	0.0352	0.0452	0.0452	0.0704	0.0553	0.0553	0.0553
dnm	0.0402	0.0452	0.0603	0.0603	0.0553	0.0503	0.0503	0.0503
gtemp	0.0603	0.0553	0.0553	0.0653	0.0553	0.0503	0.0452	0.0553
dCov-c	0.0553	0.0653	0.0603	0.0603	0.0553	0.0603	0.0754	0.0754
FPCA	0.0452	0.0452	0.0452	0.0402	0.0503	0.0452	0.0452	0.0503
KMSZ	0.0151	0.0000	0.0151	0.0151	0.0201	0.0251	0.0251	0.0251
KMSZ-p	0.0804	0.0754	0.0553	0.0452	0.0302	0.0352	0.0352	0.0402
PSS($X \sim Y$)	0.0452	0.0553	0.0603	0.0955	0.0653	0.0452	0.0452	0.0603
PSS($Y \sim X$)	0.0553	0.0704	0.0553	0.0754	0.0452	0.0302	0.0603	0.0452
PSS(Omnibus)	0.1005	0.1156	0.1106	0.1558	0.1005	0.0704	0.1005	0.1005
FMDD($X \sim Y$)	0.0553	0.0603	0.0653	0.0653	0.0653	0.0603	0.0553	0.0603
FMDD($Y \sim X$)	0.0452	0.0452	0.0503	0.0503	0.0553	0.0603	0.0603	0.0603
FMDD(Omnibus)	0.0553	0.0653	0.0704	0.0704	0.0754	0.0804	0.0754	0.0754
wavHSIC	0.0402	0.0402	0.0553	0.0653	0.0352	0.0302	0.0352	0.0302
median$\{\beta_X\}$	1.014	1.020	0.996	0.997	1.024	1.030	1.006	1.008
median$\{\beta_Y\}$	0.752	0.765	0.750	0.754	0.774	0.783	0.770	0.772

Table 5: Empirical powers for all methods under Setting 2 with correlated noise. The last two rows provide the medians of the selected β_X and β_Y for wavHSIC.

Setting 2 with	$n = 50$				$n = 200$			
correlated noise	$m = 64$		$m = 256$		$m = 64$		$m = 256$	
Power	SNR=4	SNR=8	SNR=4	SNR=8	SNR=4	SNR=8	SNR=4	SNR=8
Pearson	0.0854	0.0854	0.0804	0.0804	0.1508	0.1407	0.1256	0.1307
dnm	0.0653	0.0653	0.0754	0.0704	0.1558	0.1608	0.1558	0.1457
gtemp	0.0653	0.0653	0.0905	0.0905	0.0854	0.0804	0.0754	0.0854
dCov-c	0.1005	0.0955	0.0854	0.0854	0.2663	0.2714	0.2714	0.2814
FPCA	0.0854	0.0804	0.0754	0.0804	0.1658	0.1759	0.1809	0.1809
KMSZ	0.5427	0.5628	0.5126	0.5327	1.0000	1.0000	1.0000	1.0000
KMSZ-p	0.8241	0.8141	0.8191	0.8291	1.0000	1.0000	1.0000	1.0000
PSS($X \sim Y$)	0.0955	0.0653	0.0905	0.1005	0.1156	0.1106	0.1005	0.1156
PSS($Y \sim X$)	0.0553	0.0653	0.0704	0.0603	0.0704	0.0653	0.0603	0.0754
PSS(Omnibus)	0.1407	0.1156	0.1558	0.1508	0.1809	0.1457	0.1508	0.1809
FMDD($X \sim Y$)	0.1055	0.0905	0.0854	0.0854	0.2412	0.2513	0.2714	0.2714
FMDD($Y \sim X$)	0.0804	0.0905	0.0704	0.0704	0.2111	0.2111	0.2412	0.2412
FMDD(Omnibus)	0.1106	0.1055	0.0905	0.0905	0.2613	0.2714	0.3166	0.3015
wavHSIC	0.9950	0.9950	0.9899	0.9899	1.0000	1.0000	1.0000	1.0000
median$\{\beta_X\}$	1.011	1.022	0.990	0.995	1.023	1.033	1.011	1.014
median$\{\beta_Y\}$	0.749	0.765	0.757	0.760	0.769	0.781	0.764	0.766

Table 6: Empirical powers for all methods under Setting 3 with correlated noise. The last two rows provide the medians of the selected β_X and β_Y for wavHSIC.

Setting 3 with	$n = 50$				$n = 200$			
correlated noise	$m = 64$		$m = 256$		$m = 64$		$m = 256$	
Power	SNR=4	SNR=8	SNR=4	SNR=8	SNR=4	SNR=8	SNR=4	SNR=8
Pearson	0.0402	0.0402	0.0603	0.0603	0.0503	0.0503	0.0452	0.0553
dnm	0.0704	0.0653	0.0704	0.0704	0.0653	0.0653	0.0754	0.0754
gtemp	0.0854	0.0804	0.0704	0.0603	0.0553	0.0653	0.0704	0.0804
dCov-c	0.1055	0.1106	0.1005	0.1005	0.0905	0.0804	0.0804	0.0854
FPCA	0.0854	0.0905	0.1005	0.1005	0.0704	0.0653	0.0754	0.0704
KMSZ	0.0151	0.0101	0.0302	0.0352	0.1256	0.1357	0.1357	0.1407
KMSZ-p	0.0905	0.0854	0.1256	0.1256	0.1709	0.1759	0.1809	0.1960
PSS($X \sim Y$)	0.0905	0.0905	0.0754	0.0754	0.0653	0.0804	0.0503	0.0653
PSS($Y \sim X$)	0.0704	0.0653	0.0754	0.0854	0.0503	0.0603	0.0503	0.0402
PSS(Omnibus)	0.1558	0.1558	0.1457	0.1457	0.1156	0.1407	0.1005	0.0955
FMDD($X \sim Y$)	0.0804	0.0804	0.0905	0.0905	0.0653	0.0754	0.0704	0.0754
FMDD($Y \sim X$)	0.1005	0.1005	0.1005	0.0955	0.0754	0.0754	0.0754	0.0754
FMDD(Omnibus)	0.1106	0.1055	0.1005	0.1055	0.0804	0.0854	0.0754	0.0854
wavHSIC	0.4221	0.4472	0.4422	0.4422	0.9849	0.9849	0.9899	0.9899
median$\{\beta_X\}$	1.022	1.030	0.986	0.987	1.019	1.029	0.996	0.998
median$\{\beta_Y\}$	0.799	0.810	0.800	0.802	0.801	0.807	0.795	0.798

7 Real Data Application

We applied our proposed method to study human brain functional connectivity using the MEG dataset collected by the HCP. MEG measures magnetic fields generated by human neuronal activities with a high temporal resolution. Before source reconstruction, the signals from all MEG sensors outside head were preprocessed following the HCP MEG pipeline reference (`www.humanconnectome.org/software/hcp-meg-pipelines`) and the preprocessed data are publicly accessible from the HCP website. To obtain the electric activity signals from cortex regions, we applied the source reconstruction procedure of MEG signals to the cerebral cortex atlas provided by Glasser et al. (2016) using the linearly constrained minimum variance beamforming method in the MATLAB package `FieldTrip`.

To study the functional dependency between cortex regions under some motor activities, we focused on motor task trials where subjects moved their right hands. There were $n = 61$ subjects in the trials. For each subject, $8,004$ signal curves were obtained by denoising and source reconstruction procedures with around 75 repeated trials. Within each trial, the signals were recorded about every 2 ms from -1.2 to 1.2 seconds, where the time 0 is the starting time of the motion. Since the motion in each trial usually lasts no longer than about 0.75 seconds and typically a subject finished the previous movement and received a new cue between times -0.25 and 0 of the next trial, we considered the time domain $[-0.2521, 0.7525]$ which covers the time period of interest, with $m = 512$ sampled time points in total.

We applied the proposed method wavHSIC to perform an independence test for every pair of the MEG signals. To implement wavHSIC, we chose the CDJV wavelet basis functions with vanishing moment $D = 4$ which leads to $\alpha \approx 1.6179$. For each signal, the smoothness parameter β was selected by the method in Section 5. For comparison, we also provided the results for the model-based test KMSZ, KMSZ-p and two model-free tests, Pearson and FPCA. KMSZ, KMSZ-p and FPCA were based on top FPC scores which cumulatively account for 95% of the variation of each signal. The p-value for testing the independence between each pair of

signals were obtained by 1,999 permutations for wavHSIC, FPCA and KMSZ-p. We did not include PSS and FMDD here due to their extended computing times. See Table 7 below for an illustration.

The empirical cumulative distribution functions for the p-values of the five methods are given in Figure 1, which shows that wavHSIC is more sensitive to detecting connectivity than the other methods. To evaluate and compare the five methods at the presence of multiple testing, we set the same discovery rate at 60% to control the number of edges, or sparsity, of each brain connectivity network, which is important in evaluating the reliability of brain network metrics (e.g. Van Wijk et al., 2010; Tsai, 2018). In this analysis, we focus on sensorimotor areas 4, 3a, 3b, 1 and 2 on the left and right hemispheres as illustrated in Figure 3 (c) which are most related to motor task trials (Glasser et al., 2016). With a controlled discovery rate, we expect an excellent connectivity detection method to identify plenty of edges within these areas.

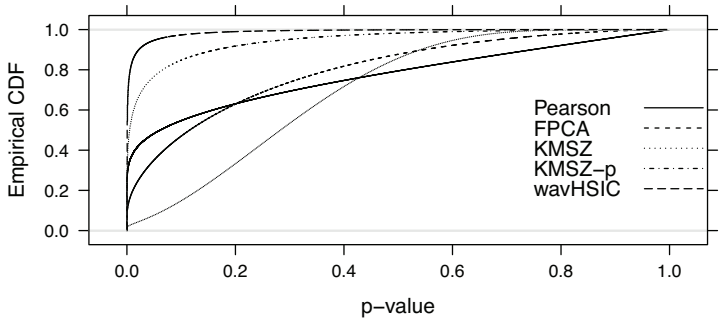


Figure 1: Empirical cumulative distribution function for the p-values for testing the independence between every pair of the 8,004 signals for each method.

Figures 2 and 3 (a) provide the functional connectivity networks within these sensorimotor areas obtained by the five methods. The nodes in each area were ordered from the superio-medial cortex to infero-lateral cortex following the atlas "atlas_MMP1.0_4k.mat" in `FieldTrip`. Compared with KMSZ, KMSZ-p and wavHSIC, Pearson and FPCA are substantially less sensitive to detecting functional connectivity and their corresponding networks are less structured (see Figure 2 (a) and (b)). This demonstrates the superior performances of

both KMSZ, KMSZ-p and wavHSIC in identifying connectivity patterns within these areas which are anatomically connected and functionally related to the motion task trials. Different from the overall homogeneous pattern in the network for KMSZ, several structured dark strips appear in the network obtained by KMSZ-p and wavHSIC within sensorimotor areas 4, 3a, 3b and 1 in the right hemisphere (see Figures 2 (c–d) and 3 (a)). These dark strips are much clearer in Figure 3 (a) than in Figure 2 (d). This indicates that wavHSIC can more clearly identify two sub-areas in sensorimotor areas 4, 3a, 3b and 1 in the right hemisphere, the top

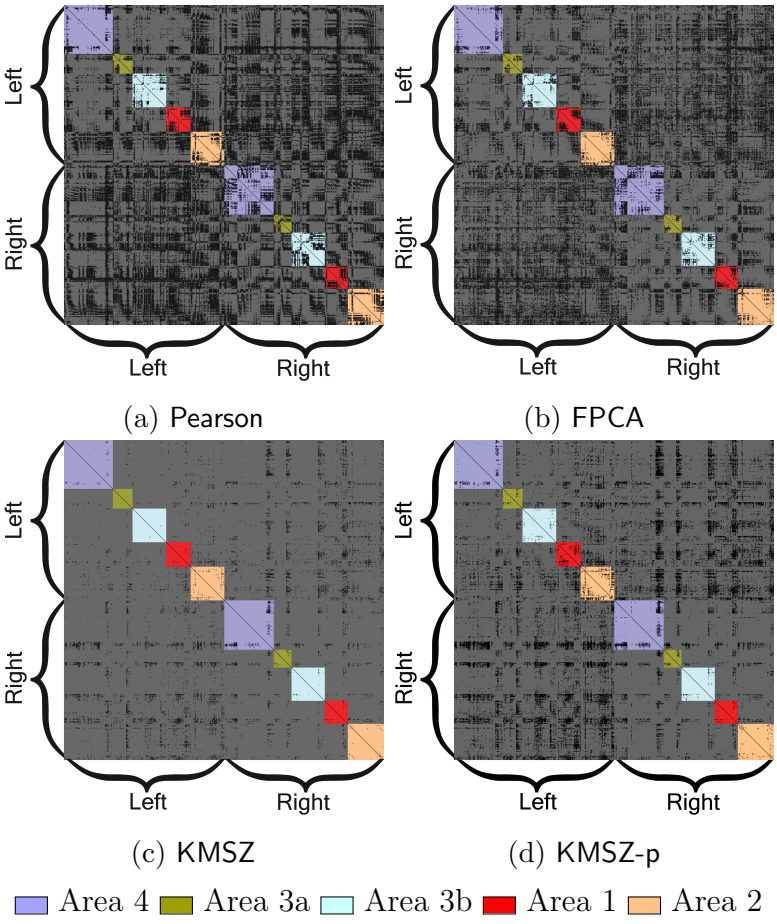
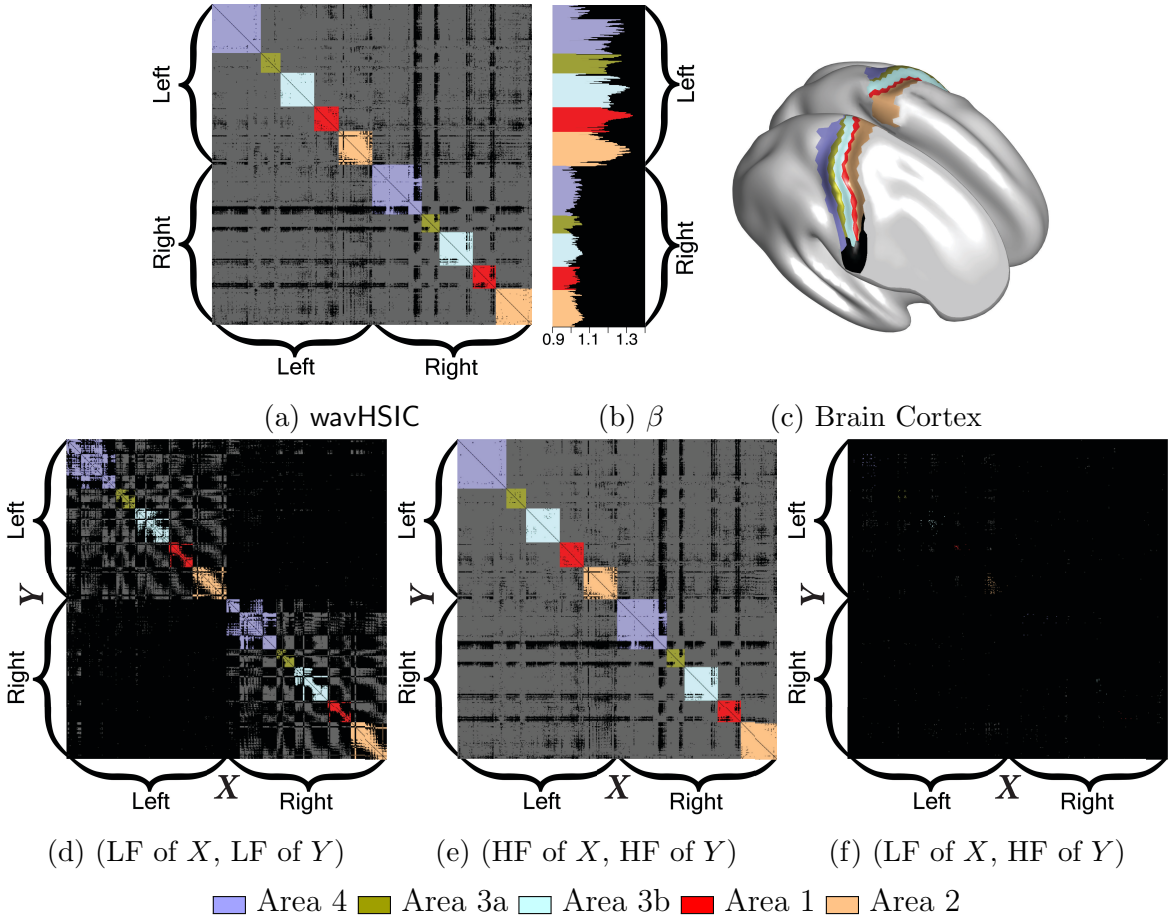


Figure 2: Functional connectivity networks of the five sensorimotor areas in the left and right hemispheres. In the adjacency matrices in (a), (b), (c), (d) obtained by the four methods respectively, a bright entry indicates significant dependency between the corresponding signal pairs while a dark one indicates otherwise.



(a) wavHSIC (b) β (c) Brain Cortex

(d) (LF of X, LF of Y) (e) (HF of X, HF of Y) (f) (LF of X, HF of Y)

Figure 3: Functional connectivity networks of the five sensorimotor areas in the left and right hemispheres with the same color scheme in Figure 2. The adjacency matrix (a) is obtained by wavHSIC with the smoothness parameters β, selected by the method in Section 5, illustrated in the barplot (b). The black subregion in (c) corresponds to face and eye portions and the rest of the colored area corresponds to upper limbs, trunk and lower limbs portions in the right hemisphere. The adjacency matrices in (d), (e), (f) are obtained by applying wavHSIC to low($\leq$ 4Hz)/high($>$ 4Hz)-pass-filtered signals with the same β values in (b) and the same p-value threshold in (a).

left (TL) and bottom right (BR) corners respectively in these corresponding colored squares as in Figure 3 (a). The signals within these four TL sub-areas or within these four BR sub-areas are strongly connected, while the connectivities between these TL and BR sub-areas are generally weak. According to Glasser et al. (2016), the four BR sub-areas in the same

hemisphere correspond to face and eye portions while the four TL sub-areas correspond to upper limbs, trunk and lower limbs portions. Since the motor task involved in this dataset is raising the right hand, the connectivity patterns detected by wavHSIC are intuitively and anatomically interpretable.

Next we illustrate how to identify dependency structures between and within different frequency bands using wavHSIC. Explicitly, we first split the denoised wavelet coefficients of each brain signal into two parts, the low-frequency part (LF, $j \leq 3$) and high-frequency part (HF, $j > 3$), which approximately correspond to the Delta band ($\leq$ 4Hz) and the Theta to the Ultra-Gamma bands ($>$ 4Hz) respectively (e.g., Buzsaki, 2006). Then for each pair of signals (X, Y) as illustrated in Figure 3 (d), (e) and (f), we applied wavHSIC to (LF of X, LF of Y), (HF of X, HF of Y), and (LF of X, HF of Y) respectively. Their corresponding functional connectivity patterns are shown in Figure 3 (d), (e) and (f) respectively. Note that the results for (HF of X, LF of Y) are included in Figure 3 (f) by switching the roles of X and Y. Apparently, the network in Figure 3 (e) is very similar to that in Figure 3 (a), which indicates that the functional dependency induced by this motor task mainly lies at high frequencies. Moreover, Figure 3 (f) shows that there is essentially no dependency between the low-frequency and high-frequency signals. Lastly, Figure 3 (d) reveals that some dependency can be detected at low frequencies, but only within the same hemisphere. This is probably due to the fact that functional Delta oscillations appear to be implicated in the synchronization of brain activity with autonomic functions of vegetative nervous system, but is not affected by a specific task (Knyazev, 2012).

To compare computing times of these methods together with PSS and FMDD, we randomly selected *one* pair of signals and then repeatedly executed each of them 20 times on a Windows 10 desktop with AMD Ryzen7 3800X CPU and 16GB RAM. A summary of their averaged computing times (in seconds) is given in Table 7. The long computational times of PSS and FMDD make it difficult to study dependency between *every* pair and create corresponding functional connectivity networks, so we did not include them in the analysis above.

Table 7: Mean computing times (in seconds) based on *one* randomly selected pair of signals for the seven methods in comparison. The values in parentheses are standard deviations.

Method	Pearson	FPCA	KMSZ	KMSZ-p	wavHSIC	PSS	FMDD
Time	0.003(0.002)	0.103(0.072)	0.020(0.003)	0.158(0.082)	0.162(0.017)	43.076(0.728)	15.280(0.253)

8 Discussion

In this paper, we propose a model-free wavelet-based independence test for two random functions of which sample paths belong to possibly different Besov spaces. Our method is built upon HSIC endowed with characteristic kernels, which is zero if and only if the two random functions are independent. Since the Besov space with wavelet basis functions provides an effective modeling environment for sample paths with various levels of smoothness, HSIC with characteristic kernels induced by wavelet coefficients is capable of capturing the dependency at different frequencies. Therefore, the proposed method is especially powerful when the two random functions are dependent only at high frequencies, as demonstrated in Section 6. If the dependency is strong at low frequencies, our simulation not presented here shows that the proposed method is not substantially advantageous over FPCA.

In the application to MEG functional connectivity, the proposed method by construction is only able to identify the *unconditional* dependency between two signal curves. Although metrics that reflect unconditional functional connectivity are still widely used in neuroscience (see, e.g., Marzetti et al., 2019, for a review), a conditional independence measure or test will be more convincing to identify the functional connectivity between two signal curves given all others in the brain. To address this problem, there have been some advances in functional graphical models. Most of the existing methods reply on either Gaussianity (e.g., Zhu et al., 2016; Qiao et al., 2019, 2020; Zapata et al., 2019; Solea and Li, 2020; Zhao et al., 2021) or regression models (e.g. Lundborg et al., 2021), while a few exceptions assume additive structures (e.g., Li and Solea, 2018; Lee et al., 2021; Solea and Dette, 2021). Developing a

conditional independence test with these assumptions relaxed would be an interesting future research topic.

SUPPLEMENTARY MATERIAL

The supplementary material includes background materials on distance-induced characteristic kernels and Besov spaces, technical proofs of Theorems 1–4 and additional simulations.

References

Abramovich, F. and B. W. Silverman (1998). Wavelet decomposition approaches to statistical inverse problems. *Biometrika 85*(1), 115–129.

Antoniadis, A. and T. Sapatinas (2007). Estimation and inference in functional mixed-effects models. *Computational Statistics & Data Analysis 51*(10), 4793–4813.

Buzsaki, G. (2006). *Rhythms of the Brain*. Oxford University Press.

Chakraborty, S. and X. Zhang (2019). A new framework for distance and kernel-based metrics in high dimensions. *arXiv preprint arXiv:1909.13469*.

Chen, F., Q. Jiang, Z. Feng, and L. Zhu (2020). Model checks for functional linear regression models based on projected empirical processes. *Computational Statistics & Data Analysis 144*, 106897.

Cohen, A., I. Daubechies, B. Jawerth, and P. Vial (1993). Multiresolution analysis, wavelets and fast algorithms on an interval. *Comptes rendus de l'Académie des sciences. Série 1, Mathématique 316*(5), 417–421.

Daubechies, I. (1992). *Ten lectures on wavelets*, Volume 61. SIAM.

DeVore, R. A. and G. G. Lorentz (1993). *Constructive approximation*, Volume 303. Springer-Verlag Berlin Heidelberg.

Donoho, D. L. (1995). Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Applied and Computational Harmonic Analysis 2*(2), 101–126.

Donoho, D. L., I. M. Johnstone, G. Kerkyacharian, and D. Picard (1995). Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society: Series B (Methodological) 57*(2), 301–337.

Dubin, J. A. and H.-G. Müller (2005). Dynamical correlation for multivariate longitudinal data. *Journal of the American Statistical Association 100*(471), 872–881.

Eubank, R. L. and T. Hsing (2008). Canonical correlation for stochastic processes. *Stochastic Processes and their Applications 118*(9), 1634–1661.

Ferraty, F. and P. Vieu (2006). *Nonparametric functional data analysis: theory and practice*. Springer, New York.

Glasser, M. F., T. S. Coalson, E. C. Robinson, C. D. Hacker, J. Harwell, E. Yacoub, K. Ugurbil, J. Andersson, C. F. Beckmann, M. Jenkinson, S. M. Smith, and D. C. Van Essen (2016, Aug). A multi-modal parcellation of human cerebral cortex. *Nature 536*(7615), 171–178.

Gretton, A., O. Bousquet, A. Smola, and B. Schölkopf (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In *International Conference on Algorithmic Learning Theory*, pp. 63–77. Springer.

Gretton, A., K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola (2008). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, pp. 585–592.

Guo, W. (2002). Functional mixed effects models. *Biometrics 58*(1), 121–128.

He, G., H.-G. Müller, and J.-L. Wang (2003). Functional canonical analysis for square integrable stochastic processes. *Journal of Multivariate Analysis 85*(1), 54–77.

He, J., O. Carmichael, E. Fletcher, B. Singh, A.-M. Iosif, O. Martinez, B. Reed, A. Yonelinas, and C. DeCarli (2012). Influence of functional connectivity and structural MRI measures on episodic memory. *Neurobiology of Aging 33*(11), 2612–2620.

Huang, J. Z., C. O. Wu, and L. Zhou (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika 89*(1), 111–128.

Johnstone, I. M. (2019). Gaussian estimation: sequence and wavelet models. `statweb.stanford.edu/~imj/GE_09_16_19.pdf`.

Johnstone, I. M. and D. Paul (2014). Adaptation in some linear inverse problems. *Stat 3*(1), 187–199.

Johnstone, I. M. and B. W. Silverman (1997). Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society: Series B (Methodological) 59*(2), 319–351.

Johnstone, I. M. and B. W. Silverman (2005). Empirical Bayes selection of wavelet thresholds. *Annals of Statistics 33*(4), 1700–1752.

Knyazev, G. G. (2012). EEG delta oscillations as a correlate of basic homeostatic and motivational processes. *Neuroscience & Biobehavioral Reviews 36*(1), 677–695.

Kokoszka, P., I. Maslova, J. Sojka, and L. Zhu (2008). Testing for lack of dependence in the functional linear model. *Canadian Journal of Statistics 36*(2), 207–222.

Kosorok, M. R. (2009). Discussion of: Brownian distance covariance. *Annals of Applied Statistics 3*(4), 1270–1278.

Kovac, A. and B. W. Silverman (2000). Extending the scope of wavelet regression methods by coefficient-dependent thresholding. *Journal of the American Statistical Association 95*(449), 172–183.

Lee, C., X. Zhang, and X. Shao (2020). Testing conditional mean independence for functional data. *Biometrika 107*(2), 331–346.

Lee, K.-Y., D. Ji, L. Li, T. Constable, and H. Zhao (2021). Conditional functional graphical models. *Journal of the American Statistical Association*, in press.

Leurgans, S. E., R. A. Moyeed, and B. W. Silverman (1993). Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological) 55*(3), 725–740.

Li, B. and E. Solea (2018). A nonparametric graphical model for functional data with application to brain networks based on fMRI. *Journal of the American Statistical Association 113*(524), 1637–1655.

Lundborg, A. R., R. D. Shah, and J. Peters (2021). Conditional independence testing in Hilbert spaces with applications to functional data analysis. *arXiv preprint arXiv:2101.07108*.

Marzetti, L., A. Basti, F. Chella, A. D'Andrea, J. Syrjälä, and V. Pizzella (2019). Brain functional connectivity through phase coupling of neuronal oscillations: a perspective from magnetoencephalography. *Frontiers in Neuroscience 13*, 964.

Morettin, P. A., A. Pinheiro, and B. Vidakovic (2017). *Wavelets in functional data analysis*. Springer.

Morris, J. S. (2015). Functional regression. *Annual Review of Statistics and Its Application 2*(1), 321–359.

Ogden, R. T. (1997). *Essential wavelets for statistical applications and data analysis*. Springer Science & Business Media.

Patilea, V., C. Sánchez-Sellero, and M. Saumard (2016). Testing the predictor effect on a functional response. *Journal of the American Statistical Association 111*(516), 1684–1695.

Qiao, X., S. Guo, and G. M. James (2019). Functional graphical models. *Journal of the American Statistical Association 114*(525), 211–222.

Qiao, X., C. Qian, G. M. James, and S. Guo (2020). Doubly functional graphical models in high dimensions. *Biometrika 107*(2), 415–431.

Ramsay, J. and B. Silverman (2005). *Functional data analysis*. Springer, New York.

Sang, P., L. Wang, and J. Cao (2019). Weighted empirical likelihood inference for dynamical correlations. *Computational Statistics & Data Analysis 131*, 194–206.

Sejdinovic, D., B. Sriperumbudur, A. Gretton, and K. Fukumizu (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics 41*(5), 2263–2291.

Shen, Q. and J. Faraway (2004). A F test for linear models with functional responses. *Statistica Sinica 14*, 1239–1257.

Shin, H. and S. Lee (2015). Canonical correlation analysis for irregularly and sparsely observed functional data. *Journal of Multivariate Analysis 134*, 1–18.

Solea, E. and H. Dette (2021). Nonparametric and high-dimensional functional graphical models. *arXiv preprint arXiv:2103.10568*.

Solea, E. and B. Li (2020). Copula Gaussian graphical models for functional data. *Journal of the American Statistical Association*, in press.

Székely, G. J. and M. L. Rizzo (2013). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis 117*, 193–213.

Székely, G. J., M. L. Rizzo, and N. K. Bakirov (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics 35*(6), 2769–2794.

Taqqu, M. S. (1975). Weak convergence to fractional Brownian motion and to the Rosenblatt process. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 31*(4), 287–302.

Tsai, S.-Y. (2018). Reproducibility of structural brain connectivity and network metrics using probabilistic diffusion tractography. *Scientific Reports 8*(1), 1–12.

Van Wijk, B. C., C. J. Stam, and A. Daffertshofer (2010). Comparing brain networks of different size and connectivity density using graph theory. *PLOS ONE 5*(10), e13701.

Vidakovic, B. (2009). *Statistical modeling by wavelets*, Volume 503. John Wiley & Sons.

Wang, Y. (1996). Function estimation via wavelet shrinkage for long-memory data. *Annals of Statistics 24*(2), 466–484.

Zapata, J., S.-Y. Oh, and A. Petersen (2019). Partial Separability and Functional Graphical Models for Multivariate Gaussian Processes. *arXiv preprint arXiv:1910.03134*.

Zhao, B., S. Zhai, Y. S. Wang, and M. Kolar (2021). High-dimensional functional graphical model structure learning via neighborhood selection approach. *arXiv preprint arXiv:2105.02487*.

Zhou, Y., S.-C. Lin, and J.-L. Wang (2018). Local and global temporal correlations for longitudinal data. *Journal of Multivariate Analysis 167*, 1–14.

Zhu, H., N. Strawn, and D. B. Dunson (2016). Bayesian graphical models for multivariate functional data. *Journal of Machine Learning Research 17*(204), 1–27.

Supplementary Material for "A Wavelet-Based Independence Test for Functional Data with an Application to MEG Functional Connectivity"

Rui Miao, Xiaoke Zhang *

Department of Statistics, George Washington University

and
Raymond K. W. Wong †
Department of Statistics, Texas A&M University

August 11, 2022

S1 Background Materials

S1.1 Distance-Induced Characteristic Kernels

Characteristic kernels are required to construct HSIC for two random functions under the RKHS framework. Such a kernel can be generated by a semi-metric of strong negative type.

Definition S1 (Strong Negative Type Semi-Metric)**.** *A semi-metric $\rho : \mathcal{Z} \times \mathcal{Z} \to [0, \infty)$ defined on a non-empty set $\mathcal{Z}$ is of negative type if $\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \rho(z_i, z_j) \leq 0$ for all $z_1, \ldots, z_n \in \mathcal{Z}$ and $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ such that $\sum_{i=1}^{n} \alpha_i = 0$, $n \geq 2$. Furthermore, it is of strong negative type if for any two probability measures P and P' on $\mathcal{Z}$ such that $\int_{\mathcal{Z}} \rho(z, z_0) \mathrm{d}P(z), \int_{\mathcal{Z}} \rho(z, z_0) \mathrm{d}P'(z) < \infty$ for some $z_0 \in \mathcal{Z}$, we have $\int_{\mathcal{Z}} \int_{\mathcal{Z}} \rho(z_1, z_2) \, \mathrm{d}(P - P')(z_1) \mathrm{d}(P - P')(z_2) = 0$ if and only if $P = P'$.*

Proposition S1 shows that a kernel induced by a strong negative type semi-metric is characteristic.

Proposition S1. *Let ρ be a semi-metric defined on $\mathcal{Z}$ and $z_0 \in \mathcal{Z}$. The induced kernel $\kappa_\rho(z, z') = \rho(z, z_0) + \rho(z', z_0) - \rho(z, z')$, $z, z' \in \mathcal{Z}$, is symmetric and positive definite. More-*

*The research of Xiaoke Zhang is partially supported by National Science Foundation grant DMS-1832046.
†The research of Raymond K. W. Wong is partially supported by National Science Foundation grants DMS-1806063, DMS-1711952 and CCF-1934904.

over, for all probability measures P such that $\int_{\mathcal{Z}} \rho(z, z_0) dP(z) < \infty$ for some $z_0 \in \mathcal{Z}$, κ_ρ is characteristic if and only if ρ is of strong negative type.

Obviously distance-induced kernels are symmetric. For the proof of Proposition S1, see Lemma 2.1 of Berg et al. (1984) for positive definiteness and Lyons (2013) and Sejdinovic et al. (2013) for the characteristic property. Since the set $\mathcal{Z}$ of interest often contains zero, in this paper we always set $z_0 = 0$ for any distance-induced kernel κ_ρ for simplicity and convenience.

S1.2 Besov Spaces and Norms

The Besov space is a generalization of the Sobolev space, which is widely used in nonparametric regression under the RKHS framework. A Besov space $B^\alpha_{p,q}[0,1], p, q, \alpha > 0$, contains all functions of which Besov norm $\|\cdot\|_{B^\alpha_{p,q}}$ is finite. Explicitly, with any integer $r \geq 1$, define the rth order difference of a function f by

$$\Delta_h^r(f, x) = \sum_{k=0}^{r} \binom{r}{k} (-1)^{r-k} f(x + kh),$$

and its rth order modulus of continuity by

$$\omega_r(f, t)_p = \sup_{0 \leq h \leq t} \|\Delta_h^r(f, \cdot) \mid_{[0,1-rh]} \|_{L^p},$$

where $\Delta_h^r(f, \cdot) \mid_{[0,1-rh]}$ represents $\Delta_h^r(f, \cdot)$ restricted on $[0, 1 - rh]$ and $\|\cdot\|_{L^p}$ is the L^p norm. Then the Besov norm of f is defined by

$$\|f\|_{B^\alpha_{p,q}} = \|f\|_{L^p} + |f|_{B^\alpha_{p,q}}, \quad \text{where} \quad |f|_{B^\alpha_{p,q}} = \left[\int_0^\infty \left\{ \frac{\omega_r(f, t)_p}{t^\alpha} \right\}^q \frac{\mathrm{d}t}{t} \right]^{\frac{1}{q}}.$$

For the same α, the Besov norms generated by different values of $r > \alpha$ are equivalent when $p > 1$ (DeVore and Lorentz, 1993). In this paper we always assume $p > 1$ and $r = \lfloor \alpha \rfloor + 1$ where $\lfloor \alpha \rfloor$ is the greatest integer less than or equal to α.

The Besov norm (semi-norm) generalizes some traditional smoothness measures, such as the Sobolev semi-norm $|\cdot|_{W^k_p}$

$$|f|_{W^k_p} = \left(\int_0^1 \left| D^k f \right|^p \mathrm{d}x \right)^{1/p}, \quad 1 \leq p \leq \infty,$$

where D^k is kth order weak-derivative operator.

S2 Technical Proofs

S2.1 Proof of Theorem 1

We first list two lemmas on some properties of negative type semi-metrics, which will be needed in the proof of Theorem 1.

Definition S2 (Radial Positive Definite Function)**.** *A real function F defined on $[0,\infty)$ is called radial positive definite on the semi-metric space $(\mathcal{Z},\rho)$ if F is continuous and*

$$\sum_{j=1}^{n}\sum_{k=1}^{n} F(\rho(z_j,z_k))c_jc_k \geq 0,$$

for all choices of $n \geq 1$ points $z_1,\ldots,z_n \in \mathcal{Z}$. We denote the set of all radial positive definite functions by $\mathrm{RPD}(\mathcal{Z})$.

Lemma S1. *The following hold in any semi-metric space $\mathcal{Z}$.*

(a) $\mathrm{RPD}(\mathcal{Z})$ *is never empty.*

(b) If $F_1, F_2 \in \mathrm{RPD}(\mathcal{Z})$*, then* $F_1 \cdot F_2 \in \mathrm{RPD}(\mathcal{Z})$.

(c) If $F_j \in \mathrm{RPD}(\mathcal{Z})$ *and* $0 \leq c_j < \infty$*,* $j = 1,\ldots,n$*, then* $\sum_{j=1}^{n} c_jF_j \in \mathrm{RPD}(\mathcal{Z})$.

(d) If $F_j \in \mathrm{RPD}(\mathcal{Z}), j = 1,2,\ldots$ *and the F_j converge point-wise to a continuous limit F, then* $F \in \mathrm{RPD}(\mathcal{Z})$.

(e) For space $(L^p, \|\cdot\|_p)$*,* $(\ell^p, \|\cdot\|_p)$ *with* $0 < p \leq 2$*, then* $\exp(-t^\alpha)$ *is RPD for* $0 < \alpha \leq p$.

Lemma S1 is a combination of Theorems 4.4 and 4.10 of Wells and Williams (2012).

Lemma S2 (Theorem 4.5, Wells and Williams (2012))**.** *In a semi-metric space $(\mathcal{Z},\rho)$, the following are equivalent:*

(a) ρ is of negative type;

(b) the function $\exp(-\lambda t)$ *belongs to* $\mathrm{RPD}(\mathcal{Z},\rho)$ *for* $\lambda > 0$*;*

(c) $(\mathcal{Z},\rho^{1/2})$ *is isometrically embeddable in a Hilbert space.*

Lemma S3 (Theorem 4.7, Wells and Williams (2012))**.** *If semi-metric ρ is of negative type on $\mathcal{Z}$, then ρ^r is of negative type for any $0 < r < 1$.*

Proof of Theorem 1. By Proposition S1, it suffices to prove that $\rho_{b^\alpha_{p,q}}$ is of strong negative type. Lemmas S1 (e) and S2 (a) ensure that $\bar{\rho}_j(f,g) := \|\boldsymbol{\theta}_j^f - \boldsymbol{\theta}_j^g\|_p^q, j = -1, 0, 1, \ldots$ are of negative type for $q \le p \le 2$. By Lemma S2, the function $F_j(t) = \exp(-2^{sjq}t)$ belongs to $\mathrm{RPD}(B^{\alpha'}_{p,q}[0,1])$, where $s = \alpha + 1/2 - 1/p$. For any finite product, by Lemma S1 (b)

$$\prod_{j=-1}^{n} F_j(\bar{\rho}_j) = \exp\left\{ -\sum_{j=-1}^{n} 2^{sjq}\bar{\rho}_j \right\} \tag{S1}$$

belongs to $\mathrm{RPD}(B^{\alpha'}_{p,q}[0,1])$. Lemma S1 (d) ensures the continuous sequence limit of (S1), i.e., $\exp(-\rho_{b^\alpha_{p,q}}) \in \mathrm{RPD}(B^{\alpha'}_{p,q}[0,1])$ as $n \to \infty$. Therefore $\sum_{j\ge -1} 2^{sjq}\bar{\rho}_j$ is of negative type on $B^{\alpha'}_{p,q}[0,1]$. By Lemma S2 (c), the $\left(B^{\alpha'}_{p,q}[0,1], \left(\sum_{j\ge -1} 2^{sjq}\bar{\rho}_j\right)^{1/2}\right)$ is a metric space isometrically embeddable in a Hilbert space. By the same procedure of Remark 3.19 in Lyons (2013), the map

$$P \mapsto \left(f \mapsto \int_{B^{\alpha'}_{p,q}[0,1]} \left(\sum_{j\ge -1} 2^{sjq}\bar{\rho}_j \right)^{r/2} (f,g)\mathrm{d}P(g) \right)$$

is injective for any $r \in (0,\infty)\backslash 2\mathbb{N}$, where $\mathbb{N}$ is the set of natural numbers (Linde, 1986). The result folllows from the fact that $(\sum_{j\ge -1} 2^{sjq}\bar{\rho}_j)^{r/2}$ is of negative type for any $r \in (0,2)$ by Lemma S3. □

S2.2 Proof of Theorem 2

Proof of Theorem 2. Here we prove a more general result where τ_j and $j^Z_\#$ involved in the penalty $\mathrm{pen}_j(k) = k\zeta_Z\{1 + \sqrt{2(1+2\varsigma_Z)\log(\tau_j m_j/k)}\}^2$ in Step 1 are replaced by $\tau_j = \tau_Z 2^{2\beta'_Z(j - j^Z_\#)_+}$ and $j^Z_\# = (1+(\varsigma_Z+1/2)/\beta'_Z)/(\alpha_Z+\varsigma_Z+1/2)\cdot\log_2 \delta_Z^{-1}$ for any $\beta_Z < \beta'_Z \le \alpha_Z$ respectively. Apparently Theorem 2 is a special case where $\beta'_Z = \alpha_Z$.

For notational simplicity, we omit the subscript Z and subject index i in all terms; namely we replace $\delta_{Z,j}$ by δ_j, δ_Z by δ, ς_Z by ς, C_Z by C, α_Z by α, β_Z by β and β'_Z by β' respectively. We further replace $\boldsymbol{\theta}^{e_i^Z}$ by $\boldsymbol{\theta}$ and $\theta^{e_i^Z}_{j,k}$ by $\theta_{j,k}$ respectively.

We first decompose the loss function by

$$\mathrm{E}\left(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2_{b^\beta} \mid \boldsymbol{\theta} \right) = \sum_{j\ge -1} 2^{2\beta j} \mathrm{E}\left(\|\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j\|^2 \mid \boldsymbol{\theta}_j \right).$$

By Theorem 11.11 in Johnstone (2019), there exist constants $a(\zeta)$ and $b(\zeta)$ that depend on ζ,

and $M_j = M_j(\varsigma, \tau_j)$ that depends on ς and τ_j such that

$$\mathrm{E}\left(\|\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j\|_2^2 \mid \boldsymbol{\theta}_j\right) \leq b(\zeta)\xi_1 M_j \delta_j^2 + a(\zeta)\mathcal{R}_j(\boldsymbol{\theta}_j, \delta_j),$$

where $\mathcal{R}_j(\boldsymbol{\theta}_j, \delta_j) = \min_{K \subseteq \{0,\ldots,2^{j-1}\}} \left[\sum_{k \notin K} \theta_{j,k}^2 + \delta_j^2 \mathrm{pen}_j(\|\boldsymbol{\theta}_j\|_0)\right]$. Therefore,

$$\mathrm{E}\left(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_{b^\beta}^2 \mid \boldsymbol{\theta}\right) = b(\zeta)\xi_1 \sum_j 2^{2\beta j} M_j \delta_j^2 + a(\zeta) \sum_j 2^{2\beta j} \mathcal{R}_j(\boldsymbol{\theta}_j, \delta_j),$$

and it suffices to study the upper bounds of (I) $= \sum_j 2^{2\beta j} M_j \delta_j^2$ and (II) $= \sum_j 2^{2\beta j} \mathcal{R}_j(\boldsymbol{\theta}_j, \delta_j)$ respectively.

Bound of (I) By (11.67) in Johnstone (2019), $M_j \leq \tau^{-1} c_{\zeta,\tau} 2^{-2\varsigma j} 2^{-2\beta'(j-j_\#)_+}$, where $c_{\zeta,\tau}$ is a constant that depends on ζ and τ. Thus

$$\begin{aligned}
\text{(I)} = \sum_j 2^{2\beta j} M_j \delta_j^2 &\leq \tau^{-1} c_{\varsigma,\tau} \delta^2 \left(\sum_{j=-1}^{j_\#} 2^{2\beta j} + 2^{2\beta' j_\#} \sum_{j > j_\#} 2^{-2(\beta'-\beta)j} \right) \\
&= \tau^{-1} c_{\varsigma,\tau} \delta^2 \left(\sum_{j=-1}^{j_\#} 2^{2\beta j} + 2^{2\beta j_\#} \sum_{j > j_\#} 2^{-2(\beta'-\beta)(j-j_\#)} \right) \\
&= \tau^{-1} c_{\varsigma,\tau} \delta^2 \left(2^{-2\beta} + \frac{2^{2\beta(j_\#+1)} - 1}{2^{2\beta} - 1} + 2^{2\beta j_\#} \sum_{j=1}^{\infty} 2^{-2(\beta'-\beta)(j-j_\#)} \right) \\
&\leq \tau^{-1} c_{\varsigma,\tau} \delta^2 \left(2^{-2\beta} + \frac{2^{2\beta} (\delta^{-w})^{2\beta}}{2^{2\beta} - 1} + (\delta^{-w})^{2\beta} / (1 - 2^{-2(\beta'-\beta)}) \right) \\
&\leq c_{\varsigma,\tau,\beta,\beta'} \delta^{2(1-w\beta)},
\end{aligned}$$

where $w = (\alpha + \varsigma + 1/2)^{-1}[1 + (\varsigma + 1/2)/\beta]$ and $c_{\zeta,\tau}$ is a constant that depends on ς, τ, β and β'.

Bound of (II) According to (11.40) in Johnstone (2019),

$$\sup_{\boldsymbol{\theta}_j : \|\boldsymbol{\theta}\|_{b^\alpha} \leq C} \mathcal{R}_j(\boldsymbol{\theta}_j, \delta_j) \leq c_{\zeta,\xi_1,\varsigma} \log \tau_j r_j(C_j, \delta_j),$$

where $C_j = 2^{-\alpha j}, m_j = 2^j$, $c_{\zeta,\xi_1,\varsigma}$ is a constant that depends on ζ, ξ_1 and ς, and

$$r_j(C_j, \delta_j) = \begin{cases} C_j^2, & \text{if } C_j \le \delta_j m_j^{1/2}, \\ m_j \delta_j^2, & \text{if } C_j \ge \delta_j m_j^{1/2}. \end{cases}$$

Notice that $\log \tau_j = \log \tau + 2\beta'(\log 2)(j - j_\#)_+$, so we have

$$\begin{aligned} \text{(II)} \quad &\le c_{\zeta,\xi_1,\varsigma} \left\{ (\log \tau) \textstyle\sum_{j \ge -1} 2^{2\beta j} r_j(C_j, \delta_j) + 2\beta'(\log 2) \sum_{j > j_\#} (j - j_\#) 2^{2\beta j} r_j(C_j, \delta_j) \right\} \\ &= c_{\zeta,\xi_1,\varsigma} \left\{ (\log \tau) \textstyle\sum_{j \ge -1} Q_j + 2\beta'(\log 2) \sum_{j > j_\#} (j - j_\#) Q_j \right\}, \end{aligned} \tag{S2}$$

where $Q_j = 2^{2\beta j} r_j(C_j, \delta_j)$. Next we handle (III) $= \sum_{j \ge -1} Q_j$ and (IV) $= \sum_{j > j_\#} (j - j_\#) Q_j$ individually.

- (III) $= \sum_{j \ge -1} Q_j$. We calculate Q_j respectively for $j \ge -1$. Define $j_* = (\alpha + \varsigma + 1/2)^{-1} \log_2(C/\delta)$.

  1° When $j \le j_*$, $C_j \ge \delta_j m_j^{1/2}$, so that

  $$Q_j = 2^{2\beta j} m_j \delta_j^2 = 2^{(2\beta + 2\varsigma + 1) j} \delta^2.$$

  2° When $j \ge j_*$, $C_j \le \delta_j m_j^{1/2}$, so that

  $$Q_j = 2^{2\beta j} C_j^2 = 2^{-2(\alpha - \beta) j} C^2.$$

  Combining 1° and 2°, we have

  $$Q_j = \begin{cases} Q^* 2^{(2\beta + 2\varsigma + 1)(j - j_*)}, & j \le j_*, \\ Q^* 2^{-(\alpha - \beta)(j - j_*)}, & j \ge j_*, \end{cases}$$

  where $Q^* = C^{2(1-r)} \delta^{2r}$ with $r = (\alpha - \beta)/(\alpha + \varsigma + 1/2)$. Therefore, (III)$\le c_1 Q^*$.

- (IV) $= \sum_{j > j_\#} (j - j_\#) Q_j$. When m is sufficiently large, $\delta \asymp m^{-1/2} \to 0$, and $j_\# > j_*$ since

$1 + (\varsigma + 1/2)/\beta' > 1$. Thus for m large enough,

$$(\text{IV}) \leq \sum_{j \geq j_*} (j - j_*) Q_j = Q^* \sum_{j \geq j^*} 2^{-(\alpha-\beta)(j-j_*)} \leq c_2 Q^*.$$

Hence by (S2), (II)$\leq c_3 C^{2(1-r)} \delta^{2r}$ where the constant c_3 depends on $c_1, c_2, \zeta, \xi_1, \varsigma, \tau$ and β'. Combining the upper bounds for (I) and (II) respectively, we have

$$\sup_{\boldsymbol{\theta}: \|\boldsymbol{\theta}\|_{b^\beta} \leq C} \mathrm{E}\left(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_{b^\beta}^2 \mid \boldsymbol{\theta} \right) \leq b(\zeta) \xi_1 c_{\varsigma,\tau,\beta,\beta'} \delta^{2(1-w\beta)} + c_3 C^{2(1-r)} \delta^{2r} = O(\delta^{2r}),$$

since

$$1 - w\beta = 1 - \frac{(1 + (\varsigma + 1/2)/\beta')\beta}{\alpha + \varsigma + 1/2} \geq 1 - \frac{(1 + (\varsigma + 1/2)/\beta)\beta}{\alpha + \varsigma + 1/2} = \frac{\alpha - \beta}{\alpha + \varsigma + 1/2} = r.$$

□

S2.3 Proof of Theorem 3

We first present a lemma that will be used to prove Theorem 3.

Lemma S4. *Let $\{(X_i(\cdot), Y_i(\cdot)\}_{i=1}^n$ be i.i.d. fully observed random samples from probability measure $P_{XY} = P_X P_Y$ defined on $\mathcal{X} \otimes \mathcal{Y}$. Then as $n \to \infty$,*

$$n\gamma(P_{n,XY}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}}) \rightsquigarrow \sum_{r=1}^{\infty} \sum_{s=1}^{\infty} \mu_r \nu_s N_{rs}^2, \tag{S3}$$

where $N_{rs} \sim N(0,1), r, s \in \mathbb{N}$ are i.i.d. and $\{\mu_r\}_{r=1}^{\infty}$ and $\{\nu_s\}_{s=1}^{\infty}$ are eigenvalues of the integral kernel operators $S_{\tilde{\kappa}_{\mathcal{X}}}$ and $S_{\tilde{\kappa}_{\mathcal{Y}}}$, respectively. If $P_{XY} \neq P_X P_Y$, then $n\gamma(P_{n,XY}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}}) \to \infty$ in probability as $n \to \infty$.

Lemma S4 is exactly Theorem 33 of Sejdinovic et al. (2013), which provides the weak convergence result of HSIC for fully observed random functions.

Proof of Theorem 3. According to Lemma S4, it suffices to prove that the difference between HSIC based on original curves $\{X_i(\cdot), Y_i(\cdot)\}_{i=1}^n$ and HSIC based on denoised curves $\{\hat{X}_i, \hat{Y}_i\}_{i=1}^n$

is $o_p(1/n)$, where $\{\hat{X}_i, \hat{Y}_i\}_{i=1}^n$ are obtained by Step 1 in Section 3. By Definition 1,

$$
\begin{aligned}
&n\left|\gamma(P_{n,XY},\kappa_{\mathcal{X}},\kappa_{\mathcal{Y}})-\gamma(P_{n,\hat{X}\hat{Y}},\kappa_{\mathcal{X}},\kappa_{\mathcal{Y}})\right| = n^{-1}\left|\|\boldsymbol{\kappa}_{\mathcal{X}}^\top\mathbf{H}\boldsymbol{\kappa}_{\mathcal{Y}}\|^2_{\mathcal{H}(\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}})}-\|\hat{\boldsymbol{\kappa}}_{\mathcal{X}}^\top\mathbf{H}\hat{\boldsymbol{\kappa}}_{\mathcal{Y}}\|^2_{\mathcal{H}(\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}})}\right| \\
=&n^{-1}\left|\|\boldsymbol{\kappa}_{\mathcal{X}}^\top\mathbf{H}\boldsymbol{\kappa}_{\mathcal{Y}}\|_{\mathcal{H}(\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}})}-\|\hat{\boldsymbol{\kappa}}_{\mathcal{X}}^\top\mathbf{H}\hat{\boldsymbol{\kappa}}_{\mathcal{Y}}\|_{\mathcal{H}(\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}})}\right|\left(\|\boldsymbol{\kappa}_{\mathcal{X}}^\top\mathbf{H}\boldsymbol{\kappa}_{\mathcal{Y}}\|_{\mathcal{H}(\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}})}+\|\hat{\boldsymbol{\kappa}}_{\mathcal{X}}^\top\mathbf{H}\hat{\boldsymbol{\kappa}}_{\mathcal{Y}}\|_{\mathcal{H}(\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}})}\right) \\
\le&n^{-1}\|\boldsymbol{\kappa}_{\mathcal{X}}^\top\mathbf{H}\boldsymbol{\kappa}_{\mathcal{Y}}-\hat{\boldsymbol{\kappa}}_{\mathcal{X}}^\top\mathbf{H}\hat{\boldsymbol{\kappa}}_{\mathcal{Y}}\|_{\mathcal{H}(\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}})}\left(\|\boldsymbol{\kappa}_{\mathcal{X}}^\top\mathbf{H}\boldsymbol{\kappa}_{\mathcal{Y}}\|_{\mathcal{H}(\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}})}+\|\hat{\boldsymbol{\kappa}}_{\mathcal{X}}^\top\mathbf{H}\hat{\boldsymbol{\kappa}}_{\mathcal{Y}}\|_{\mathcal{H}(\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}})}\right) \\
\le&2n^{-1/2}\|\boldsymbol{\kappa}_{\mathcal{X}}^\top\mathbf{H}\boldsymbol{\kappa}_{\mathcal{Y}}-\hat{\boldsymbol{\kappa}}_{\mathcal{X}}^\top\mathbf{H}\hat{\boldsymbol{\kappa}}_{\mathcal{Y}}\|_{\mathcal{H}(\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}})}\times n^{-1/2}\|\boldsymbol{\kappa}_{\mathcal{X}}^\top\mathbf{H}\boldsymbol{\kappa}_{\mathcal{Y}}\|_{\mathcal{H}(\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}})}+n^{-1}\|\boldsymbol{\kappa}_{\mathcal{X}}^\top\mathbf{H}\boldsymbol{\kappa}_{\mathcal{Y}}-\hat{\boldsymbol{\kappa}}_{\mathcal{X}}^\top\mathbf{H}\hat{\boldsymbol{\kappa}}_{\mathcal{Y}}\|^2_{\mathcal{H}(\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}})}
\end{aligned}
$$

where $\boldsymbol{\kappa}_{\mathcal{X}}^\top=[\kappa_{\mathcal{X}}(\cdot,X_1),\ldots,\kappa_{\mathcal{X}}(\cdot,X_n)]$, $\boldsymbol{\kappa}_{\mathcal{Y}}^\top=[\kappa_{\mathcal{Y}}(\cdot,Y_1),\ldots,\kappa_{\mathcal{Y}}(\cdot,Y_n)]$, $\hat{\boldsymbol{\kappa}}_{\mathcal{X}}^\top=\left[\kappa_{\mathcal{X}}(\cdot,\hat{X}_1),\ldots,\kappa_{\mathcal{X}}(\cdot,\hat{X}_n)\right]$, $\hat{\boldsymbol{\kappa}}_{\mathcal{Y}}^\top=\left[\kappa_{\mathcal{Y}}(\cdot,\hat{Y}_1),\ldots,\kappa_{\mathcal{Y}}(\cdot,\hat{Y}_n)\right]$.

By (S3),

$$
n^{-1/2}\|\boldsymbol{\kappa}_{\mathcal{X}}^\top\mathbf{H}\boldsymbol{\kappa}_{\mathcal{Y}}\|_{\mathcal{H}(\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}})}\rightsquigarrow\sqrt{\sum_{r=1}^{\infty}\sum_{s=1}^{\infty}\mu_r\nu_s N_{rs}^2}=O_p(1), \tag{S4}
$$

so it suffices to prove that $\|\boldsymbol{\kappa}_{\mathcal{X}}^\top\mathbf{H}\boldsymbol{\kappa}_{\mathcal{Y}}-\hat{\boldsymbol{\kappa}}_{\mathcal{X}}^\top\mathbf{H}\hat{\boldsymbol{\kappa}}_{\mathcal{Y}}\|_{\mathcal{H}(\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}})}=o_p\left(n^{1/2}\right)$.

Notice that $\|\boldsymbol{\kappa}_{\mathcal{X}}^\top\mathbf{H}\boldsymbol{\kappa}_{\mathcal{Y}}-\hat{\boldsymbol{\kappa}}_{\mathcal{X}}^\top\mathbf{H}\hat{\boldsymbol{\kappa}}_{\mathcal{Y}}\|_{\mathcal{H}(\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}})}$ can be bounded by the following inequality:

$$
\begin{aligned}
&\|\boldsymbol{\kappa}_{\mathcal{X}}^\top\mathbf{H}\boldsymbol{\kappa}_{\mathcal{Y}}-\hat{\boldsymbol{\kappa}}_{\mathcal{X}}^\top\mathbf{H}\hat{\boldsymbol{\kappa}}_{\mathcal{Y}}\|_{\mathcal{H}(\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}})}=\|\boldsymbol{\kappa}_{\mathcal{X}}^\top\mathbf{H}\left(\boldsymbol{\kappa}_{\mathcal{Y}}-\hat{\boldsymbol{\kappa}}_{\mathcal{Y}}\right)+\left(\boldsymbol{\kappa}_{\mathcal{X}}-\hat{\boldsymbol{\kappa}}_{\mathcal{X}}\right)\mathbf{H}\hat{\boldsymbol{\kappa}}_{\mathcal{Y}}^\top\|_{\mathcal{H}(\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}})} \\
\le&\|\boldsymbol{\kappa}_{\mathcal{X}}^\top\mathbf{H}\left(\boldsymbol{\kappa}_{\mathcal{Y}}-\hat{\boldsymbol{\kappa}}_{\mathcal{Y}}\right)\|_{\mathcal{H}(\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}})}+\|\left(\boldsymbol{\kappa}_{\mathcal{X}}-\hat{\boldsymbol{\kappa}}_{\mathcal{X}}\right)\mathbf{H}\hat{\boldsymbol{\kappa}}_{\mathcal{Y}}^\top\|_{\mathcal{H}(\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}})} \\
\le&\|\boldsymbol{\kappa}_{\mathcal{X}}^\top\mathbf{H}\left(\boldsymbol{\kappa}_{\mathcal{Y}}-\hat{\boldsymbol{\kappa}}_{\mathcal{Y}}\right)\|_{\mathcal{H}(\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}})}+\|\left(\boldsymbol{\kappa}_{\mathcal{X}}-\hat{\boldsymbol{\kappa}}_{\mathcal{X}}\right)\mathbf{H}\boldsymbol{\kappa}_{\mathcal{Y}}^\top\|_{\mathcal{H}(\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}})}+\|\left(\boldsymbol{\kappa}_{\mathcal{X}}-\hat{\boldsymbol{\kappa}}_{\mathcal{X}}\right)\mathbf{H}\left(\boldsymbol{\kappa}_{\mathcal{Y}}-\hat{\boldsymbol{\kappa}}_{\mathcal{Y}}\right)^\top\|_{\mathcal{H}(\kappa_{\mathcal{X}}\otimes\kappa_{\mathcal{Y}})} \\
=&\mathrm{tr}^{\frac{1}{2}}\left(\boldsymbol{\Gamma}^X\mathbf{H}\langle\boldsymbol{\kappa}_{\mathcal{Y}}-\hat{\boldsymbol{\kappa}}_{\mathcal{Y}},\boldsymbol{\kappa}_{\mathcal{Y}}^\top-\hat{\boldsymbol{\kappa}}_{\mathcal{Y}}^\top\rangle_{\mathcal{H}(\kappa_{\mathcal{Y}})}\mathbf{H}\right)+\mathrm{tr}^{\frac{1}{2}}\left(\boldsymbol{\Gamma}^Y\mathbf{H}\langle\boldsymbol{\kappa}_{\mathcal{X}}-\hat{\boldsymbol{\kappa}}_{\mathcal{X}},\boldsymbol{\kappa}_{\mathcal{X}}^\top-\hat{\boldsymbol{\kappa}}_{\mathcal{X}}^\top\rangle_{\mathcal{H}(\kappa_{\mathcal{X}})}\mathbf{H}\right) \\
&+\mathrm{tr}^{\frac{1}{2}}\left(\langle\boldsymbol{\kappa}_{\mathcal{X}}-\hat{\boldsymbol{\kappa}}_{\mathcal{X}},\boldsymbol{\kappa}_{\mathcal{X}}^\top-\hat{\boldsymbol{\kappa}}_{\mathcal{X}}^\top\rangle_{\mathcal{H}(\kappa_{\mathcal{X}})}\mathbf{H}\langle\boldsymbol{\kappa}_{\mathcal{Y}}-\hat{\boldsymbol{\kappa}}_{\mathcal{Y}},\boldsymbol{\kappa}_{\mathcal{Y}}^\top-\hat{\boldsymbol{\kappa}}_{\mathcal{Y}}^\top\rangle_{\mathcal{H}(\kappa_{\mathcal{Y}})}\mathbf{H}\right) \\
=&\mathrm{tr}^{\frac{1}{2}}\left(\check{\boldsymbol{\Gamma}}^X\langle\boldsymbol{\kappa}_{\mathcal{Y}}-\hat{\boldsymbol{\kappa}}_{\mathcal{Y}},\boldsymbol{\kappa}_{\mathcal{Y}}^\top-\hat{\boldsymbol{\kappa}}_{\mathcal{Y}}^\top\rangle_{\mathcal{H}(\kappa_{\mathcal{Y}})}\right)+\mathrm{tr}^{\frac{1}{2}}\left(\check{\boldsymbol{\Gamma}}^Y\langle\boldsymbol{\kappa}_{\mathcal{X}}-\hat{\boldsymbol{\kappa}}_{\mathcal{X}},\boldsymbol{\kappa}_{\mathcal{X}}^\top-\hat{\boldsymbol{\kappa}}_{\mathcal{X}}^\top\rangle_{\mathcal{H}(\kappa_{\mathcal{X}})}\right) \\
&+\mathrm{tr}^{\frac{1}{2}}\left(\langle\boldsymbol{\kappa}_{\mathcal{X}}-\hat{\boldsymbol{\kappa}}_{\mathcal{X}},\boldsymbol{\kappa}_{\mathcal{X}}^\top-\hat{\boldsymbol{\kappa}}_{\mathcal{X}}^\top\rangle_{\mathcal{H}(\kappa_{\mathcal{X}})}\mathbf{H}\langle\boldsymbol{\kappa}_{\mathcal{Y}}-\hat{\boldsymbol{\kappa}}_{\mathcal{Y}},\boldsymbol{\kappa}_{\mathcal{Y}}^\top-\hat{\boldsymbol{\kappa}}_{\mathcal{Y}}^\top\rangle_{\mathcal{H}(\kappa_{\mathcal{Y}})}\mathbf{H}\right) \\
\le&\mathrm{tr}^{\frac{1}{2}}(\check{\boldsymbol{\Gamma}}^X)\mathrm{tr}^{\frac{1}{2}}(\langle\boldsymbol{\kappa}_{\mathcal{Y}}-\hat{\boldsymbol{\kappa}}_{\mathcal{Y}},\boldsymbol{\kappa}_{\mathcal{Y}}^\top-\hat{\boldsymbol{\kappa}}_{\mathcal{Y}}^\top\rangle_{\mathcal{H}(\kappa_{\mathcal{Y}})})+\mathrm{tr}^{\frac{1}{2}}(\check{\boldsymbol{\Gamma}}^Y)\mathrm{tr}^{\frac{1}{2}}(\langle\boldsymbol{\kappa}_{\mathcal{X}}-\hat{\boldsymbol{\kappa}}_{\mathcal{X}},\boldsymbol{\kappa}_{\mathcal{X}}^\top-\hat{\boldsymbol{\kappa}}_{\mathcal{X}}^\top\rangle_{\mathcal{H}(\kappa_{\mathcal{X}})}) \qquad (*) \\
&+\mathrm{tr}^{\frac{1}{2}}(\langle\boldsymbol{\kappa}_{\mathcal{X}}-\hat{\boldsymbol{\kappa}}_{\mathcal{X}},\boldsymbol{\kappa}_{\mathcal{X}}^\top-\hat{\boldsymbol{\kappa}}_{\mathcal{X}}^\top\rangle_{\mathcal{H}(\kappa_{\mathcal{X}})})\mathrm{tr}^{\frac{1}{2}}(\langle\boldsymbol{\kappa}_{\mathcal{Y}}-\hat{\boldsymbol{\kappa}}_{\mathcal{Y}},\boldsymbol{\kappa}_{\mathcal{Y}}^\top-\hat{\boldsymbol{\kappa}}_{\mathcal{Y}}^\top\rangle_{\mathcal{H}(\kappa_{\mathcal{Y}})})=o_p(n^{1/2}),
\end{aligned}
$$

where $\check{\boldsymbol{\Gamma}}^X=\mathbf{H}\boldsymbol{\Gamma}^X\mathbf{H}$ and $\check{\boldsymbol{\Gamma}}^Y=\mathbf{H}\boldsymbol{\Gamma}^Y\mathbf{H}$ are centered Gram matrices.

In (*) we used the fact that for symmetric positive definite matrices $\mathbf{A}$ and $\mathbf{B}$,

$$
\mathrm{tr}\mathbf{A}\mathbf{B}=\mathrm{vec}(\mathbf{A})^\top\mathrm{vec}(\mathbf{B})\le\|\mathbf{A}\|_F\|\mathbf{B}\|_F=\sqrt{\mathrm{tr}\mathbf{A}^2\mathrm{tr}\mathbf{B}^2}\le\mathrm{tr}\mathbf{A}\mathrm{tr}\mathbf{B}.
$$

The last equation holds due to the facts below with $(\mathcal{Z}, Z, z) = (\mathcal{X}, X, x)$ or $(\mathcal{Y}, Y, y)$:

- $\mathrm{tr}(\check{\mathbf{\Gamma}}^Z) = O_p(n)$ because $\int_{\mathcal{Z}} \check{\kappa}_{\mathcal{Z}}(z,z)\mathrm{d}P_Z(z) < \infty$ which is ensured by the assumptions in Theorem 2.

- $\mathrm{tr}\langle \boldsymbol{\kappa}_{\mathcal{Z}} - \hat{\boldsymbol{\kappa}}_{\mathcal{Z}}, \boldsymbol{\kappa}_{\mathcal{Z}}^\top - \hat{\boldsymbol{\kappa}}_{\mathcal{Z}}^\top \rangle_{\mathcal{H}(\kappa_{\mathcal{Z}})} = \sum_{i=1}^n \|\kappa_{\mathcal{Z}}(\cdot, Z_i) - \kappa_{\mathcal{Z}}(\cdot, \hat{Z}_i)\|^2_{\mathcal{H}(\kappa_{\mathcal{Z}})} = o_p(1)$, because

$$\begin{aligned}
&\|\kappa_{\mathcal{Z}}(\cdot, Z_i) - \kappa_{\mathcal{Z}}(\cdot, \hat{Z}_i)\|^2_{\mathcal{H}(\kappa_{\mathcal{Z}})} = \kappa_{\mathcal{Z}}(Z_i, Z_i) + \kappa_{\mathcal{Z}}(\hat{Z}_i, \hat{Z}_i) - 2\kappa_{\mathcal{Z}}(Z_i, \hat{Z}_i) \\
=&2\|Z_i\|_{b^{\beta_Z}} + 2\|\hat{Z}_i\|_{b^{\beta_Z}} - 2\left(\|Z_i\|_{b^{\beta_Z}} + \|\hat{Z}_i\|_{b^{\beta_Z}} - \|Z_i - \hat{Z}_i\|_{b^{\beta_Z}}\right) = 2\|Z_i - \hat{Z}_i\|_{b^{\beta_Z}},
\end{aligned}$$

  and $\|Z_i - \hat{Z}_i\|_{b^{\beta_Z}} = o_p(n^{-1}), i = 1, \ldots, n$ ensured by Theorem 2 and (4) in Theorem 3.

□

S2.4 Proof of Theorem 4

We first introduce a few notations. To perform a permutation test, let $\mathcal{S}(n) = \{\sigma_1, \ldots, \sigma_{n!}\}$ be the cyclic group of $\{1, \ldots, n\}$. For a permutation σ randomly selected from $\mathcal{S}(n)$, let $\gamma(P^{\sigma}_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}}) = n^{-2}\mathrm{tr}(\mathbf{\Gamma}^{\hat{X}}\mathbf{H}\mathbf{\Gamma}^{\hat{Y}}(\sigma)\mathbf{H})$, where $\mathbf{\Gamma}^{\hat{Y}}(\sigma)$ is generated by $\mathbf{\Gamma}^{\hat{Y}}$ with rows and columns permuted according to σ. Let R be the rank of $\gamma(P_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}})$ in all possible permuted HSICs. Then we reject $H_0 : P_{XY} = P_X P_Y$ if $p_{\hat{X}\hat{Y}} = R/n! \leq \alpha$, where $p_{\hat{X}\hat{Y}}$ denotes the p-value of the permutation test enumerating all permutations and α is the level of significance.

In practice, it is impractical to consider all permutations from $\mathcal{S}(n)$. Hence we use a Monte-Carlo approximation by randomly choosing B permutations $\sigma_1, \ldots, \sigma_B \in \mathcal{S}(n)\backslash\{\mathrm{id}\}$ where id refers to no permutation and calculating $\gamma(P_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}}), \gamma(P^{\sigma_1}_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}}), \ldots, \gamma(P^{\sigma_B}_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}})$. With a notational abuse, let R be the rank of $\gamma(P_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}})$ and we reject H_0 if $\hat{p}_{\hat{X}\hat{Y}} = R/(B+1) \leq \alpha$, where $\hat{p}_{\hat{X}\hat{Y}}$ is the p-value of the permutation test enumerating a finite sample of size B from $\mathcal{S}(n)$.

If the value of $\gamma(P_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}})$ repeats in $\{\gamma(P^{\sigma_1}_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}}), \ldots, \gamma(P^{\sigma_B}_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}})\}$ for several times with $B \leq n!$, the rank R of $\gamma(P_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}})$ is determined by the following two ways proposed by Rindt et al. (2020).

- Breaking ties at random: R is distributed uniformly on ranks of $\gamma(P^{\sigma}_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}})$ that have the same value of $\gamma(P_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}})$;

- Breaking ties conservatively: R is the largest among ranks of $\gamma(P^{\sigma}_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}})$ that have the same value of $\gamma(P_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}})$.

Next we list two lemmas which will be useful to prove Theorem 4.

Lemma S5. *For σ randomly selected from $\mathcal{S}(n)$, $\gamma(P_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}}) \to 0$ in probability as $n \to \infty$.*

Lemma S5 is a direct application of Theorem 3 of Rindt et al. (2020) for $d = 2$.

Lemma S6. *Suppose that the alternative hypothesis* $H_1 : P_{XY} \neq P_X P_Y$ *is true and noises are i.i.d. Let* $\{t_n^1(\hat{\mathcal{D}}) \geq \cdots \geq t_n^{n!}(\hat{\mathcal{D}})\}$ *be ordered values of HSIC computed on all permutations of denoised curves* $\{\gamma(P^{\sigma_1}_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}}), \ldots, \gamma(P^{\sigma_{n!}}_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}})\}$. *Let* $a = \lfloor n!\alpha \rfloor$ *for any level of significance* $\alpha \in (0,1)$. *Then* $t_n^a(\hat{\mathcal{D}}) \to 0$ *in probability as* $n \to \infty$.

Lemma S6 is a direct application of Theorem 4 of Rindt et al. (2020) for $d = 2$.

Proof of Theorem 4. Denote the fully observed dataset by $\mathcal{D} = \{(X_i, Y_i) : i = 1, \ldots, n\}$ and the denoised dataset by $\hat{\mathcal{D}} = \{(\hat{X}_i, \hat{Y}_i) : i = 1, \ldots, n\}$. For a permutation $\sigma \in \mathcal{S}(n)$, denote the permuted datasets by $\sigma(\mathcal{D})$ and $\sigma(\hat{\mathcal{D}})$, resulting in permuted HSIC $\gamma(P^{\sigma}_{n,XY}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}})$ and $\gamma(P^{\sigma}_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}})$ respectively.

If $H_0 : P_{XY} = P_X P_Y$ **is true,** then for any $\sigma \in \mathcal{S}(n)$, $\mathcal{D}$ and $\sigma(\mathcal{D})$ have the same distribution and $\hat{\mathcal{D}}$ and $\sigma(\hat{\mathcal{D}})$ have the same distribution due to the facts that the noise across subjects are i.i.d and that the denoising procedure in Section 3 is separately for each subject. For B permutations $\sigma_1, \ldots, \sigma_B$ randomly selected from $\mathcal{S}(n)\backslash\{\text{id}\}$, $(\mathcal{D}, \sigma_1(\mathcal{D}), \ldots, \sigma_B(\mathcal{D}))$ is an exchangeable vector, and thus $\left(\gamma(P_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}}), \gamma(P^{\sigma_1}_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}}), \ldots, \gamma(P^{\sigma_B}_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}})\right)$ is exchangeable.

By breaking ties at random, each entry is equally likely to have any given rank, so the rank of $\gamma(P_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}})$ is uniformly distributed in $\{1, \ldots, B\}$. Therefore the type I error rate can be controlled for any level of significance $\alpha \in (0,1)$. Breaking ties conservatively can result in an even smaller Type I error rate.

If $H_1 : P_{XY} \neq P_X P_Y$ **is true,** then by the definition of $t_n^a(\hat{\mathcal{D}})$ in Lemma S6, we reject $H_0 : P_{XY} = P_X P_Y$ if $\gamma(P_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}}) > t_n^a(\hat{\mathcal{D}})$. For any $\alpha \in (0,1)$,

$$\lim_{n\to\infty} P(p_{\hat{X}\hat{Y}} \leq \alpha) \geq \lim_{n\to\infty} P(\gamma(P_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}}) > t_n^a(\hat{\mathcal{D}})) = 1,$$

since $\gamma(P_{n,\hat{X}\hat{Y}}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}}) \to \gamma(P_{XY}, \kappa_{\mathcal{X}}, \kappa_{\mathcal{Y}}) > 0$ in probability as $n \to \infty$ by the proof of Theorem 3.

For a finite number B of permutations, the p-value $\hat{p}_{\hat{X}\hat{Y}} = (1 + U)/(B + 1)$ where $U \sim \text{Binomial}(B, p_{\hat{X}\hat{Y}})$. If $U = 0$, then $\hat{p}_{\hat{X}\hat{Y}} = 1/(B + 1) \leq \alpha$ and we reject the null hypothesis.

Since $P(p_{\hat{X}\hat{Y}} \leq \epsilon_1) \geq 1 - \epsilon_2$ for some $\epsilon_1, \epsilon_2 > 0$. For n large enough, we have

$$\begin{aligned} P(\hat{p}_{\hat{X}\hat{Y}}) &\geq P(\hat{p}_{\hat{X}\hat{Y}} = 1/(B+1) \mid p_{\hat{X}\hat{Y}} \leq \epsilon_1) P(p_{\hat{X}\hat{Y}} \leq \epsilon_1) \\ &\geq (1-\epsilon_1)^B (1-\epsilon_2). \end{aligned}$$

Then the consistency of the permutation test is proved by letting $\epsilon_1, \epsilon_2 \to 0$. □

S3 Additional Simulation

S3.1 Performance of wavHSIC for Irregular Design

In this section, we present the results of a simulation study where subjects are not measured at the same regular grid with $m = 2^{J+1}$ for some integer J.

Similar to Section 6, we had 199 simulation runs and in each simulation run $\{(X_i(t), Y_i(t)) : t \in [0,1], i = 1, \ldots, n\}$ where $n = 50$ or 200 were generated under Settings 1–3. For each subject i, $i = 1, \ldots, n$, the numbers of measurements per subject, m_i^X and m_i^Y, were both sampled from either DiscreteUnif$\{50, \ldots, 70\}$ or DiscreteUnif$\{220, \ldots, 280\}$. Given m_i^X and m_i^Y, the measurement times $\{T_{il}^X : l = 1 \ldots, m_i^X\}$ and $\{T_{il}^Y : l = 1 \ldots, m_i^Y\}$ were sampled independently on ContinuousUnif$[0,1]$. Since the number of measurements per subject and measurement times may be different across subjects, their notations here have an additional subscript "i" compared to those in Section 3. We added white Gaussian noise to all measurements with signal-to-noise ratio SNR=4 or 8. Therefore, the observed data were $\{\tilde{X}_i(T_{il}^X) = X_i(T_{il}^X) + e_{il}^X : l = 1, \ldots, m_i^X\}$ and $\{\tilde{Y}_i(T_{il}^Y) = Y_i(T_{il}^Y) + e_{il}^Y : l = 1, \ldots, m_i^Y\}$.

Before the two steps in Section 3.2, we performed the linear interpolation method by Kovac and Silverman (2000) to interpolate data onto a common and regular grid of $[0,1]$ with $m = 2^{J+1}$ for some integer J. When $m_i^X, m_i^Y \sim$ DiscreteUnif$\{50, \ldots, 70\}$, we chose $m = 64$; when $m_i^X, m_i^Y \sim$ DiscreteUnif$\{220, \ldots, 280\}$, we chose $m = 256$. The results are given in Table S1. Compared with Tables 1 - 3, wavHSIC now performs slightly worse in controlling the Type I error rate and achieving a high power, but it is overall satisfactory.

Table S1: Rejection rates of wavHSIC when subjects are not measured at the same dyadic grid with $m = 2^{J+1}$. Medians of selected β_X and β_Y are provided for each setting.

		$n = 50$				$n = 200$			
		$m \sim 50-70$		$m \sim 220-280$		$m \sim 50-70$		$m \sim 220-280$	
		SNR=4	SNR=8	SNR=4	SNR=8	SNR=4	SNR=8	SNR=4	SNR=8
Setting 1	Type I error rate	0.0553	0.0503	0.0955	0.0905	0.0452	0.0452	0.0402	0.0603
	median$\{\beta_X\}$	1.066	1.126	0.972	0.988	1.072	1.131	0.990	1.005
	median$\{\beta_Y\}$	0.791	0.866	0.723	0.745	0.807	0.871	0.742	0.764
Setting 2	Power	0.8291	0.9296	0.9648	0.9598	1.0000	1.0000	1.0000	1.0000
	median$\{\beta_X\}$	1.059	1.124	0.979	0.994	1.073	1.133	0.989	1.009
	median$\{\beta_Y\}$	0.788	0.860	0.723	0.745	0.801	0.870	0.738	0.758
Setting 3	Power	0.1859	0.2563	0.2814	0.2714	0.5578	0.7739	0.8291	0.8040
	median$\{\beta_X\}$	1.074	1.133	0.961	0.981	1.066	1.124	0.974	0.995
	median$\{\beta_Y\}$	0.830	0.906	0.777	0.794	0.830	0.896	0.769	0.789

S3.2 Simulation Settings in Lee et al. (2020, Supplementary Material)

In this section, we run an additional simulation study under the same settings in Lee et al. (2020, Supplementary Material, Section 1.2) to compare our method wavHSIC with PSS and FMDD. We also include KMSZ, KMSZ-p, dCov-c and FPCA here due to their competitive performances shown in Section 6. Here we use the same strategies for tuning parameters as in Section 6. For wavHSIC in following examples, we perform linear interpolation method by Kovac and Silverman (2000) to interpolate data onto a regular grid of $[0, 1]$ with $m = 2^{J+1} = 64$.

Example S1. *(Lee et al., 2020, Supplementary Material, Example 1) We generated functional response Y by a quadratic form of covariate X,*

$$Y_i(t) = c \cdot \left\{ X_i(t)^2 - 1 \right\} + \epsilon_i(t),$$

where X_i and ϵ_i, $i = 1, \ldots, n$ are independent Brownian motion and Brownian bridge, respectively. X is independent of Y when $c = 0$, while the alternative is satisfied when $c = 0.5$. Sampling points are $t = 1/200, 3/200, \ldots, 199/200$, with sample size $n = 40$ or 100. The results are given in Table S2.

Table S2 shows that KMSZ, PSS($Y \sim X$), FMDD($Y \sim X$) perform essentially the same as that in Lee et al. (2020, Supplementary Material, Table 1). Even the tests PSS($X \sim Y$) and FMDD($X \sim Y$) with the response and covariate switched can control type I error rates when $c = 0$, but when $c = 0.5$ their powers are much lower than that of PSS($Y \sim X$) and of FMDD($Y \sim X$) respectively. Two omnibus tests PSS(Omnibus) and FMDD(Omnibus) cannot control type I error probabilities when $c = 0$. For two distance covariance methods, dCov-c cannot control type I error rate well when $\alpha = 0.05, 0.01$, while FPCA has an accurate size for any combination of (α, n) when $c = 0$. When $c = 0.5$, the powers of these two methods are uniformly better than or comparable with PSS and FMDD. Our wavHSIC can almost always control the type I error rates when $c = 0$ and is uniformly more powerful than all the other methods for all (α, n) when $c = 0.5$.

Example S2. *(Lee et al., 2020, Supplementary Material, Example 2) We generate*

$$X_i(t) = \frac{4}{\pi} \sum_{k=1,3,\ldots,21} Z_{i,k} \sin(2\pi k t),$$
$$Y_i(t) = \frac{4}{\pi} \sum_{k=3,5,7,9} Z_{i,k}^2 \sin(2\pi k t) + 4\epsilon_i(t),$$

where $Z_{i,k}, k = 1, \ldots, 21$, $i = 1, \ldots, n$ are i.i.d. $N(0, 1)$ random variables and $\epsilon_i(t)$, $i = 1, \ldots, n$ are standard Brownian bridges on $[0, 1]$. Sampling points are $t = 1/200, 3/200, \ldots, 199/200$, with sample size $n = 40$ or 100. The results are given in Table S3.

Table S3 shows that KMSZ, PSS($Y \sim X$), FMDD($Y \sim X$) perform almost the same as those in Lee et al. (2020, Supplementary Material, Table 2). Permutation based KMSZ-p performs better than KMSZ when the sample size n is small, but for small nominal levels $\alpha = 0.05$ or 0.01, the powers of KMSZ-p are not as good as those of KMSZ for $n = 100$. Similar to Table S2, the tests PSS($X \sim Y$) and FMDD($X \sim Y$) have much lower powers than PSS($Y \sim X$) and FMDD($Y \sim X$) respectively. Between the two distance covariance based methods, FPCA performs better than dCov-c. FPCA performs better than other model-based methods for $n = 40$, and its powers lie between PSS and FMDD when $n = 100$. Our proposed method wavHSIC has uniformly higher powers than the other methods. Interestingly, the median of β_X are always 0 by our tuning parameter selection strategy, which indicates that the distance variances across low to high frequencies for $X(t)$ are successfully detected as equally distributed.

Table S2: Rejection rates of six test methods for Example S1. For wavHSIC, medians of selected β_X and β_Y are provided for each setting.

$c = 0$	$\alpha = 0.1$		$\alpha = 0.05$		$\alpha = 0.01$	
Type I error rate	$n = 40$	$n = 100$	$n = 40$	$n = 100$	$n = 40$	$n = 100$
dCov-c	0.1106	0.0955	0.0804	0.0653	0.0251	0.0553
FPCA	0.1055	0.0955	0.0452	0.0603	0.0101	0.0101
KMSZ	0.0352	0.0754	0.0101	0.0302	0.0000	0.0050
KMSZ-p	0.1156	0.0854	0.0553	0.0452	0.0000	0.0151
PSS($Y \sim X$)	0.1407	0.1005	0.0704	0.0402	0.0402	0.0050
PSS($X \sim Y$)	0.0854	0.1106	0.0302	0.0503	0.0050	0.0000
PSS(Omnibus)	0.2060	0.2010	0.1005	0.0905	0.0452	0.0050
FMDD($Y \sim X$)	0.1005	0.0905	0.0653	0.0653	0.0151	0.0151
FMDD($X \sim Y$)	0.1206	0.0704	0.0603	0.0452	0.0101	0.0151
FMDD(Omnibus)	0.1256	0.1005	0.0804	0.0653	0.0151	0.0201
wavHSIC	0.1055	0.0955	0.0352	0.0553	0.0050	0.0101
median$\{\beta_X\}$	0.957	0.958	0.957	0.958	0.957	0.958
median$\{\beta_Y\}$	1.624	1.629	1.624	1.629	1.624	1.629
$c = 0.5$	$\alpha = 0.1$		$\alpha = 0.05$		$\alpha = 0.01$	
Power	$n = 40$	$n = 100$	$n = 40$	$n = 100$	$n = 40$	$n = 100$
dCov-c	0.8141	1.0000	0.7286	0.9950	0.5477	0.9950
FPCA	0.9598	1.0000	0.8995	1.0000	0.5678	0.9899
KMSZ	0.2362	0.3015	0.1256	0.2261	0.0352	0.0754
KMSZ-p	0.3719	0.3367	0.2412	0.2714	0.1055	0.1055
PSS($Y \sim X$)	0.4925	1.0000	0.3417	1.0000	0.1608	0.9347
PSS($X \sim Y$)	0.1005	0.0955	0.0553	0.0452	0.0101	0.0201
PSS(Omnibus)	0.5327	1.0000	0.3719	1.0000	0.1709	0.9347
FMDD($Y \sim X$)	0.6734	1.0000	0.3970	0.9799	0.0704	0.6332
FMDD($X \sim Y$)	0.1709	0.0955	0.0854	0.0553	0.0201	0.0101
FMDD(Omnibus)	0.6734	1.0000	0.3970	0.9799	0.0704	0.6332
wavHSIC	1.0000	1.0000	1.0000	1.0000	0.8492	1.0000
median$\{\beta_X\}$	0.957	0.958	0.957	0.958	0.957	0.958
median$\{\beta_Y\}$	1.278	1.266	1.278	1.266	1.278	1.266

Table S3: Rejection rates of six test methods for Example S1. For wavHSIC, medians of selected β_X and β_Y are provided for each setting.

	$\alpha = 0.1$		$\alpha = 0.05$		$\alpha = 0.01$	
Power	$n = 40$	$n = 100$	$n = 40$	$n = 100$	$n = 40$	$n = 100$
dCov-c	0.3266	0.7035	0.2362	0.5980	0.1407	0.3719
FPCA	0.5779	0.9045	0.4573	0.8392	0.2864	0.6281
KMSZ	0.1910	0.0804	0.1910	0.3317	0.0955	0.2211
KMSZ-p	0.3568	0.3668	0.2563	0.2764	0.1055	0.1307
PSS($Y \sim X$)	0.0352	0.5477	0.1256	0.6734	0.1910	0.7487
PSS($X \sim Y$)	0.0151	0.0050	0.0603	0.0503	0.1156	0.1005
PSS(Omnibus)	0.0503	0.5528	0.1759	0.6834	0.2714	0.7638
FMDD($Y \sim X$)	0.5528	0.9950	0.3568	0.9598	0.1055	0.5980
FMDD($X \sim Y$)	0.1709	0.1357	0.1005	0.0905	0.0151	0.0201
FMDD(Omnibus)	0.5528	0.9950	0.3568	0.9598	0.1055	0.5980
wavHSIC	0.9598	1.0000	0.8844	1.0000	0.6131	0.9950
median$\{\beta_X\}$	0.000	0.000	0.000	0.000	0.000	0.000
median$\{\beta_Y\}$	0.618	0.579	0.618	0.579	0.618	0.579

References

Berg, C., J. P. R. Christensen, and P. Ressel (1984). *Harmonic analysis on semigroups: theory of positive definite and related functions*, Volume 100. Springer.

DeVore, R. A. and G. G. Lorentz (1993). *Constructive approximation*, Volume 303. Springer-Verlag Berlin Heidelberg.

Johnstone, I. M. (2019). Gaussian estimation: sequence and wavelet models. `statweb.stanford.edu/~imj/GE_09_16_19.pdf`.

Kovac, A. and B. W. Silverman (2000). Extending the scope of wavelet regression methods by coefficient-dependent thresholding. *Journal of the American Statistical Association 95*(449), 172–183.

Lee, C., X. Zhang, and X. Shao (2020). Testing conditional mean independence for functional data. *Biometrika 107*(2), 331–346.

Linde, W. (1986). On Rudin's equimeasurability theorem for infinite dimensional Hilbert spaces. *Indiana University Mathematics Journal 35*(2), 235–243.

Lyons, R. (2013). Distance covariance in metric spaces. *Annals of Probability 41*(5), 3284–3305.

Rindt, D., D. Sejdinovic, and D. Steinsaltz (2020). Consistency of permutation tests for HSIC and dHSIC. *arXiv preprint arXiv:2005.06573*.

Sejdinovic, D., B. Sriperumbudur, A. Gretton, and K. Fukumizu (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics 41*(5), 2263–2291.

Wells, J. H. and L. R. Williams (2012). *Embeddings and extensions in analysis*, Volume 84. Springer Science & Business Media.