Intrinsic Sliced Wasserstein Distances for Comparing Collections of Probability Distributions on Manifolds and Graphs

Anonymous Authors[1]

Abstract

Collections of probability distributions arise in a variety of applications ranging from user activity pattern analysis to brain connectomics. In practice these distributions can be defined over diverse domain types including finite intervals, circles, cylinders, spheres, other manifolds, and graphs. This paper introduces an approach for detecting differences between two collections of distributions over such general domains. To this end, we propose the intrinsic slicing construction that yields a novel class of Wasserstein distances on manifolds and graphs. These distances are Hilbert embeddable, allowing us to reduce the distribution collection comparison problem to a more familiar mean testing problem in a Hilbert space. We provide two testing procedures one based on resampling and another on combining p-values from coordinate-wise tests. Our experiments in various synthetic and real data settings show that the resulting tests are powerful and the p-values are well-calibrated.

1. Introduction

Distributional data defined over general domains such as manifolds and graphs arise in a variety of statistical applications. In this paper we consider the problem of comparing two collections of distributions over such a general domain. Our goal is to test for homogeneity—whether all of the distributions come from the same *meta-distribution*—in an interpretable manner. While conceptually similar to two-sample testing, this is a higher order notion in the sense that our units of analysis are distributions/histograms.

For instance, given collections of personal activity histograms (over cylinder: time of day $\times$ intensity) for two sub-populations, one may be interested in determinining whether there are statistically significant differences between activity patterns of these sub-populations. As another example, consider normalized counts of events per region on a daily basis. Collected over a year, this gives a set of 365 daily probability distributions over the region adjacency graph, and one may wish to compare the collection of distributions from weekdays to those from weekends. As typical with two-sample tests, testing specific aspects of homogeneity is preferable: for example, in regular two-sample testing, detecting that the means are unequal provides interpretable insights, whereas a general test that only says there are unspecified differences between the distributions is less useful for interpretation.

Limited settings of this problem tackling distributions over the interval/circle have been considered in the literature (Dubey & Müller, 2019), yet the general case of distributions over graphs and manifolds is open. The requirement to test for specific differences is non-trivial on general domains: what is the equivalent of mean for a collection of distributions? While Fréchet mean (see e.g. (Peyré & Cuturi, 2019)) may seem like the natural choice, there are a number of problems with testing Fréchet mean equality. First, the existence and uniqueness of the Fréchet mean is not guaranteed, and it can be sensitive to small changes. Second, computing the Fréchet mean is expensive and can become prohibitive when resampling is used to compute the null distribution. Finally, resampling poses conceptual problems: using permutation null will detect differences beyond the equality of Fréchet means (same problem exists in regular two-sample testing, see e.g. (Huang et al., 2006)), and using bootstrap requires designing the null case, which is highly non-trivial.

We attack this problem using insights from recent developments that utilize *Hilbert embeddings* for simplifying distributional data problems (see e.g. (Solomon et al., 2014; Petersen & Müller, 2016)). The simplification comes as a result of linearity of Hilbert spaces, which allows adapting existing statistical approaches to distributional data. A crucial requirement on the embedding is that the distance in the embedding space should give a meaningful distance between measures; it is this property that renders quantities computed in the embedding space such as means and variances meaningful. While transportation based distances are efficient at capturing many aspects of distributional data such as horizontal variation (Panaretos & Zemel, 2019; Peyré & Cuturi, 2019; Bigot, 2020), yet the transportation theoretic approaches hit a roadblock beyond the real line case due to

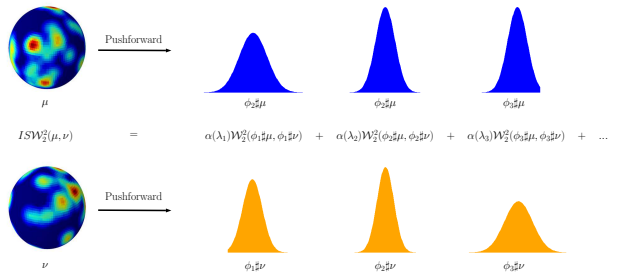


Figure 1. Schematic of the proposed intrinsic slicing construction. Given two probability measures on the sphere (here the darkest blue corresponds to zero mass), different aspects of their dissimilarities become apparent after pushforward to the real line using the eigenfunctions of the Laplace-Beltrami operator, $\{\phi_i\}$, in this case spherical harmonics. As a particular example of our general construction, the (squared) intrinsic sliced 2-Wasserstein distance $ISW_2^2(\cdot,\cdot)$ is the weighted sum of the dissimilarities of the corresponding pushforwards of μ and ν as measured by squared 2-Wasserstein distance $\mathcal{W}_2^2(\cdot,\cdot)$ *on the real line*.

their Hilbert non-embeddability (Peyré & Cuturi, 2019).

We overcome these difficulties by introducing a new slicing construction on manifolds and graphs (Figure 1) inspired by the sliced 2-Wasserstein ($\mathcal{W}_2$) distances in high dimensional spaces (Kolouri et al., 2019b;a). Our construction leverages eigenvalues and eigenfunctions of the Laplace-Beltrami operator on manifolds and graph Laplacians to capture the intrinsic geometry and connectivity of the data domain. We apply this slicing construction to obtain a novel class of intrinsic sliced 2-Wasserstein distances on manifolds and graphs. The resulting distances are Hilbert embeddable, have a number of desirable properties, and can be truncated to obtain finite-dimensional embeddings.

Using the corresponding embedding allows us to reduce the distribution collection comparison problem to the comparison of means in a high-dimensional Euclidean space. At the theoretical level our test checks equality of Fréchet means along slicings (see discussion around Example 1). These means are transparently tied to the input data, whereby rejections lead to interpretable insights. We provide two approaches for hypothesis testing and verify via extensive experiments that these tests are powerful, and the p-values are well-calibrated.

2. Related Work

Our framework is not simply a higher order version of a two-sample kernel test (Gretton et al., 2012) since we test for equality of a specific aspect of meta-distributions. This renders our null hypothesis different from Gretton et al. (2012), and we need a different set of techniques both for proofs and computations. For example, testing in (Gretton et al., 2012) can use the permutation null which is valid due to the stronger null hypothesis of equal distributions. In contrast, with our null hypothesis we cannot use the permutation null and have to resort to a bootstrap procedure. Other approaches such as the general Hilbert embedding framework of (Petersen & Müller, 2016) is not tied to a distance between probability distributions and so can be problematic for capturing the location and variability aspects of distribution collections. In addition, (Petersen & Müller, 2016) has difficulties in higher dimensions and does not provide constructions suitable for manifolds or graphs.

The Sliced Wasserstein (SW) distance and its generalized variant (Kolouri et al., 2019a, GSW) sets up the idea of approximating Wasserstein distances using multiple nonlinear projections, it is presented in extrinsic terms (i.e. Euclidean space) and can suffer from the curse of dimensionality when a low dimensional data manifold lives in a high-dimensional space. Our choice of eigenfunctions for projection is very different from the one-parameter function families in GSW and allows us to rigorously prove a number of general and testing-specific properties. Moreover, the GSW construction does not directly apply to graphs. While the tree-sliced variant of GSW (Le et al., 2019) can be applied in an intrinsic manner (the clustering variant), it relies on a different type of distance, in the limit related to the euclidean/geodesic distance. This can be seen by comparing our lower bound to theirs: our lower bound for ISW is in terms of the MMD using the spectral distance (Proposition 4). The recently-proposed Sobolev transport (Le et al., 2022) does consider measures supported on graphs, but in absence of a slicing construction it is limited to testing for unspecific differences and requires extensive compute due to much slower permutation based calibration (see Section 6). Deshpande et al. (2019) proposed Max Sliced Wasserstein (MSW) distance—taking maximum over projected $\mathcal{W}_2$ distances vs the average $\mathcal{W}_2$ distances of SW—in the context of generative models. Finally, the robust sliced Wasserstein distance of Lai & Zhao (2017) does make use of the geometric properties of the underlying manifold. However, their goal is to compute a correspondence between two manifolds by mapping them into $\mathbb{R}^d$ using eigenmaps and treating the mapped manifolds as measures in $\mathbb{R}^d$ and minimizing some version of Euclidean slicing. *None of these works consider the problem of comparing collections of distributions, provide ways for obtaining calibrated p-values, or prove properties of the distances that make them desirable for hypothesis testing.*

While statistical manifold theory (Murray & Rice, 1993, SMT) provides a powerful framework for studying parametric families of probability distributions, we focus on more general spaces of probability measures, such as the space of all Borel probability measures on a manifold or graph. The Fisher information metric is defined in terms of a parametric family of probability distributions and may not be easily extended to such nonparametric spaces of measures. In the language of SMT, assume that we observe (each

observation is a distribution, e.g. a histogram) two collections of parametric distributions: $F(x, \beta_k), \beta_k \sim G(\theta)$, and $F(x, \gamma_k), \gamma_k \sim H(\gamma)$. Thus, each collection is generated from a distribution of its parameters; our test is interested in finding out whether these distributions G and H are the same, and is easily extendable to nonparametric settings.

3. Preliminaries

Given a compact metric space $\mathcal{X}$, let $\mathcal{P}(\mathcal{X})$ denote the set of Borel probability measures on $\mathcal{X}$. Our main interest is in the case where $\mathcal{X}$ is a graph or a manifold with the shortest/geodesic distance as the metric, and thus the compactness restriction. The 2-Wasserstein distance can be defined on $\mathcal{P}(\mathcal{X})$ using the metric of $\mathcal{X}$ as the ground distance (Peyré & Cuturi, 2019; Panaretos & Zemel, 2019), giving $\mathcal{W}_2^{\mathcal{X}} : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}_{\geq 0}$; due to the repeated use of the real line case we use the shorthand $\mathcal{W}_2 = \mathcal{W}_2^{\mathbb{R}}$. Central to our study are distributions on the space of probability measures $\mathcal{P}(\mathcal{P}(\mathcal{X})) = (\mathcal{P}(\mathcal{X}), \mathcal{B}(\mathcal{P}(\mathcal{X})))$, where $\mathcal{B}(\mathcal{P}(X))$ is the Borel σ-algebra generated by the topology induced by $\mathcal{W}_2^{\mathcal{X}}$ (Bigot, 2020). To avoid confusion, we will refer to the elements of $\mathcal{P}(\mathcal{P}(\mathcal{X}))$ as *meta-distributions*.

Let $P, Q \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$, and assume that we are given two collections of probability measures $\{\mu_i\}_{i=1}^{N_1}$ and $\{\nu_i\}_{i=1}^{N_2}$ that are drawn from P and Q: $\mu_i \sim P$ and $\nu_i \sim Q$ in an independent-and-identically-distributed (hereafter i.i.d.) manner. Our goal is to use this sample to test the null hypothesis of whether $P = Q$. While this is conceptually a two-sample test, note that our data points are distributions; in practice, the distributions μ_i or ν_i are given by histograms.

Remark 1. Let us compare this with the usual two-sample testing. Consider $P \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$ constructed as follows. Let $\mu^* \in \mathcal{P}(\mathcal{X})$ be a fixed probability measure. Let $x_1, x_2, ...x_A \sim \mu^*$ and construct the histogram summarizing this sample: $\frac{1}{A}\sum_{a=1}^{A} \delta_{x_a}$. Now, $\frac{1}{A}\sum_{a=1}^{A} \delta_{x_a} \in \mathcal{P}(\mathcal{X})$ is one sample drawn from P. In our testing scenario one gets the collection $\{\mu_i\}_{i=1}^{N_1}$, where each histogram is obtained as above: $\mu_i \sim P$. Similarly, consider $Q \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$ of the same type based on some other fixed $\nu^* \in \mathcal{P}(\mathcal{X})$, and let $\{\nu_i\}_{i=1}^{N_2}$ the corresponding collection of histograms. Testing whether $P = Q$ in the limit boils down to $\mu^* = \nu^*$. When compared to the usual two-sample testing this may seem rather inefficient, requiring A times more samples (resp. $N_1 A$ and $N_2 A$ samples from μ^* and ν^*). However, in our setup it is *not assumed* that the histograms in the collections come from meta-distributions of the above simple type (i.e. all μ_i are generated by drawing from the same underlying distribution μ^*). In fact, the target use-case for our approach is when these histograms are collected by observing different individuals who have their *person-specific* behaviors/distributions. □

Let $\mathcal{D}(\cdot,\cdot) : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}_{\geq 0}$ be a distance between probability distributions. $\mathcal{D}(\cdot,\cdot)$ is called *Hilbertian* (this is just a naming convention; no implication that the map is a Hilbert map) if there exist a Hilbert space $\mathcal{H}$ and a map $\eta : \mathcal{P}(\mathcal{X}) \to \mathcal{H}$ such that $\mathcal{D}(\mu, \nu) = \|\eta(\mu) - \eta(\nu)\|_{\mathcal{H}}$. For example, it is well-known that 2-Wasserstein distance on $\mathcal{X} = \mathbb{R}$ is Hilbertian (Peyré & Cuturi, 2019) (also see Section 4.2) and Maximum Mean Discrepancy (MMD) on any $\mathcal{X}$ is Hilbertian (Gretton et al., 2012); however, the 2-Wasserstein distance $\mathcal{W}_2^{\mathcal{X}}$ on general $\mathcal{X}$ is not Hilbertian (Peyré & Cuturi, 2019).

Since the map η takes every measure on $\mathcal{X}$ to a point in $\mathcal{H}$, we see that a meta-distribution $P \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$ gives a rise to a measure on $\mathcal{H}$ given by pushforward operation, $\eta\#P = P \circ \eta^{-1} \in \mathcal{P}(\mathcal{H})$. In addition, if a finite dimensional approximation $\eta_D : \mathcal{P}(\mathcal{X}) \to \mathbb{R}^D$ of η is available, then $\eta_D\#P$ is a measure on $\mathbb{R}^D$. This observation is enormously useful: problems about the elements of the rather abstract space $\mathcal{P}(\mathcal{P}(\mathcal{X}))$ are reduced to problems about familiar measures on $\mathcal{H}$ or even $\mathbb{R}^D$. For example, the usual notions of mean and variance can be applied to the measure $\eta\#P$ to gain insights about the meta-distribution P. The validity of these insights hinges on the η-map coming from a Hilbertian distance, as distances are central to the statistical quantities of interest.

Testing for $\eta\#P = \eta\#Q$ can serve as a proxy for our original testing problem of $P = Q$. As typical with two-sample tests, various aspects of the equality $\eta\#P = \eta\#Q$ can be tested, such as the mean or variance equality; unspecific tests of equality can be applied as well. We will concentrate on testing certain aspects of the equality so that one can easily drill down on the results. This is similar to the regular two-sample testing where checking for equality of, say, means is often preferable as it gives immediately interpretable insights, whereas a general test that only says there are unspecified differences between the distributions is less useful for interpretation. To obtain succint and interpretable tests we concentrate on the mean of the resulting pushforward measure in $\mathcal{H}$.

Definition 1. For a meta-distribution $P \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$, we define its *Hilbert centroid* with respect to the Hilbertian distance $\mathcal{D}$ as $C_{\eta\#P} = \mathbb{E}_{\mu\sim P}[\eta(\mu)] \in \mathcal{H}$, assuming it exists.

Our testing procedure is based on checking the equality $C_{\eta\#P} = C_{\eta\#Q}$, or more explicitly: $\mathbb{E}_{\mu\sim P}[\eta(\mu)] = \mathbb{E}_{\nu\sim Q}[\eta(\nu)]$. Intuitively, each "dimension" of the map η probes some aspect of the two involved meta-distributions and makes sure that they are in agreement in expectation. One of our testing approaches will use the statistic

$$\mathbb{T}(P, Q) = \|C_{\eta\#P} - C_{\eta\#Q}\|_{\mathcal{H}}^2. \tag{3.1}$$

to capture the deviations from equality; this quantity can be

written directly in terms of pairwise distances.

Proposition 1. *For $P, Q \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$, the following holds:*

$$\mathbb{T}(P,Q) = \mathbb{E}_{\mu\sim P,\nu\sim Q}[\mathcal{D}^2(\mu,\nu)] - \frac{1}{2}\mathbb{E}_{\mu,\mu'\sim P}[\mathcal{D}^2(\mu,\mu')] - \frac{1}{2}\mathbb{E}_{\nu,\nu'\sim Q}[\mathcal{D}^2(\nu,\nu')].$$

Next we give an example of what Hilbert centroid equality means in an important special case.

Example 1. Let $\mathcal{X} = [0,T] \subset \mathbb{R}$ with $\mathcal{D}$ being the 2-Wasserstein distance $\mathcal{W}_2$. Given a probability measure $\mu \in \mathcal{P}([0,T])$, let F_μ be its cumulative distribution function: $F_\mu(x) = \mu([0,x]) = \int_0^x d\mu$. The generalized inverse of cumulative distribution function (CDF) is defined by $F_\mu^{-1}(s) := \inf\{x \in [0,T] : F_\mu(x) > s\}$. The squared 2-Wasserstein distance has a rather simple expression in terms of the inverse CDF (Peyré & Cuturi, 2019):

$$\mathcal{W}_2^2(\mu,\nu) = \int_0^1 (F_\mu^{-1}(s) - F_\nu^{-1}(s))^2 ds. \tag{3.2}$$

This formula immediately establishes the Hilbertianity of $\mathcal{W}_2$ through the map $\eta : \mathcal{P}([0,T]) \to L_2([0,T])$ defined by $\eta(\mu) = F_\mu^{-1}$. Note that η is invertible for increasing normalized functions in the embedding space. Using this insight, we see that the corresponding "average measure" of $P \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$ can be introduced via $P_{\text{av}} = \eta^{-1}(\mathbb{E}_{\mu\sim P}[\eta(\mu)])$. It is easy to prove that P_{av} satisfies the following: $P_{\text{av}} = \arg\min_{\rho\in\mathcal{P}(\mathcal{X})} \mathbb{E}_{\mu\sim P}[\mathcal{W}_2(\mu,\rho)^2]$, which is the definition of the Fréchet mean, see for example (Peyré & Cuturi, 2019). *In this setting, $C_{\eta\#P} = C_{\eta\#Q}$ boils down to having the same Fréchet means, $P_{\text{av}} = Q_{\text{av}}$.* □

We will later see that the Hilbert embedding corresponding to the intrinsic sliced 2-Wasserstein distance is assembled of embeddings like in Example 1 applied after pushforwards (see Figure 1 for an intuition). This means that the resulting equality $C_{\eta\#P} = C_{\eta\#Q}$ becomes more stringent, making it a better proxy for detecting the deviations from $P = Q$ without losing the interpretability aspect.

4. Intrinsic Sliced 2-Wasserstein Distance

We introduce a Hilbertian version of $\mathcal{W}_2$ on manifolds and graphs via a construction we call *intrinsic slicing* due to its use of the domain's intrinsic geometric properties. To focus our discussion we concentrate on the manifold case, as the graph case is simpler and is obtained by replacing the Laplace-Beltrami operator by the graph Laplacian.

Let $\lambda_\ell, \phi_\ell; \ell = 0, 1, \ldots.$ be the eigenvalues and eigenfunctions of the Laplace-Beltrami operator on $\mathcal{X}$ with Neumann boundary conditions. The eigenfunctions are sorted by increasing eignevalue and assumed to be orthonormal with respect to some fixed (e.g. uniform) measure on $\mathcal{X}$; also $\phi_0 = const$ and $\lambda_0 = 0$. One can define the spectral kernel $k(x,y) = \sum_\ell \alpha(\lambda_\ell)\phi_\ell(x)\phi_\ell(y)$ and the corresponding spectral distance on the manifold $d(x,y) = k(x,x) + k(y,y) - 2k(x,y) = \sum \alpha(\lambda_\ell)(\phi_\ell(x) - \phi_\ell(y))^2$, where $\alpha : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a function that controls contribution from each spectral band. By setting $\alpha(\lambda) = e^{-t\lambda}$ for some $t > 0$, we get the heat/diffusion kernel and the corresponding diffusion distance (Coifman & Lafon, 2006). Another important case is $\alpha(\lambda) = 1/\lambda^2$ if $\lambda > 0$ and $\alpha(0) = 0$, which gives the biharmonic kernel and distance (Lipman et al., 2010). In both of these constructions $\alpha(\cdot)$ is a decreasing function, allowing the smoother low-frequency (i.e. smaller λ_ℓ) eigenfunctions to contribute more.

4.1. Definition and properties

A real-valued function $\phi : \mathcal{X} \to \mathbb{R}$ can be used to map the manifold $\mathcal{X}$ onto the real line. Any probability measure $\mu \in \mathcal{P}(\mathcal{X})$ can likewise be projected onto the real line using the pushforward of ϕ, which we denote by $\phi\sharp\mu = \mu \circ \phi^{-1} \in \mathcal{P}(\mathbb{R})$. While the pushforward notions used here and in previous sections are conceptually the same, for clarity we use $\sharp$ for measures and $\#$ for meta-distributions. We define intrinsic slicing as follows.

Definition 2. Given a function $\alpha : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ and a probability distance $\mathcal{D}(\cdot,\cdot)$ on $\mathcal{P}(\mathbb{R})$, we define the intrinsic sliced distance $IS\mathcal{D}(\cdot,\cdot)$ on $\mathcal{P}(\mathcal{X})$ by

$$IS\mathcal{D}^2(\mu,\nu) = \sum_\ell \alpha(\lambda_\ell)\mathcal{D}^2(\phi_\ell\sharp\mu, \phi_\ell\sharp\nu).$$

The choice of the Laplacian eigenfunctions in the definition can be justified by a number of their properties. Eigenfunctions are intrinsic quantities of a manifold and are ordered by smoothness. Thus, they allow capturing the intrinsic connectivity of the underlying domain. Furthermore, due to the orthogonality of eigenfunctions, their pushforwards can capture complementary aspects of the distribution.

While the definition is general, our focus in this paper is on the case when $\mathcal{D} = \mathcal{W}_2$; we remind that we always use $\mathcal{W}_2$ to denote the 2-Wasserstein distance on $\mathcal{P}(\mathbb{R})$. We call the resulting distance *Intrinsic Sliced 2-Wasserstein Distance*, and denote it by $IS\mathcal{W}_2$. First, we discuss the convergence of the infinite sum in Definition 2.

Proposition 2. *If $\mathcal{X}$ is a smooth compact n-dimensional manifold and $\sum_\ell \lambda_\ell^{(n-1)/2}\alpha(\lambda_\ell) < \infty$, then $IS\mathcal{W}_2$ is well-defined.*

Next, we prove a number of properties of $IS\mathcal{D}$.

Proposition 3. *If $\mathcal{D}$ is a Hilbertian probability distance such that $IS\mathcal{D}$ is well-defined, then (i) $IS\mathcal{D}$ is Hilbertian, and (ii) $IS\mathcal{D}$ satisfies the following metric properties: non-negativity, symmetry, the triangle inequality, and $IS\mathcal{D}(\mu,\mu) = 0$.*

Proof. By Hilbertian property of $\mathcal{D}$, there exists a Hilbert space $\mathcal{H}^0$ and a map $\eta^0 : \mathcal{P}(\mathbb{R}) \to \mathcal{H}^0$ such that $\mathcal{D}(\rho_1, \rho_2) = \|\eta^0(\rho_1) - \eta^0(\rho_2)\|_{\mathcal{H}'}$ for all $\rho_1, \rho_2 \in \mathcal{P}(\mathbb{R})$. Plugging this into Definition 2 we have $IS\mathcal{D}(\mu, \nu) = \|\eta(\mu) - \eta(\nu)\|_{\mathcal{H}}$, where $\mathcal{H} = \oplus_\ell \mathcal{H}^0$ and the ℓ-th component of $\eta(\mu)$ is $\sqrt{\alpha(\lambda_\ell)}\eta_0(\phi_\ell \sharp \mu) \in \mathcal{H}$. The second part of Proposition 3 directly follows from the Hilbert property. □

Since $\mathcal{W}_2$ is Hilbertian on $\mathcal{P}(\mathbb{R})$, the application of Proposition 3 yields that $IS\mathcal{W}_2$ is also Hilberitan, making it possible to use $IS\mathcal{W}_2$ for our hypothesis tests in Section 5.

For a simple choice of distance $\mathcal{D}$ on $\mathcal{P}(\mathbb{R})$, namely the absolute mean difference, the corresponding intrinsic sliced distance is the well-known MMD (Gretton et al., 2012).

Proposition 4. *Let $\mathcal{D}(\rho_1, \rho_2) = |\mathbb{E}_{x\sim\rho_1}[x] - \mathbb{E}_{y\sim\rho_2}[y]|$ for $\rho_1, \rho_2 \in \mathcal{P}(\mathbb{R})$, then the corresponding $IS\mathcal{D}$ is equivalent to the MMD with the spectral kernel $k(\cdot, \cdot)$.*

When $k(x, y)$ is the heat kernel, the sliced distance in Proposition 4 is very much like the MMD with the Gaussian kernel, with the parameter t in $\alpha(\lambda) = e^{-t\lambda}$ controlling the kernel width. Indeed, the two kernels coincide on $\mathbb{R}^d$, and on general manifolds Varadhan's formula gives asymptotic equivalence for small t (Berger, 2003).

An interesting insight derived from the above result is that $IS\mathcal{W}_2$ is in a sense a "stronger" distance than MMD that uses the corresponding spectral kernel. The $IS\mathcal{W}_2$ compares the quantiles of the pushforward distributions (Eq. (3.2)), whereas MMD compares their expectations only. We formalize this notion in the next result, also providing a theoretical reason for preferring $IS\mathcal{W}_2$ for hypothesis testing.

Proposition 5. *$MMD(\mu, \nu) \leq IS\mathcal{W}_2(\mu, \nu)$ when the same $\alpha(\cdot)$ is used in both constructions.*

We are now in a position to prove that $IS\mathcal{W}_2$ is a true metric.

Theorem 1. *If $\alpha(\lambda) > 0$ for all $\lambda > 0$, then $IS\mathcal{W}_2$ is a metric on $\mathcal{P}(\mathcal{X})$.*

We remind that 2-Wasserstein distance can be defined directly on $\mathcal{P}(\mathcal{X})$ using the geodesic distance as the ground metric; we denote this distance as $\mathcal{W}_2^{\mathcal{X}}$. Lipschitz properties of the eigenfunctions imply the following:

Proposition 6. *There exists a constant c depending only on $\mathcal{X} \subseteq \mathbb{R}^n$ such that for all $\mu, \nu \in \mathcal{P}(\mathcal{X})$ the inequality $IS\mathcal{W}_2(\mu, \nu) \leq c\mathcal{W}_2^{\mathcal{X}}(\mu, \nu)\sqrt{\sum_\ell \lambda_\ell^{(n+3)/2}\alpha(\lambda_\ell)}$ holds.*

Our final result looks at the quantity $\mathbb{T}$ defined using $IS\mathcal{W}_2$ by Eq. (3.1). We will be using $\mathbb{T}$ computed on finite collections of measures as a test statistic in the next section. We show that it enjoys robustness with respect to small perturbations of the measures in the collection.

Proposition 7. *Let $\{\mu_i\}_{i=1}^N$ and $\{\nu_i\}_{i=1}^N$ be two collections of probability measures on $\mathcal{P}(\mathcal{X})$, such that $\forall i, \mathcal{W}_2^{\mathcal{X}}(\mu_i, \nu_i) \leq \epsilon$, then $\mathbb{T}(\{\mu_i\}_{i=1}^N, \{\nu_i\}_{i=1}^N) \leq C^2\epsilon^2$. Here $C = c\sqrt{\sum_\ell \lambda_\ell^{(n+3)/2}\alpha(\lambda_\ell)}$ from previous proposition and is assumed to be finite.*

This bound implies that if the distributions in a collection undergo horizontal shifts that are small as measured by the geodesic distance $\mathcal{W}_2^{\mathcal{X}}$, then $\mathbb{T}$ is small as well.

4.2. Approximate Hilbert Embedding

An important aspect of $IS\mathcal{W}_2$ is that its Hilbert map $\eta : \mathcal{P}(\mathcal{X}) \to \mathcal{H}$ can be approximated by a finite-dimensional embedding $\eta_D : \mathcal{P}(\mathcal{X}) \to \mathbb{R}^D$ such that $IS\mathcal{W}_2(\mu, \nu) \approx \|\eta_D(\mu) - \eta_D(\nu)\|_{\mathbb{R}^D}$. This is useful for practical computation and for one of our hypothesis testing approaches.

Using the formula for $IS\mathcal{W}_2$ on $\mathcal{P}(\mathbb{R})$ in terms of the quantile function, Eq. (3.2), the Hilbert map is defined by $\eta^0(\mu) = F_\mu^{-1}$. We have $\mathcal{W}_2(\mu, \nu) = \|\eta^0(\mu) - \eta^0(\nu)\|_{L_2(\mathbb{R})}$, where the norm involves integration. We can discretize the integral using the Riemann sum for equidistant knots $s_k = \frac{k-1}{D'}, k = 1, ..., D'$, define the approximate embedding $\eta_{D'}^0 : \mathcal{P}(\mathbb{R}) \to \mathbb{R}^{D'}$ as:

$$\eta_{D'}^0 : \mu \to \frac{1}{\sqrt{D'}}[F_\mu^{-1}(s_1), ..., F_\mu^{-1}(s_{D'})]. \tag{4.1}$$

Now, $\mathcal{W}_2(\mu, \nu) \approx \|\eta_{D'}^0(\mu) - \eta_{D'}^0(\nu)\|_{\mathbb{R}^{D'}}$ with approximation quality depending on the embedding dimension D'.

To approximate the Hilbert map for $IS\mathcal{W}_2$ we truncate the series defining $IS\mathcal{W}_2$ and use a finite number of eigenfunctions for pushforward: $\phi_\ell, \ell = 1, ..., L$, where ϕ_0 is dropped since it is a constant. By inspecting the proof of Proposition 3 and using Eq. (4.1), we can define $\eta_D : \mathcal{P}(\mathcal{X}) \to \mathbb{R}^D$ with $D = LD'$ as the concatenation of L maps:

$$(\eta_D)_\ell : \mu \to \sqrt{\frac{\alpha(\lambda_\ell)}{D'}}[F_{\phi_\ell \sharp \mu}^{-1}(s_1), ..., F_{\phi_\ell \sharp \mu}^{-1}(s_{D'})].$$

Spectral decompositions of Laplace-Beltrami operators for general manifolds (Coifman & Lafon, 2006; Reuter et al., 2009) or graph Laplacians can be computed numerically. For applications involving simple manifolds, eigenvalues and eigenfunctions can be computed analytically (see Appendix A.3).

The hyperparameters L and D' capture distinct aspects of the complexity of the distribution. A larger L allows access to higher order eigenfunctions that possess greater spatial oscillations, linking to the finer details of the geometry and topology of the underlying domain and distribution collection. On the other hand, D' captures how well the pushforward distribution is represented, which is determined by the number of quantiles. Both hyperparameters can have a significant impact on the success of the testing process.

5. Hypothesis Testing

Let $\{\mu_i\}_{i=1}^{N_1}$ and $\{\nu_i\}_{i=1}^{N_2}$ be two i.i.d. collections of measures drawn from $P, Q \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$ respectively. Our goal is to use these samples to test the null hypothesis $H_0 : C_{\eta\#P} = C_{\eta\#Q}$, where η is the Hilbert embedding of the sliced distance ISW_2 on $\mathcal{P}(\mathcal{X})$.

5.1. Resampling Based Test

We use the quantity $\mathbb{T}(\cdot,\cdot)$ from Eq. (3.1) as the test statistic. Its sample version is computed by replacing the expectations by the empirical means, and excluding the diagonal terms to achieve unbiasedness

$$\hat{\mathbb{T}} \equiv \sum_{i,j:i\neq j} \frac{ISW_2^2(\mu_i,\mu_j)}{2N_1(N_1-1)} + \sum_{i,j:i\neq j} \frac{ISW_2^2(\nu_i,\nu_j)}{2N_2(N_2-1)} - \sum_{i,j} \frac{ISW_2^2(\mu_i,\nu_j)}{N_1N_2}. \quad (5.1)$$

Note that $\mathbb{E}\hat{\mathbb{T}} = \mathbb{T}(P,Q)$. In practice, the ISW_2 values are computed from the approximate embedding: $ISW_2(\rho_1,\rho_2) \approx \|\eta_D(\rho_1) - \eta_D(\rho_2)\|_{\mathbb{R}^D}$. We denote the resulting statistic by $\tilde{\mathbb{T}}_{L,D'}$.

The difference between $\tilde{\mathbb{T}}_{L,D'}$ and the population version (i.e. $\mathbb{T} - \tilde{\mathbb{T}}_{L,D'}$) can be decomposed as $(\mathbb{T} - \hat{\mathbb{T}}) + (\hat{\mathbb{T}} - \hat{\mathbb{T}}_L) + (\hat{\mathbb{T}}_L - \tilde{\mathbb{T}}_{L,D'})$, where the summands inside the terms $\hat{\mathbb{T}}_L$ and $\tilde{\mathbb{T}}_{L,D'}$ correspond to partial sums that approximate $ISW_2^2(\cdot,\cdot)$ by $\sum_{l=1}^{L}\alpha(\lambda_l)\mathcal{W}_2^2(\phi_l\sharp\cdot,\phi_l\sharp\cdot)$, and $\mathcal{W}_2^2(\phi_l\sharp\cdot,\phi_l\sharp\cdot)$ by $\|\eta_{D'}(\phi_l\sharp\cdot) - \eta_{D'}(\phi_l\sharp\cdot)\|^2$, respectively. We show in Appendix A.4 that a) summands in the second and third terms in the sum can be made infinitesmally small by choosing large enough L and D', respectively; b) an asymptotic result for the first difference can be obtained by extending the tools from (Gretton et al., 2012; Serfling, 2009). These results are based on several assumptions detailed in Appendix A.4. Combining the two results, we establish asymptotic distributions of $\tilde{\mathbb{T}}_{L,D'}$:

Theorem 2. *Assume relevant conditions (see Appendix A.4) hold. Define $N = N_1 + N_2$, and suppose that as $N_1, N_2 \to \infty$, we have $N_1/N \to \rho_1, N_2/N \to \rho_2 = 1-\rho_1$, for some fixed $0 < \rho_1 < 1$. With $L \geq L_N, D' \geq D_N$ chosen in an appropriate way (see Appendix A.4), under $H_0 : C_{\eta\#P} = C_{\eta\#Q}$ we have*

$$N\tilde{\mathbb{T}}_{L,D'} \rightsquigarrow \sum_{m=1}^{\infty} \gamma_m(A_m^2 - 1),$$

where $A_m \sim N(0,1)$ for $m = 1,2,\ldots$, and γ_m are the eigenvalues of a certain operator that depends on P and Q. Further, under $H_1 : C_{\eta\#P} \neq C_{\eta\#Q}$, $\sqrt{N}\left(\tilde{\mathbb{T}}_{L,D'} - \mathbb{T}\right)$ is asymptotically Gaussian with mean 0 and finite variance.

We evaluate the power performance of the testing procedure based on $\tilde{\mathbb{T}}_{L,D'}$ for the sequence of contiguous alternatives $H_{1N} = \{(P,Q) : C_{\mu\#P} = C_{\mu\#Q} + \delta_N, l = 1,2,\ldots\}$, where the deviation from null is quantified collectively by pushforward differences $\delta_{\ell N} \in \mathcal{H}, \delta_N = \oplus_\ell(\sqrt{\alpha_\ell}\delta_{\ell N})$ that are made to approach 0 as $N \to \infty$. The following theorem establishes consistency of our testing procedure against a family of such local alternatives.

Theorem 3. *Assume the same conditions and choice of L, D' as Theorem 2. Then for the sequence of contiguous alternatives H_{1N} such that $N\|\delta_N\|_{\mathcal{H}^*}^2 \to \infty$, the test based on $\tilde{\mathbb{T}}_{L,D'}$ is consistent for any $\alpha \in (0,1)$, that is as $N \to \infty$ the asymptotic power approaches 1.*

Testing Procedure In practice, to obtain the p-value for the $\tilde{\mathbb{T}}_{L,D'}$-statistic we use a bootstrap procedure. Remember that $\tilde{\mathbb{T}}_{L,D'}$ is computed via the approximate embedding η_D with $D = LD'$. The collection $\{\mu_i\}_{i=1}^{N_1}$ is mapped to the collection $\{X_i = \eta_D(\mu_i)\}_{i=1}^{N_1}$ of vectors in $\mathbb{R}^D$ drawn in an i.i.d. manner from $\eta_D\#P = P \circ \eta_D^{-1} \in \mathcal{P}(\mathbb{R}^D)$. Similarly, for the other collection we have a sample $\{Y_i = \eta_D(\nu_i)\}_{i=1}^{N_2}$ drawn from $\eta_D\#Q$. Now, the null $H_0 : C_{\eta\#P} = C_{\eta\#Q}$ implies that the means of the distributions $\eta_D\#P$ and $\eta_D\#Q$ coincide in $\mathbb{R}^D$.

The bootstrap null distribution for $\tilde{\mathbb{T}}_{L,D'}$ can be obtained as follows. Let $\bar{X}$ and $\bar{Y}$ be the sample means; construct the combined sample $\{X_i - \bar{X} + \frac{\bar{X}+\bar{Y}}{2}\}_{i=1}^{N_1} \bigcup \{Y_i - \bar{Y} + \frac{\bar{X}+\bar{Y}}{2}\}_{i=1}^{N_2}$. This centers both samples at $\frac{\bar{X}+\bar{Y}}{2}$. Now, from the combined sample we select with replacement N_1 (resp. N_2) samples to make bootstrap sample $\{X_i^b\}_{i=1}^{N_1}$ (resp. $\{Y_i^b\}_{i=1}^{N_2}$). Repeat this process B times (we take $B = 1000$ in our experiments), and collect the null test statistic values $\tilde{\mathbb{T}}_{L,D'}^b = \tilde{\mathbb{T}}_{L,D'}(\{X_i^b\}_{i=1}^{N_1}, \{Y_i^b\}_{i=1}^{N_2})$ for $b = 1,\ldots,B$. The approximate p-value is then given by: $p = \frac{1}{B+1}\left(|\{b : \tilde{\mathbb{T}}_{L,D'}^b \geq \tilde{\mathbb{T}}_{L,D'}\}| + 1\right)$.

Remark 2. Permutation testing cannot be applied here as it would detect differences beyond the mean inequality. Such differences help reject the stronger null hypothesis $H_0 : P = Q$ which can be the intent in some situations. However, such a rejection will not allow pinpointing the aspect responsible for the difference; see (Huang et al., 2006).

5.2. Testing via p-value Combination

The bootstrap test above incurs a high computational cost and the granularity of the p-values is determined by the number of resamples, which can be too coarse in massive multiple comparison settings often seen in industrial applications. Thus, we propose an approach that avoids resampling.

As explained above, testing $H_0 : C_{\eta\#P} = C_{\eta\#Q}$ can be interpreted as testing whether the means of the distributions $\eta_D\#P$ and $\eta_D\#Q$ coincide in $\mathbb{R}^D$. To this

end, we adopt the approach proposed by (Rustamov & Klosowski, 2020) in a spatial statistics context. First, we apply the Behrens-Fisher-Welch t-test (without assuming equality of variances) to each coordinate of the samples $\{X_i = \eta_D(\mu_i)\}_{i=1}^{N_1}$ and $\{Y_i = \eta_D(\nu_i)\}_{i=1}^{N_2}$ to obtain the p-values $p_k, k = 1, 2, ..., D$. Second, an overall p-value is computed via the harmonic mean p-value combination method which is robust to dependencies (Good, 1958; Wilson, 2019): $p^H = H\left(D/(\frac{1}{p_1} + \frac{1}{p_2} + \cdots + \frac{1}{p_D})\right)$, where the function H has a known form described in (Wilson, 2019). Another approach for combining p-values is the Cauchy combination test (Liu & Xie, 2020), but in our numerical experiments we found that the Cauchy combination approach encounters problems when any of the p-values is very close to 1, which can happen in our setting due to the form of the embedding η_D. Therefore, in contrast to (Rustamov & Klosowski, 2020), for us the harmonic combination is the only appropriate choice.

To guarantee size control, we establish a version of Theorem 1 from (Liu & Xie, 2020) for the harmonic mean p-value. Assume that a test statistic $Z \in \mathbb{R}^D$ has null distribution with zero mean and every pair of coordinates of Z follows bivariate Gaussian distribution. Compute the coordinate-wise two-sided p-values $p_k = 2(1 - \Phi(|Z_k|))$ where Φ is the standard Gaussian CDF.

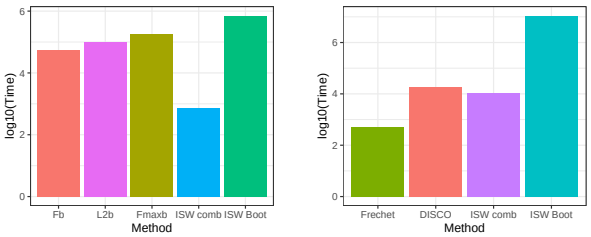
Theorem 4. *Let $p_k, k = 1, ..., D$ be the null p-values as above and p^H computed via harmonic mean approach, then* $\lim_{\alpha \to 0} \text{Prob}\{p^H \leq \alpha\}/\alpha = 1$.

In Appendix A.5 we show that this theorem applies in our setting, so the proposed procedure asymptotically controls the size of the test for small α. Our experimental results show that the control is already achieved for moderate sample sizes and the commonly used $\alpha = 0.05$.

6. Experiments

We compare the performance of our tests with several existing methods, across synthetic and real data from a number of domains, and settings of the embedding parameters L, D'. For evaluation, we use empirical power at different degrees of departure from the null hypothesis (Captured by δ in the plots in Fig. 3), calculated by averaging the proportion of rejections at level $\alpha = 0.05$ over 1000 independent datasets, with samples divided into two groups of sizes $n_1 = 60, n_2 = 40$. To ensure the tests are well-calibrated, we calculate nominal sizes assuming the two sample groups are drawn from the same meta-distribution. See Appendix B.1 for more details on the experiment settings, and discussions on computational complexity.

Finite intervals We use embedding dimensions $L = 3, D' = 10$ to compare our method against 11 functional



(a) (b)

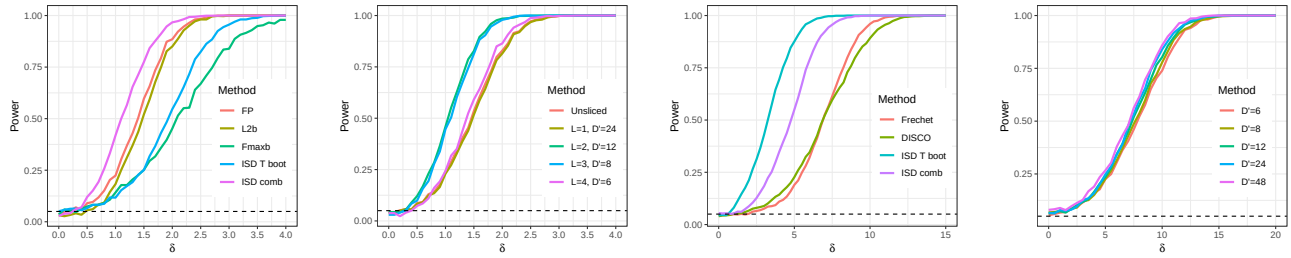
Figure 2. Computation times (at $\log_{10}$ scale) for (a) real line and (b) circle experiments.

ANOVA tests—we report results for 3 of them in Figure 3a (see Appendix B for complete results). All methods maintain nominal size for $\delta = 0$ (Figure 3a). While the combination test (ISD comb) based on our proposal outperformed all the other tests across all values of δ, the bootstrap test that uses the overall $\mathbb{T}$ statistic (ISD T boot) performs better than Fmaxb but worse than others. In terms of computation time, ISD comb takes the least amount of time (Fig. 2a).

We also compare the p-value combination test based on an *unsliced* 24-dimensional inverse CDF embedding with sliced $I\mathcal{SW}_2$-based tests (Figure 3b). We use multiple pairs of (L, D') values, all of them giving overall embeddings of dimension $D = LD' = 24$. The performance of an $I\mathcal{SW}_2$-based test that uses slicing over only the first eigenfunction is almost as good as the unsliced version. With more eigenfunctions, the powers first improve considerably, then become similar to the unsliced version again.

Manifold domains We consider data from distributions on circles and cylinders. For circular data, we take von Mises distributions with randomly chosen parameters as our samples. For an angle x (measured in radians), the von Mises probability density function is given by $f(x|\mu, \kappa) = \exp[\kappa \cos(x - \mu)](2\pi I_0(\kappa))^{-1}$, where $I_0(\kappa)$ is the modified Bessel function of order 0. We fix $\kappa = 2$, and use $\mu \equiv \mu_i \sim N(0, 0.1^2)$, $\mu \equiv \nu_i \sim N(\delta, 0.1^2)$ for samples from group 1 and 2 respectively—with $\delta \in [0, 15] \times \pi/180$ (i.e. 0 to 15 degrees converted to radians). As each observation vector, we take 100 random draws from each sample-specific distribution. For our embeddings, we use $L = 10, D' = 20$. Since the competing methods cannot handle circular geometry directly, to implement them we cut the circle into an interval. Figure 3c shows that all methods maintain nominal size, but both our tests maintain considerably higher power than existing methods for all δ. Computationally, ISD comb has comparable order of magnitude as the other two methods (Fig. 2b).

To generate cylindrical data, we use the distribution proposed by Mardia & Sutton (1978). Samples from each distribution have the form of a bivariate random vector (Θ, X), where the first (circular) marginal Θ has a von Mises distri-



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

Figure 3. Performances on synthetic data. Dotted lines indicates nominal size of all tests ($\alpha = 0.05$). (a) comparison with existing methods on finite interval—a test based on basis function representation (Gorecki & Smaga, 2015, FP), a sum-type ℓ_2 norm-based test (L2b) (Zhang, 2013), and a max-type test (Zhang et al., 2019) that uses the maximum of coordinate-wise F statistic (Fmaxb), (b) unsliced vs. different settings of (L, D') on finite interval. (c) Circular data, comparing with Fréchet ANOVA (Dubey & Müller, 2019), and the DISCO nonparametric test (Rizzo & Székely, 2010); (d) harmonic combination tests on cylindrical data for $L = 4$. .

bution, and X is a Gaussian conditional on $\Theta = \theta$ (details in Appendix). We draw the mean parameters for each of the two coordinate-wise distributions from Unif$(0, 1)$ and Unif$(\delta, \delta + 1)$ for sample groups 1 and 2 respectively, with $\delta \in [0, 30] \times \pi/180$. We then use 500 random draws from each sample distribution to obtain histograms. To evaluate the effects of choosing L, D' we calculate our embeddings for $L \in \{2, 3, 4, 5\}, D' \in \{6, 8, 12, 24, 48\}$. The choice of L has small effect on performance, so we report results for $L = 4$ in Figure 3d . Higher values of D' result in some increase in power.

Comparison to existing slicing methods As an example showcasing the importance of using intrinsic distances, consider the following graph setup: points A and B are adjacent on a planar regular 200-gon, with the edge AB removed. The intrinsic distance between A and B is large and the (extrinsic) Euclidean distance is small. No matter what Euclidean projection we use, the SW cost of moving probability mass from A to B would be small, leading to testing power loss when we compare two distribution sets that concentrate one around A and the other around B.

To numerically analyze this case we generate distributions on this graph by putting Unif(0,1) weights at each of the vertices, and added an additional weight of 10 to the bin at point A to obtain the first set of distributions with a mode at A. For the second set we put the weight of 10 at B instead to obtain a set of distributions with mode at B.

Table 1 shows comparison of $I\mathcal{SW}_2$ with SW in two dimensions, GSW with 3 choices of defining functions (circular, homogeneous polynomial of degree 3, and degree 5), MSW, and Sobolev transport (Le et al., 2022, ST). We use $L = 10, D' = 8$ to produce all embeddings, and implement the sliced version of ST by aggregating ST distances originating from 10 random vertices. $I\mathcal{SW}_2$ achieves the highest possible power, maintains nominal type-I error, and has

Method	Power	Size	Time (sec)
$I\mathcal{SW}_2$	1.00	0.029	630.67
SW	0.128	0.029	579.3
GSW-circle	0.128	0.029	564.82
GSW-poly3	0.025	0.025	779.69
GSW-poly5	0.044	0.035	1009.85
MSW	1.00	0.63	61764.22
Sobolev	1.00	0.048	21474.88

Table 1. Comparison of multiple slicing techniques.

computational time comparable to SW/GSW. ST and MSW achieve the same power as $I\mathcal{SW}_2$, but take much longer. This is because unlike slicing-based methods, they need to rely on permutation testing in absence of embeddings. For MSW, the high power comes at the price of a size that is much higher than the acceptable threshold of 0.05.

NHANES data on physical activity monitoring This data (NHANES, 2008) contains physical activity pattern readings for 6839 individuals. Data for each individual corresponds to activity monitor intensity values for $24 \times 60 = 1440$ minutes throughout the day, for 7 days. We capture this activity pattern into a cylindrical histogram with time and intensity (capped at 10,000) dimensions, having 96 and 100 bins respectively. Since the time dimension is periodic, normalized counts of this histogram can be considered as person-specific probability distributions over the cylinder $S^1(T_1) \times [0, T_2)$, with $T_1 = 96, T_2 = 100$. To check if activity patterns vary across different groups of individuals, we first split individuals into age-specific groups: 6–15, 16–25, ...,76–85, then sample 100 males and 100 females from each split. For our analysis, we consider $L = 3$ indices along the two directions, i.e. $\ell_1, \ell_2 = 1, 2, 3$ and $D' = 5$. We summarize the p-value combination test results in Table 6, *below the diagonal*. We run the Benjamini-Hochberg (Benjamini & Hochberg, 1995) procedure on the resulting p-values at

Age Groups	6–15	16–25	26–35	36–45	46–55	56–65	66–75	76–85
6–15		0.394	0.098	0.555	0.882	0.985	0.919	0.997
16–25	**1.2e-13**		0.575	0.967	0.126	0.921	0.911	0.977
26–35	**3.1e-21**	**2.7e-04**		0.459	0.197	0.996	0.919	0.565
36–45	**6.1e-22**	**7.9e-08**	**0.042**		0.864	0.637	0.849	0.991
46–55	**8.2e-22**	**4.7e-05**	**0.011**	0.343		0.841	0.165	0.554
56–65	**1.3e-25**	**0.001**	**0.001**	**5.6e-05**	**0.003**		0.991	0.962
66–75	**3.6e-35**	**7.8e-12**	**1.5e-11**	**4.6e-15**	**1.8e-13**	**0.001**		0.989
76–85	**3.8e-46**	**1.4e-26**	**1.7e-30**	**8.4e-37**	**2.1e-35**	**1.3e-17**	**6.5e-09**	

Table 2. Activity intensity comparison across age groups in the NHANES data. Below diagonal: p-values for to the actual data comparisons. Above diagonal: null p-values obtained by combining and randomly splitting the two groups. Bold entries correspond to rejected hypotheses with the BH procedure at FDR level 0.1.

Crime Type	Tue vs Thu	Tue vs Sat
Theft	0.428	**4.2e-06**
Deceptive Pract.	0.313	**0.001**
Battery	0.430	**0.001**
Robbery	0.119	**0.003**
Narcotics	0.854	**0.004**
Criminal Dam.	0.855	**0.02**
Other Offense	0.931	0.052
Burglary	0.142	0.261
Assault	0.997	0.38
Motor Veh. Theft	0.858	0.416

Table 3. Chicago Crime analysis p-values. Bold entries correspond to rejected hypotheses with the BH procedure at FDR level 0.1.

the false discovery rate of 0.1, and the rejected hypotheses are indicated by p-values in bold. $I\mathcal{SW}_2$ detects statistically significant differences between all pairs of groups, except the 36–45 and 46–55 groups. As expected, the control p-values—obtained by mixing male and female samples in each age group and splitting arbitrarily—do not concentrate near zero. More details are given in Appendix B.2.

Chicago Crime We use this dataset (City of Chicago, 2022) to show a practical usage of our method on histograms over graphs. Each beat (geographic area subdivision used by police) corresponds to a vertex, and two vertices are connected by an edge if the corresponding beats share a geographic boundary. For each crime type and day, the normalized counts of that crime type for each beat gives a daily probability distribution over the graph. Our goal is to compare the collection of distributions of, say, theft occurring on Tuesday to those of Thursday and Saturday. The Tuesday versus Thursday comparison is intended as a null case, as we do not expect to see any differences between them (Rustamov & Klosowski, 2020). Table 7 results $I\mathcal{SW}_2$ outcomes using 100-dimensional embeddings ($L = 20, D' = 5$).We detect statistically significant differences between Tuesday and Saturday patterns for six categories of crime, and as expected, no differences between Tuesday and Thursday patterns. For more details and another graph-based application, see Appendices B.3 and B.4, respectively.

7. Conclusion

A few limitations of the proposed framework are worth mentioning. There is scope of exploration for choosing the parameters L and D' in a principled manner. Empirical computation of eigenfunctions for general manifolds will introduce approximation errors that need to be tackled by expanding our theoretical results. Finally, per Theorem 1, $I\mathcal{SW}_2$ is theoretically a true metric. But in practice, the reflexiveness property (i.e. $d(x, y) = 0 \Leftrightarrow x = y$) is lost when a finite number of slices have to be used. This is also the case for all SW distances. In spite of that they remain highly effective practically. As our experiments demonstrate, $I\mathcal{SW}_2$ substantially improves upon the performance of SW distances.

There are two potential societal impacts of our proposal. Privacy-sensitive situations—such as analyzing manifold-valued personal data—can have be privacy risks associated with releasing the embeddings vectors. Additional studies are warranted to quantify such risks and generate differentially private embeddings. Any difference between data from different demographic groups found using our procedure should be evaluated in light of potential biases in the data collection phase.

The current paper applies the $I\mathcal{SW}_2$ framework in the context of Hypothesis tests for meta-distributions on general domains, due to the high relevance of this setup in real data situations. It is worth pointing out that the assumption-lean nature of $I\mathcal{SW}_2$ can enable adoption of many other statistical or ML methods to general spaces. The $I\mathcal{SW}_2$ embeddings can be used to do other hypothesis tests, such as tests for uniformity on a hypersphere (García-Portugués & Verdebout, 2018) or tests for equality of covariances by extending Wasserstein covariances on general domains. They can also be used as input features for supervised learning problems where prediction targets live in a general domain. Given that rigorous methodology for such problems has only been proposed recently (Han et al., 2020), development of prediction models for manifold-valued data free of restrictive assumptions is an attractive future line of research. $I\mathcal{SW}_2$ also is directly applicable as a loss function for generative modeling. For example, if one is generating distributions over a graph or manifold, $I\mathcal{SW}_2$ can serve as a loss function similarly to how other sliced distances are used in this context over Euclidean domains.

References

Aine, C. J., Bockholt, H. J., Bustillo, J. R., Cañive, J. M., Caprihan, A., Gasparovic, C., Hanlon, F. M., Houck, J. M., Jung, R. E., Lauriello, J., Liu, J., Mayer, A. R., Perrone-Bizzozero, N. I., Posse, S., Stephen, J. M., Turner, J. A., Clark, V. P., and Calhoun, V. D. Multimodal neuroimaging in schizophrenia: Description and dissemination. *Neuroinformatics*, 15(4):343–364, Oct 2017. ISSN 1559-0089. doi: 10.1007/s12021-017-9338-9. URL https://doi.org/10.1007/s12021-017-9338-9. B.4

Arroyo Relión, J. D., Kessler, D., Levina, E., and Taylor, S. F. Network classification with applications to brain connectomics. *Ann. Appl. Stat.*, 13(3):1648–1677, 09 2019. doi: 10.1214/19-AOAS1252. URL https://doi.org/10.1214/19-AOAS1252. B.4, B.4

Bakhvalov, N. S. On the approximate calculation of multiple integrals. *J. Complexity*, 31:502–516, 2015. English translation; the original appeared in Vestnik MGU, Ser. Math. Mech. Astron. Phys. Chem, 4, 3–18, 1959. A.4

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 00359246. 6, B.2, B.3

Berger, M. *A Panoramic View of Riemannian Geometry*. Springer-Verlag Berlin Heidelberg, 2003. 4.1

Bigot, J. Statistical data analysis in the wasserstein space. *ESAIM: ProcS*, 68:1–19, 2020. doi: 10.1051/proc/202068001. URL https://doi.org/10.1051/proc/202068001. 1, 3

Borden, B. and Luscombe, J. *Essential Mathematics for the Physical Sciences*, volume Volume I: Homogeneous boundary value problems, Fourier methods, and special functions, chapter Spherical harmonics and friends. Morgan & Claypool Publishers, 2017. Pages 6–1 to 6–26. A.3

Brown, L. and Steinerberger, S. On the Wasserstein distance between classical sequences and the Lebesgue measure. *Trans. Amer. Math. Soc.*, in press, 2020. https://arxiv.org/abs/1909.09046. A.4

City of Chicago. Chicago data portal: Crimes - 2018, 2022. URL https://data.cityofchicago.org/Public-Safety/Crimes-2018/3i3m-jwuy. 6

Coifman, R. R. and Lafon, S. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5 – 30, 2006. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2006.04.006. URL http://www.sciencedirect.com/science/article/pii/S1063520306000546. Special Issue: Diffusion Maps and Wavelets. 4, 4.2, A.3

Dehling, H. and Fried, R. Asymptotic distribution of two-sample empirical U-quantiles with applications to robust tests for shifts in location. *Journal of Multivariate Analysis*, 105(1):124–140, 2012. A.4

Deshpande, I., Hu, Y.-T., Sun, R., Pyrros, A., Siddiqui, N., Koyejo, S., Zhao, Z., Forsyth, D., and Schwing, A. G. Max-sliced wasserstein distance and its use for gans. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10640–10648, 2019. 2

Dubey, P. and Müller, H.-G. Frćhet analysis of variance for random objects. *Biometrika*, 106(4):803–821, 10 2019. ISSN 0006-3444. doi: 10.1093/biomet/asz052. URL https://doi.org/10.1093/biomet/asz052. 1, 3, 4

García-Portugués, E. and Verdebout, T. An overview of uniformity tests on the hypersphere, 2018. 7

Gavish, N., Nyquist, P., and Peletier, M. Large deviations and gradient flows for the Brownian one-dimensional hard-rod system. https://arxiv.org/abs/1909.02054, 2019. A.4

Good, I. J. Significance tests in parallel and in series. *Journal of the American Statistical Association*, 53(284):799–813, 1958. ISSN 01621459. 5.2

Gorecki, T. and Smaga, L. A Comparison of Tests for the One-Way ANOVA Problem for Functional Data. *Computational Statistics*, 30:987–1010, 2015. 3, 4

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, March 2012. ISSN 1532-4435. 2, 3, 4.1, 5.1, A.2, A.2, A.4

Han, K., Müller, H.-G., and Park, B. U. Additive functional regression for densities as responses. *Journal of the American Statistical Association*, 115:997–1010, 2020. 7

Hu, J., Shi, Y., and Xu, B. The gradient estimate of a neumann eigenfunction on a compact manifold with boundary. *Chinese Annals of Mathematics, Series B*, 36(6):991–1000, Nov 2015. ISSN 1860-6261. doi: 10.1007/s11401-015-0924-6. URL https://doi.org/10.1007/s11401-015-0924-6. A.2

Huang, Y., Xu, H., Calian, V., and Hsu, J. C. To permute or not to permute. *Bioinformatics*, 22(18):2244–2248, 07 2006. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl383. URL https://doi.org/10.1093/bioinformatics/btl383. 1, 2

Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R., and Rohde, G. Generalized sliced wasserstein distances. In Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 261–272. Curran Associates, Inc., 2019a. URL http://papers.nips.cc/paper/8319-generalized-sliced-wasserstein-distances.pdf. 1, 2

Kolouri, S., Pope, P. E., Martin, C. E., and Rohde, G. K. Sliced wasserstein auto-encoders. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019b. URL https://openreview.net/forum?id=H1xaJn05FQ. 1

Lai, R. and Zhao, H. Multiscale Nonrigid Point Cloud Registration Using Rotation-Invariant Sliced-Wasserstein Distance via Laplace–Beltrami Eigenmap. In *SIAM J. Imaging Sci.*, volume 10 of *2*, pp. 449–483, 2017. 2

Le, T., Yamada, M., Fukumizu, K., and Cuturi, M. Tree-sliced variants of wasserstein distances. In *Advances in Neural Information Processing Systems*, volume 32, 2019. 2

Le, T., Nguyen, T., Phung, D., and Anh Nguyen, V. Sobolev transport: A scalable metric for probability measures with graph metrics. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 9844–9868. PMLR, 2022. 2, 6, B.1

Lipman, Y., Rustamov, R. M., and Funkhouser, T. A. Biharmonic distance. *ACM Trans. Graph.*, 29(3), July 2010. ISSN 0730-0301. doi: 10.1145/1805964.1805971. URL https://doi.org/10.1145/1805964.1805971. 4, A.3

Liu, Y. and Xie, J. Cauchy combination test: A powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115(529):393–402, 2020. doi: 10.1080/01621459.2018.1554485. 5.2, A.5, A.5

Mardia, K. V. and Sutton, T. W. A Model for Cylindrical Variables with Applications. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(2):229–233, 1978. 6, B.1

Micchelli, C. A., Xu, Y., and Zhang, H. Universal kernels. *J. Mach. Learn. Res.*, 7:2651–2667, December 2006. ISSN 1532-4435. A.2

Murray, M. K. and Rice, J. W. *Differential Geometry and Statistics*, chapter The definition of a statistical manifold (Chapter 3.2). Chapman & Hall, 1993. 2

NHANES. 2005-2006 Data Documentation, Codebook, and Frequencies, 2008. URL https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/PAXRAW_D.htm. 6

Panaretos, V. M. and Zemel, Y. Statistical aspects of wasserstein distances. *Annual Review of Statistics and Its Application*, 6(1):405–431, 2019. doi: 10.1146/annurev-statistics-030718-104938. 1, 3

Petersen, A. and Müller, H.-G. Functional data analysis for density functions by transformation to a hilbert space. *Ann. Statist.*, 44(1):183–218, 02 2016. doi: 10.1214/15-AOS1363. URL https://doi.org/10.1214/15-AOS1363. 1, 2

Peyré, G. and Cuturi, M. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019. ISSN 1935-8237. doi: 10.1561/2200000073. URL http://dx.doi.org/10.1561/2200000073. 1, 3, 3, 1, 1

Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., Vogel, A. C., Laumann, T. O., Miezin, F. M., Schlaggar, B. L., and Petersen, S. E. Functional network organization of the human brain. *Neuron*, 72(4):665–678, Nov 2011. ISSN 1097-4199. doi: 10.1016/j.neuron.2011.09.006. URL https://pubmed.ncbi.nlm.nih.gov/22099467. 22099467[pmid]. B.4

Reed, M. and Simon, B. *Methods of modern mathematical physics. Vol. 1: Functional Analysis*. Academic Press, 1080. A.4

Reuter, M., Wolter, F.-E., Shenton, M., and Niethammer, M. Laplace-Beltrami eigenvalues and topological features of eigenfunctions for statistical shape analysis. *Computer-Aided Design*, 41(10):739–755, 2009. 4.2, A.3

Rizzo, M. L. and Székely, G. J. DISCO analysis: A nonparametric extension of analysis of variance. *Ann. Appl. Stat.*, 4:1034–1055, 2010. 3, 4

Rubner, Y., Tomasi, C., and Guibas, L. J. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000. doi: 10.1023/A:1026543900054. URL https://doi.org/10.1023/A:1026543900054. A.2

Rustamov, R. M. and Klosowski, J. T. Kernel mean embedding based hypothesis tests for comparing spatial point patterns. *Spatial Statistics*, 38:100459, 2020. ISSN 2211-6753. doi: https://doi.org/10.1016/j.spasta.2020.100459. URL http://www.sciencedirect.com/science/article/pii/S2211675320300531. 5.2, 6, B.3

Serfling, R. J. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 2009. 5.1, A.4, A.4

Solomon, J., Rustamov, R. M., Guibas, L. J., and Butscher, A. Wasserstein propagation for semi-supervised learning. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pp. 306–314. JMLR.org, 2014. URL http://proceedings.mlr.press/v32/solomon14.html. 1

Villani, C. *Topics in optimal transportation*. Graduate studies in mathematics. American mathematical society, Providence, Rhode Island, 2003. ISBN 0-8218-3312-X. A.2

Wilson, D. J. The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences*, 116(4):1195–1200, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1814092116. 5.2, A.5

Zhang, J.-T. *Analysis of Variance for Functional Data*. Chapman and Hall/CRC, first edition, 2013. 3, 4

Zhang, J.-T., Chen, M.-Y., Wu, H.-T., and Zhou, B. A new test for functional one-way ANOVA with applications to ischemic heart screening. *Computational Statistics & Data Analysis*, 132:3–17, 2019. 3, 4

Appendix

For the appendix to be self-contained, we restate results (theory and experiments) in the main paper when necessary. We start with proofs of theoretical results and some implementation details in Appendix A and details of experiments performed in Appendix B.

A. Proofs and additional results

A.1. Proofs and Notes for Section 3

Proposition 1. *For $P, Q \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$, the following equality holds:*

$$\mathbb{T}(P,Q) = \mathbb{E}_{\mu\sim P,\nu\sim Q}[\mathcal{D}^2(\mu,\nu)] - \frac{1}{2}\mathbb{E}_{\mu,\mu'\sim P}[\mathcal{D}^2(\mu,\mu')] - \frac{1}{2}\mathbb{E}_{\nu,\nu'\sim Q}[\mathcal{D}^2(\nu,\nu')], \tag{A.1}$$

where to avoid notational clutter we use $\mathcal{D}^2(\cdot,\cdot)$ as a shorthand for $(\mathcal{D}(\cdot,\cdot))^2$.

Proof. This is a straightforward application of the "kernel trick": using the Hilbert property of the distance we can rewrite,

$$\begin{aligned}
&\mathbb{E}_{\mu\sim P,\nu\sim Q}[\|\eta(\mu)-\eta(\nu)\|^2_{\mathcal{H}}] - \frac{1}{2}\mathbb{E}_{\mu,\mu'\sim P}[\|\eta(\mu)-\eta(\mu')\|^2_{\mathcal{H}}] - \frac{1}{2}\mathbb{E}_{\nu,\nu'\sim Q}[\|\eta(\nu)-\eta(\nu')\|^2_{\mathcal{H}}]\\
&\quad=\mathbb{E}_{\mu\sim P}[\|\eta(\mu)\|^2_{\mathcal{H}}] + \mathbb{E}_{\nu\sim Q}[\|\eta(\nu)\|^2_{\mathcal{H}}] - 2\langle\mathbb{E}_{\mu\sim P}[\eta(\mu)], \mathbb{E}_{\nu\sim Q}[\eta(\nu)]\rangle_{\mathcal{H}}\\
&\quad-\mathbb{E}_{\mu\sim P}[\|\eta(\mu)\|^2_{\mathcal{H}}] - \mathbb{E}_{\nu\sim Q}[\|\eta(\nu)\|^2_{\mathcal{H}}]\\
&\quad+\langle\mathbb{E}_{\mu\sim P}[\eta(\mu)], \mathbb{E}_{\mu\sim P}[\eta(\mu)]\rangle_{\mathcal{H}} + \langle\mathbb{E}_{\nu\sim Q}[\eta(\nu)], \mathbb{E}_{\nu\sim Q}[\eta(\nu)]\rangle_{\mathcal{H}}\\
&\quad=\|\mathbb{E}_{\mu\sim P}[\eta(\mu)] - \mathbb{E}_{\nu\sim Q}[\eta(\nu)]\|^2_{\mathcal{H}} = \mathbb{T}(P,Q).
\end{aligned}$$

Which gives the sought equivalence. □

A.2. Proofs and Notes for Section 4.1

Proposition 2. *If $\mathcal{X}$ is a smooth compact n-dimensional manifold and $\sum_\ell \lambda_\ell^{(n-1)/2}\alpha(\lambda_\ell) < \infty$, then $IS\mathcal{W}_2$ is well-defined.*

Proof. We use Hörmander's bound on the supremum norm of the eigenfunctions:

$$\|\phi_\ell\|_\infty \leq c\lambda_\ell^{(n-1)/4}\|\phi_\ell\|_2,$$

for some constant c that depends on the manifold. By orthonormality of the eigenfunctions we have $\forall \ell, \|\phi_\ell\|_2 = 1$. Next, note that $\mathcal{W}_2(\phi_\ell\sharp\mu, \phi_\ell\sharp\nu) \leq 2\|\phi_\ell\|_\infty$ as the maximum distance that the mass would be transported in any transportation plan involving pushforwards via ϕ_ℓ is upper bounded by $2\|\phi_\ell\|_\infty$. As a result, every term in the series defining $IS\mathcal{W}_2$ can be upper-bounded by the terms of the following series:

$$\sum_\ell 4\|\phi_\ell\|^2_\infty\alpha(\lambda_\ell) \leq \sum_\ell 4c^2\lambda_\ell^{(n-1)/2}\alpha(\lambda_\ell) \propto \sum_\ell \lambda_\ell^{(n-1)/2}\alpha(\lambda_\ell),$$

which proves the claim by the direct comparison test for convergence of series. □

Remark 3. When Weyl law applies, we have that $\lambda_\ell = \Theta(\ell^{2/n})$, which allows us to replace the above condition by $\sum_\ell \ell^{(n-1)/n}\alpha(\lambda_\ell) < \infty$. For the diffusion kernel/distance choice of $\alpha(\lambda) = e^{-t\lambda}$ the series always converges independently of the manifold dimension. For biharmonic choice of $\alpha(\lambda) = 1/\lambda^2$, the sufficient condition is the convergence of $\sum_\ell \ell^{(n-1)/n}/\lambda_\ell^2 \sim \sum_\ell \ell^{(n-1)/n}/(\ell^{2/n})^2 = \sum_\ell \ell^{(n-5)/n}$, where we applied Weyl's asymptotic again. As a result, the biharmonic choice of α is guaranteed to provide a well-defined $IS\mathcal{W}_2$ for 1 and 2-dimensional manifolds. Notice, however, that the Hörmander's bound used in the proof of the above proposition can be rather lax in some of the settings that are practically relevant, such as the product spaces of lines and circles (where all of the eigenfunctions are bounded by a constant as can be seen from Table 4), and, thus, convergence for the biharmonic choice holds more widely.

Proposition 3. *If $\mathcal{D}$ is a Hilbertian probability distance such that $IS\mathcal{D}$ is well-defined, then*

(i) $IS\mathcal{D}$ is Hilbertian, and

(ii) $IS\mathcal{D}$ satisfies the following metric properties: non-negativity, symmetry, the triangle inequality, and $IS\mathcal{D}(\mu,\mu) = 0$.

Proof. By Hilbertian property of $\mathcal{D}$, there exists a Hilbert space $\mathcal{H}^0$ and a map $\eta^0 : \mathcal{P}(\mathbb{R}) \to \mathcal{H}^0$ such that $\mathcal{D}(\rho_1, \rho_2) = \|\eta^0(\rho_1) - \eta^0(\rho_2)\|_{\mathcal{H}^0}$ for all $\rho_1, \rho_2 \in \mathcal{P}(\mathbb{R})$. Plugging this into the definition of $IS\mathcal{D}$ we have $IS\mathcal{D}(\mu, \nu) = \|\eta(\mu) - \eta(\nu)\|_{\mathcal{H}}$, where $\mathcal{H} = \oplus_\ell \mathcal{H}^0$ and the ℓ-th component of $\eta(\mu)$ is $\sqrt{\alpha(\lambda_\ell)}\eta_0(\phi_\ell \sharp \mu) \in \mathcal{H}^0$. The second part of Proposition 3 directly follows from the Hilbert property. □

Proposition 4. *When $\mu = \delta_x(\cdot), \nu = \delta_y(\cdot)$ for two points $x, y \in \mathcal{X}$, we have $IS\mathcal{W}_2(\mu, \nu) = d(x, y)$, where $d(\cdot, \cdot)$ is the spectral distance corresponding to the choice of $\alpha(\cdot)$.*

Proof. We have $\phi_\ell \sharp \delta_x = \delta_{\phi_\ell(x)}$ and similarly for y. Now $\mathcal{W}_2^2(\phi_\ell \sharp \mu, \phi_\ell \sharp \nu) = \mathcal{W}_2^2(\delta_{\phi_\ell(x)}, \delta_{\phi_\ell(y)}) = (\phi_\ell(x) - \phi_\ell(y))^2$. This last equality follows from the fact that the 2-Wasserstein on real line between delta measures is equal to the distance between the two points. Then scaling and adding up gives exactly the kernel distance $d(x, y)$ between the two points. □

Proposition 5. *Let $\mathcal{D}(\rho_1, \rho_2) = |\mathbb{E}_{x\sim\rho_1}[x] - \mathbb{E}_{y\sim\rho_2}[y]|$ for $\rho_1, \rho_2 \in \mathcal{P}(\mathbb{R})$, then the corresponding intrinsic sliced distance is equivalent to the MMD with the spectral kernel $k(\cdot, \cdot)$.*

Proof. We can rewrite the definition as follows:

$$
\begin{aligned}
IS\mathcal{D}^2(\mu, \nu) &= \sum_\ell \alpha(\lambda_\ell)(\mathbb{E}_{x\sim\phi_\ell\sharp\mu}[x] - \mathbb{E}_{y\sim\phi_\ell\sharp\nu}[y])^2 = \sum_\ell \alpha(\lambda_\ell)(\mathbb{E}_{x\sim\mu}[\phi_\ell(x)] - \mathbb{E}_{y\sim\nu}[\phi_\ell(y)])^2 \\
&= \sum_\ell \alpha(\lambda_\ell)(\mathbb{E}_{x,x'\sim\mu}[\phi_\ell(x)\phi_\ell(x')] + \mathbb{E}_{y,y'\sim\nu}[\phi_\ell(y)\phi_\ell(y')] - 2\mathbb{E}_{x\sim\mu,y\sim\nu}[\phi_\ell(x)\phi_\ell(y)]) \\
&= \mathbb{E}_{x,x'\sim\mu}[\sum_\ell \alpha(\lambda_\ell)\phi_\ell(x)\phi_\ell(x')] + \mathbb{E}_{y,y'\sim\nu}[\sum_\ell \alpha(\lambda_\ell)\phi_\ell(y)\phi_\ell(y')] \\
&\quad - 2\mathbb{E}_{x\sim\mu,y\sim\nu}[\sum_\ell \alpha(\lambda_\ell)\phi_\ell(x)\phi_\ell(y)] \\
&= \mathbb{E}_{x,x'\sim\mu}[k(x, x')] + \mathbb{E}_{y,y'\sim\nu}[k(y, y')] - 2\mathbb{E}_{x\sim\mu,y\sim\nu}[k(x, y)],
\end{aligned}
$$

where we used the spectral kernel $k(x, y) = \sum_\ell \alpha(\lambda_\ell)\phi_\ell(x)\phi_\ell(y)$. The last expression coincides with the MMD based on kernel $k(\cdot, \cdot)$; see Lemma 6 in (Gretton et al., 2012). □

Proposition 6. *$MMD(\mu, \nu) \leq IS\mathcal{W}_2(\mu, \nu)$ when the same $\alpha(\cdot)$ is used in both constructions.*

Proof. This follows directly from the fact that for $\rho_1, \rho_2 \in \mathcal{P}(\mathbb{R})$ the inequality $|\mathbb{E}_{x\sim\rho_1}[x] - \mathbb{E}_{y\sim\rho_2}[y]| \leq \mathcal{W}_1(\rho_1, \rho_2) \leq \mathcal{W}_2(\rho_1, \rho_2)$ holds. Here the first inequality follows from the centroid bound (Rubner et al., 2000), and the second inequality is the well-known ordering property of Wasserstein distances (Villani, 2003). □

Theorem 1. *If $\alpha(\lambda) > 0$ for all $\lambda > 0$, then $IS\mathcal{W}_2$ is a metric on $\mathcal{P}(\mathcal{X})$.*

Proof. In the light of the Proposition 3 it remains only to prove that $IS\mathcal{W}_2(\mu, \nu) = 0$ implies $\mu = \nu$. According to Proposition 6, $IS\mathcal{W}_2(\mu, \nu) = 0$ yields $MMD(\mu, \nu) = 0$. The assumption that $\alpha(\lambda) > 0$ for all $\lambda > 0$ implies that the spectral kernel $k(\cdot, \cdot)$ corresponding to $\alpha(\cdot)$ is universal (Micchelli et al., 2006). Universality implies the characteristic property (Gretton et al., 2012), which in turn means that $MMD(\mu, \nu) = 0$ is equivalent to $\mu = \nu$, proving the claim. □

Proposition 7. *There exists a constant c depending only on $\mathcal{X}$ such that for all $\mu, \nu \in \mathcal{P}(\mathcal{X})$ the inequality $IS\mathcal{W}_2(\mu, \nu) \leq c\mathcal{W}_2^{\mathcal{X}}(\mu, \nu)\sqrt{\sum_\ell \lambda_\ell^{(n+3)/2}\alpha(\lambda_\ell)}$ holds; here, n is the dimension of $\mathcal{X}$.*

Proof. We remind $\mathcal{W}_2^{\mathcal{X}}$ is the 2-Wasserstein distance defined directly $\mathcal{P}(\mathcal{X})$ using the geodesic distance as the ground metric. The Neumann eigenfunctions on compact manifolds satisfy the inequality $\|\nabla\phi_\ell\|_\infty \leq c_1\lambda_\ell\|\phi_\ell\|_\infty$, see (Hu et al., 2015). Applying the bound used in the proof of convergence, $\|\phi_\ell\|_\infty \leq c_2\lambda_\ell^{(n-1)/4}$, we get that ϕ_ℓ is Lipschitz with respect to the geodesic distance on $\mathcal{X}$ with the Lipschitz constant bounded by $c\lambda_\ell\lambda_\ell^{(n-1)/4} = c\lambda_\ell^{(n+3)/4}$.

Consider the optimal coupling between μ and ν whose cost equals to $\mathcal{W}_2^{\mathcal{X}}(\mu, \nu)$. Note that this coupling straightforwardly provides a coupling between the pushforwards $\phi_\ell \sharp \mu$ and $\phi_\ell \sharp \nu$. Using the Lipschitz property of eigenfunctions, we see that the cost of the pushforward coupling is smaller than $c\lambda_\ell^{(n+3)/4}\mathcal{W}_2^{\mathcal{X}}(\mu, \nu)$. Since any such coupling provides an upper bound

$\mathcal{X}$	Eigenvalues	Eigenfunctions
$[0,T]$	$(\frac{\pi\ell}{T})^2$	$\sqrt{\frac{2}{T}}\cos\frac{\pi\ell x}{T}$
$S^1(T)=[0,T] \mod T$	$(\frac{2\pi\ell}{T})^2$	$\sqrt{\frac{2}{T}}[\cos/\sin]\frac{2\pi\ell x}{T}$
$[0,T_1]\times[0,T_2]$	$(\frac{\pi\ell_1}{T_1})^2+(\frac{\pi\ell_2}{T_2})^2$	$\sqrt{\frac{4}{T_1T_2}}\cos\frac{\pi\ell_1 x}{T_1}\cos\frac{\pi\ell_2 x}{T_2}$
$S^1(T_1)\times[0,T_2]$	$(\frac{2\pi\ell_1}{T_1})^2+(\frac{\pi\ell_2}{T_2})^2$	$\sqrt{\frac{4}{T_1T_2}}[\cos/\sin]\frac{2\pi\ell_1 x}{T_1}\cos\frac{\pi\ell_2 x}{T_2}$
$S^1(T_1)\times S^1(T_2)$	$(\frac{2\pi\ell_1}{T_1})^2+(\frac{2\pi\ell_2}{T_2})^2$	$\sqrt{\frac{4}{T_1T_2}}[\cos/\sin]\frac{2\pi\ell_1 x}{T_1}[\cos/\sin]\frac{\pi\ell_2 x}{T_2}$
S^2	Spherical harmonics (Borden & Luscombe, 2017)	
Graphs/Data Clouds/Meshes	Eigen-decomposition of the Laplacian matrix	

Table 4. Eigenvalues and eigenfunctions of the Laplace-Beltrami operator with Neumann boundary conditions for simple manifolds. We exclude zero eigenvalue and the corresponding constant eigenvector; thus, all indices ℓ, ℓ_1, ℓ_2 run over positive integers. The notation $[\cos/\sin]$ means picking either the cosine or sine function—*all choices must be used, giving multiple eigenfunctions*.

on $\mathcal{W}_2(\phi_\ell\sharp\mu, \phi_\ell\sharp\nu)$, we have $\mathcal{W}_2(\phi_\ell\sharp\mu, \phi_\ell\sharp\nu) \le c\lambda_\ell^{(n+3)/4}\mathcal{W}_2^{\mathcal{X}}(\mu,\nu)$. Plugging this into the formula for ISW_2 we get the claimed bound. □

Proposition 8. *Let $\{\mu_i\}_{i=1}^N$ and $\{\nu_i\}_{i=1}^N$ be two collections of probability measures on $\mathcal{P}(\mathcal{X})$, such that $\forall i, \mathcal{W}_2^{\mathcal{X}}(\mu_i,\nu_i)\le\epsilon$, then $\mathbb{T}(\{\mu_i\}_{i=1}^N,\{\nu_i\}_{i=1}^N)\le C^2\epsilon^2$. Here $C=c\sqrt{\sum_\ell \lambda_\ell^{(n+3)/2}\alpha(\lambda_\ell)}$ from previous proposition and is assumed to be finite.*

Proof. We have

$$\begin{aligned}\mathbb{T}(\{\mu_i\}_{i=1}^N,\{\nu_i\}_{i=1}^N) &= \left\|\frac{1}{N}\sum_{i=1}^N\eta(\mu_i)-\frac{1}{N}\sum_{i=1}^N\eta(\nu_i)\right\|_{\mathcal{H}}^2 = \left\|\frac{1}{N}\sum_{i=1}^N(\eta(\mu_i)-\eta(\nu_i))\right\|_{\mathcal{H}}^2\\ &\le \frac{1}{N}\sum_{i=1}^N\|\eta(\mu_i)-\eta(\nu_i)\|_{\mathcal{H}}^2 = \frac{1}{N}\sum_{i=1}^N ISW_2^2(\mu_i,\nu_i) \le \frac{1}{N}\sum_{i=1}^N(C\mathcal{W}_2^{\mathcal{X}}(\mu_i,\nu_i))^2\\ &\le \frac{1}{N}N(C\epsilon)^2 = C^2\epsilon^2.\end{aligned}$$

□

A.3. Computational Details for Section 4.2

The case of finite intervals is the building block for the general case, so let us first consider the case of $\mathcal{X}=[0,T]$. We represent a histogram over this interval by a discrete measure of the form $\mu=\sum w_a\delta_{x_a}$ with the histogram bin centers $x_a\in[0,T]$ and weights w_a satisfying $\sum w_a=1$, where $a=1,2,...,A$. Note that it is not required for the histograms in the collections to be supported at the same bin locations. For a given histogram, let $\{x_{(a)},w_{(a)}\}_{a=1}^A$ be the locations sorted from smallest to largest and their corresponding weights; since the bin locations are unique there will not be any ties. The quantile function is computed via $F_\mu^{-1}(s):=\min\{x_{(a)}:\sum_{b\le a}w_{(b)}>s\}$. The approximate map $\eta_{D'}^0$ now can be computed using the s_k-th quantile value $F_\mu^{-1}(s_k)$ for each value of $s_k, k=1,...,D'$.

For a general domain $\mathcal{X}$, the histogram representation is the same as above: $\sum w_a\delta_{x_a}$ with the histogram bin centers $x_a\in\mathcal{X}$ and weights w_a satisfying $\sum w_a=1$, where $a=1,2,...,A$. The pushforward $\phi_\ell\sharp\mu$ gives a histogram on the real line defined by $\sum w_a\delta_{\phi_\ell(x_a)}$. Note that while x_a are distinct, their images under ϕ_ℓ do not have to be distinct, so one re-aggregates the weights to obtain $\sum_{a\in S}w'_a\delta_{\phi_\ell(x_a)}$, where S is a subset of $1,2,...,A$ and w'_a are the new weights. It is now straightforward to compute the quantile function as before and build the approximate map $(\eta_D)_\ell$. Doing so for the different values of ℓ and concatenating the resulting vectors gives η_D.

In practice, these computations can be carried out on a variety of domains—analytic manifolds, manifolds discretized as point clouds or meshes, and graphs. In most cases the spectral decomposition of the Laplace-Beltrami operator or graph Laplacian has to be computed numerically (Coifman & Lafon, 2006; Reuter et al., 2009). For applications that involve simple manifolds, the eigenvalues and eigenfunctions can be computed analytically. For completeness we list them in Table 4. Note

that we benefit from the fact that the eigen-decomposition for product spaces can be derived from the eigen-decompositions of the components.

The choice of the function $\alpha(\cdot)$ determining the contributions of each spectral band is problem specific. When working on manifolds of low dimension, the choice of $\alpha(\cdot)$ that corresponds to the biharmonic distance is convenient. While the diffusion distance provides a general choice that works on manifolds of any dimension, the biharmonic distance does not have any parameters to tune and was shown to provide an excellent alternative to the geodesic distance in low-dimensional settings (Lipman et al., 2010). When in doubt, inspecting the behavior of the distance on the underlying domain will allow assessing whether the distance is appropriate for the given problem.The importance of relying on a well-behaved spectral distance was highlighted in Proposition 4.

A.4. Proofs and Notes for Section 5.1

We remind that we will be using the following test statistic for the results that are discussed below:

$$\hat{\mathbb{T}} \equiv \sum_{i,j} \frac{ISW_2^2(\mu_i,\nu_j)}{N_1N_2} - \sum_{i,j:i\neq j} \frac{ISW_2^2(\mu_i,\mu_j)}{2N_1(N_1-1)} - \sum_{i,j:i\neq j} \frac{ISW_2^2(\nu_i,\nu_j)}{2N_2(N_2-1)}. \tag{A.2}$$

Proposition 9. *Assume conditions (i)-(iii) hold. Define $N = N_1 + N_2$, and assume that as $N_1, N_2 \to \infty$, we have $N_1/N \to \rho_1, N_2/N \to \rho_2 = 1-\rho_1$, for some fixed $0 < \rho_1 < 1$. Define a new measure R as a scaled mixture of the centered pushforward measures*

$$R = \left(\frac{1}{\rho_1}+\frac{1}{\rho_2}\right)^{-1}\left[\frac{1}{\rho_1}\left(\eta\#P - C_{\eta\#P}\right) + \frac{1}{\rho_2}\left(\eta\#Q - C_{\eta\#Q}\right)\right] = \rho_2\left(\eta\#P - C_{\eta\#P}\right) + \rho_1\left(\eta\#Q - C_{\eta\#Q}\right).$$

Suppose $\gamma_m, m = 1,2,\ldots$ are the eigenvalues of

$$\frac{1}{\rho_1\rho_2}\int_{\mathcal{H}} \langle x, x'\rangle_{\mathcal{H}} \psi_m(x') dR(x') = \gamma_m \psi_m(x).$$

Then under $H_0 : C_{\eta\#P} = C_{\eta\#Q}$ we have

$$N\hat{\mathbb{T}} \rightsquigarrow \sum_{m=1}^{\infty} \gamma_m (A_m^2 - 1), \tag{A.3}$$

where A_m are i.i.d. $\mathcal{N}(0,1)$ random variables. Under $H_1 : C_{\eta\#P} \neq C_{\eta\#Q}$ we have $\sqrt{N}(\hat{\mathbb{T}} - \mathbb{T}) \rightsquigarrow N(0,\sigma_1^2)$, where

$$\begin{aligned}\sigma_1^2 \;=\; & 4\left[\frac{1}{\rho_1}\mathbb{V}_{\mu\sim P}\mathbb{E}_{\mu'\sim P}\langle\eta(\mu),\eta(\mu')\rangle_{\mathcal{H}} + \frac{1}{\rho_2}\mathbb{V}_{\nu\sim Q}\mathbb{E}_{\nu'\sim Q}\langle\eta(\nu),\eta(\nu')\rangle_{\mathcal{H}} + \right.\\ & \left.\frac{1}{\rho_1}\mathbb{V}_{\mu\sim P}\mathbb{E}_{\nu\sim Q}\langle\eta(\mu),\eta(\nu)\rangle_{\mathcal{H}} + \frac{1}{\rho_2}\mathbb{V}_{\nu\sim Q}\mathbb{E}_{\mu\sim P}\langle\eta(\mu),\eta(\nu)\rangle_{\mathcal{H}}\right].\end{aligned} \tag{A.4}$$

Proof. Using the Hilbertianity of ISD (Proposition 3), we have

$$\begin{aligned} ISD^2(\mu_i,\mu_j) =& \|\eta(\mu_i) - \eta(\mu_j)\|_{\mathcal{H}}^2 \\ =& \|\eta(\mu_i)\|_{\mathcal{H}}^2 + \|\eta(\mu_j)\|_{\mathcal{H}}^2 - 2\langle\eta(\mu_i),\eta(\mu_j)\rangle_{\mathcal{H}}\end{aligned},$$

Consequently

$$\sum_{i,j:i\neq j} ISD^2(\mu_i,\mu_j) = 2(N_1-1)\sum_{i=1}^{N_1}\|\eta(\mu_i)\|_{\mathcal{H}}^2 - 2\sum_{i,j:i\neq j}\langle\eta(\mu_i),\eta(\mu_j)\rangle_{\mathcal{H}}.$$

Similarly,

$$\begin{aligned}\sum_{i,j:i\neq j} ISD^2(\nu_i,\nu_j) \;&=\; 2(N_2-1)\sum_{i=1}^{N_2}\|\eta(\nu_i)\|_{\mathcal{H}}^2 - 2\sum_{i,j:i\neq j}\langle\eta(\nu_i),\eta(\nu_j)\rangle_{\mathcal{H}},\\ \sum_{i,j} ISD^2(\mu_i,\nu_j) \;&=\; N_2\sum_{i=1}^{N_1}\|\eta(\mu_i)\|_{\mathcal{H}}^2 + N_1\sum_{j=1}^{N_2}\|\eta(\nu_j)\|_{\mathcal{H}}^2 - 2\sum_{i,j:i\neq j}\langle\eta(\mu_i),\eta(\nu_j)\rangle_{\mathcal{H}}.\end{aligned}$$

Putting these back into Eq. (A.2) after simplifying and cancelling out the norm-square terms we have

$$\hat{\mathbb{T}} = \frac{1}{N_1(N_1-1)} \sum_{i,j:i\neq j} \langle \eta(\mu_i), \eta(\mu_j) \rangle_{\mathcal{H}} + \frac{1}{N_2(N_2-1)} \sum_{i,j:i\neq j} \langle \eta(\nu_i), \eta(\nu_j) \rangle_{\mathcal{H}} - \frac{2}{N_1 N_2} \sum_{i,j} \langle \eta(\mu_i), \eta(\nu_j) \rangle_{\mathcal{H}}. \tag{A.5}$$

At this point, we replace the maps η by their centered versions $\tilde{\eta}(\mu) = \eta(\mu) - C_{\eta\#P}, \tilde{\eta}(\nu) = \eta(\nu) - C_{\eta\#Q}$; remember that the center of mass of $\eta\#P$ is denoted by $C_{\eta\#P}$. Accumulating the sample-level partial sums above the centering terms cancel out under $H_0 : C_{\eta\#P} = C_{\eta\#Q}$, so that each η can be replaced by $\tilde{\eta}$ in (A.5) above.

Denote $x_i \equiv \tilde{\eta}(\mu_i), y_i \equiv \tilde{\eta}(\nu_i)$ as the Hilbert-embedded samples of $X \sim \tilde{\eta}\#P, Y \sim \tilde{\eta}\#Q$, respectively. We remind now that R is a mixture of the centered pushforward measures: $R = \rho_2(\tilde{\eta}\#P) + \rho_1(\tilde{\eta}\#Q)$. Let $L_2(\mathcal{H}, R)$ be the space of real-valued functions on $\mathcal{H}$ that are square integrable with respect to R. Now we can define the following operator $S : L_2(\mathcal{H}, R) \to \mathcal{H}$,

$$(Sf)(x) := \int_{\mathcal{H}} \langle x, x' \rangle_{\mathcal{H}} f(x') dR(x').$$

Following condition (ii), $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is square-integrable under R. The above operator is thus Hilbert-Schmidt, hence compact (Reed & Simon, 1080, Theorem VI.23). Consequently, it permits an eigenfunction decomposition with respect to measure R, $\langle x, x' \rangle_{\mathcal{H}} = \sum_{m=1}^{\infty} \gamma_m \psi_m(x) \psi_m(x')$, for $x, x' \in \mathcal{H}$. Note that here $\psi_m : \mathcal{H} \to \mathbb{R}$ and

$$\int_{\mathcal{H}} \langle x, x' \rangle \psi_m(x') dR(x') = \gamma_m \psi_m(x),$$

$$\int_{\mathcal{H}} \psi_m(x) \psi_n(x) dR(x) = \delta_{mn}.$$

Due to the centering of η we also have when $\gamma_m \neq 0$,

$$\gamma_m \mathbb{E}_X[\psi_m(x)] = \int_{\mathcal{H}} \mathbb{E}_X[\langle x, x' \rangle_{\mathcal{H}}] \psi_n(x') dR(x') = 0 \quad \Rightarrow \quad \mathbb{E}_X[\psi_m(x)] = 0.$$

Similarly, $\mathbb{E}_Y[\psi_m(y)] = 0$. The V-statistic from the overall sample can now be written as an infinite sum (Serfling, 2009, Section 5.5):

$$\|\hat{C}_{\eta\#P} - \hat{C}_{\eta\#Q}\|_{\mathcal{H}}^2 = \sum_{m=1}^{\infty} \gamma_m \left(\frac{1}{N_1} \sum_{i=1}^{N_1} \psi_m(x_i) - \frac{1}{N_2} \sum_{i=1}^{N_2} \psi_m(y_i) \right)^2 := \sum_{m=1}^{\infty} \gamma_m a_m^2.$$

Our goal is to show that (a) $a_m \rightsquigarrow \mathcal{N}(0, (N\rho_1\rho_2)^{-1})$, for $\forall m$, and (b) a_m and a_n are independent when $m \neq n$.

First note that

$$\mathbb{E}(a_m) = \mathbb{E}\left(\frac{1}{N_1} \sum_{i=1}^{N_1} \psi_m(x_i) - \frac{1}{N_2} \sum_{i=1}^{N_2} \psi_m(y_i) \right) = 0.$$

In addition we have,

$$\begin{aligned}
Cov(a_m, a_n) &= \mathbb{E}(a_m a_n) - \mathbb{E}(a_m).\mathbb{E}(a_n) \\
&= \mathbb{E}(a_m a_n) \\
&= \mathbb{E}\left(\frac{1}{N_1}\sum_{i=1}^{N_1}\psi_m(x_i) - \frac{1}{N_2}\sum_{i=1}^{N_2}\psi_m(y_i)\right)\left(\frac{1}{N_1}\sum_{i=1}^{N_1}\psi_n(x_i) - \frac{1}{N_2}\sum_{i=1}^{N_2}\psi_n(y_i)\right) \\
&= \mathbb{E}_X\left(\frac{1}{N_1^2}\sum_{i=1}^{N_1}\psi_m(x_i)\psi_n(x_i)\right) + \mathbb{E}_Y\left(\frac{1}{N_2^2}\sum_{i=1}^{N_2}\psi_m(y_i)\psi_n(y_i)\right) \\
&= \frac{1}{\rho_1 N}\mathbb{E}_X\left(\frac{1}{N_1}\sum_{i=1}^{N_1}\psi_m(x_i)\psi_n(x_i)\right) + \frac{1}{\rho_2 N}\mathbb{E}_Y\left(\frac{1}{N_2}\sum_{i=1}^{N_2}\psi_m(y_i)\psi_n(y_i)\right) \\
&= \frac{1}{N}\left[\frac{1}{\rho_1}\int_{\mathcal{H}}\psi_m(x)\psi_n(x)d(\tilde{\eta}\#P)(x) + \frac{1}{\rho_2}\int_{\mathcal{H}}\psi_m(y)\psi_n(y)d(\tilde{\eta}\#Q)(y)\right] \\
&= \frac{1}{N\rho_1\rho_2}\int_{\mathcal{H}}\psi_m(z)\psi_n(z)dR(z) \\
&= \frac{1}{N\rho_1\rho_2}\delta_{mn}.
\end{aligned}$$

An application of CLT follows that (a) holds. This together with vanishing covariance proves (b). Consequently, we can apply the CLT for degenerate V-statistics (Serfling, 2009, Section 5.5.2) to obtain the limiting distribution, with $A_m \sim \mathcal{N}(0,1)$,

$$N\|\hat{C}_{\eta\#P} - \hat{C}_{\eta\#Q}\|_{\mathcal{H}}^2 \rightsquigarrow \sum_{m=1}^{\infty}\frac{\gamma_m}{\rho_1\rho_2}A_m^2.$$

Let us now look at the difference between this V-statistic and our U-statistic, i.e. $\hat{\mathbb{T}}$ in (A.5). We see that

$$\begin{aligned}
\|\hat{C}_{\eta\#P} - \hat{C}_{\eta\#Q}\|_{\mathcal{H}}^2 - \hat{\mathbb{T}} &= \frac{1}{N_1^2}\sum_{i,j}\langle x_i, x_j\rangle_{\mathcal{H}} + \frac{1}{N_2^2}\sum_{i,j}\langle y_i, y_j\rangle_{\mathcal{H}} - \frac{2}{N_1 N_2}\sum_{i,j}\langle x_i, y_j\rangle_{\mathcal{H}} \\
&\quad - \frac{1}{N_1(N_1-1)}\sum_{i,j;i\neq j}\langle x_i, x_j\rangle_{\mathcal{H}} + \frac{1}{N_2(N_2-1)}\sum_{i,j;i\neq j}\langle y_i, y_j\rangle_{\mathcal{H}} + \frac{2}{N_1 N_2}\sum_{i,j}\langle x_i, y_j\rangle_{\mathcal{H}} \\
&= -\left[\frac{1}{N_1(N_1-1)} - \frac{1}{N_1^2}\right]\sum_{i,j;i\neq j}\langle x_i, x_j\rangle_{\mathcal{H}} - \left[\frac{1}{N_2(N_2-1)} - \frac{1}{N_2^2}\right]\sum_{i,j;i\neq j}\langle y_i, y_j\rangle_{\mathcal{H}} \\
&\quad + \left(\frac{1}{N_1^2}\sum_{i=1}^{N_1}\|x_i\|_{\mathcal{H}}^2 + \frac{1}{N_2^2}\sum_{i=1}^{N_2}\|y_i\|_{\mathcal{H}}^2\right) \\
&= -K^x - K^y + B.
\end{aligned}$$

We claim that $K^x = O_p(N_1^{-2})$, $K^y = O_p(N_2^{-2})$, and $NB \xrightarrow{P} \sum_{m=1}^{\infty}\gamma_m(\rho_1\rho_2)^{-1}$. As a result,

$$\begin{aligned}
N\left[\|\hat{C}_{\eta\#P} - \hat{C}_{\eta\#Q}\|_{\mathcal{H}}^2 - \hat{\mathbb{T}}\right] &= -NO_p(N_1^{-2}) - NO_p(N_2^{-2}) + \sum_{m=1}^{\infty}\frac{\gamma_m}{\rho_1\rho_2} + o_p(1) \\
&= \sum_{m=1}^{\infty}\frac{\gamma_m}{\rho_1\rho_2} + o_p(1),
\end{aligned}$$

so that $N\hat{\mathbb{T}} \rightsquigarrow \sum_{m=1}^{\infty}\gamma_m(\rho_1\rho_2)^{-1}(A_m^2 - 1)$, and we conclude the proof by reassigning $\gamma_m \leftarrow \gamma_m(\rho_1\rho_2)^{-1}$ to obtain (A.3).

Proof of Claim. For the K-terms we have

$$\begin{aligned} K^x &= \left[\frac{1}{N_1(N_1-1)} - \frac{1}{N_1^2}\right] \sum_{i,j;i\neq j} \langle x_i, x_j\rangle_{\mathcal{H}} \\ &= \frac{1}{N_1^2(N_1-1)} \sum_{i,j;i\neq j} \langle x_i, x_j\rangle_{\mathcal{H}} \\ &= \sum_{m=1}^{\infty} \gamma_m \frac{1}{N_1} \frac{1}{N_1(N_1-1)} \sum_{i,j;i\neq j} \psi_m(x_i)\psi_m(x_j) \\ &= \sum_{m=1}^{\infty} \gamma_m K_m^x, \end{aligned}$$

where K_m^x is defined as the inner sum. Since $\mathbb{E}_X \psi_m(x) = 0$, we have $\mathbb{E}_X(K_m^x) = \frac{1}{N_1}[\mathbb{E}_X \psi_m(x)]^2 = 0$, and

$$\begin{aligned} Var_X(K_m^x) &= \mathbb{E}_X[(K_m^x)^2] \\ &= \frac{1}{N_1^2}\mathbb{E}_X\left[\frac{1}{N_1^2(N_1-1)^2}\sum_{i\neq j}\sum_{l\neq k}\psi_m(x_i)\psi_m(x_j)\psi_m(x_l)\psi_m(x_k)\right] \qquad \text{(A.6)} \\ &= \frac{1}{N_1^2}\mathbb{E}_X\left[\frac{1}{N_1^2(N_1-1)^2}\sum_{i\neq j}\psi_m^2(x_i)\psi_m^2(x_j)\right] \qquad \text{(A.7)} \\ &= \frac{1}{N_1^2}\cdot\frac{1}{N_1(N_1-1)}\left(\mathbb{E}_X[\psi_m^2(x)]\right)^2. \end{aligned}$$

The cross terms—terms involving $l \neq i$ or $k \neq j$—vanish due to the sample being iid and eigenfunctions having zero expectations. The expectation in the last line is finite by assumption (ii), so that $Var_X(K_m^x) = O(N_1^{-4})$, giving $K_m^x = O_p(N_1^{-2})$. Note that the assumption (ii) moreover implies the convergence of the big-oh coefficients, leading to $K^x = \sum_{m=1}^{\infty}\gamma_m K_m^x = O_p(N_1^{-2})$. Similarly we get $K^y = O_p(N_2^{-2})$.

For the term B, we have

$$B = \frac{1}{N_1^2}\sum_{i=1}^{N_1}\|x_i\|_{\mathcal{H}}^2 + +\frac{1}{N_2^2}\sum_{i=1}^{N_2}\|y_i\|_{\mathcal{H}}^2 = \sum_{m=1}^{\infty}\gamma_m\left[\frac{1}{N_1^2}\sum_{i=1}^{N_1}\psi_m^2(x_i) + \frac{1}{N_2^2}\sum_{i=1}^{N_2}\psi_m^2(y_i)\right] := \sum_{m=1}^{\infty}\gamma_m C_m.$$

Taking expectation,

$$\begin{aligned} \mathbb{E}_{X,Y}(C_m) &= \frac{1}{\rho_1 N}\int_{\mathcal{H}}\psi_m^2(x)d(\tilde{\eta}\#P)(x) + \frac{1}{\rho_2 N}\int_{\mathcal{H}}\psi_m^2(y)d(\tilde{\eta}\#Q)(y) \\ &= \frac{1}{N\rho_1\rho_2}\int_{\mathcal{H}}\psi_m^2(z)dR(z) \\ &= \frac{1}{N\rho_1\rho_2}. \end{aligned}$$

Thus $\mathbb{E}_{X,Y}(NB) = \sum_m \gamma_m(\rho_1\rho_2)^{-1}$. Finally,

$$NB = \sum_{m=1}^{\infty}\gamma_m\left[\frac{1}{\rho_1 N_1}\sum_{i=1}^{N_1}\psi_m^2(x_i) + \frac{1}{\rho_2 N_2}\sum_{i=1}^{N_2}\psi_m^2(y_i)\right] \xrightarrow{P} \sum_{m=1}^{\infty}\gamma_m\left[\frac{1}{\rho_1}\mathbb{E}_X\psi_m^2(x) + \frac{1}{\rho_2}\mathbb{E}_Y\psi_m^2(y)\right] = \mathbb{E}_{X,Y}(NB)$$

by the weak law of large numbers. This proves the claim for B.

Alternative Distribution. For the the limiting distribution under H_1, notice that the first two terms in (A.2) are the one-sample U-statistic calculated on the samples $\{\mu_i\}_{i=1}^{N_1}$ and $\{\nu_i\}_{i=1}^{N_2}$, respectively. Using the CLT for non-degenerate

U-statistics (Serfling, 2009, Section 5.5.1, Theorem A), we have

$$\sqrt{N_1}\left[\frac{\sum_{i,j:i\neq j}\langle\eta(\mu_i),\eta(\mu_j)\rangle_{\mathcal{H}}}{N_1(N_1-1)} - \mathbb{E}_{\mu,\mu'\sim P}\langle\eta(\mu),\eta(\mu')\rangle_{\mathcal{H}}\right] \rightsquigarrow N\left(0, 4\mathbb{V}_{\mu\sim P}\left[\mathbb{E}_{\mu'\sim P}\langle\eta(\mu),\eta(\mu')\rangle_{\mathcal{H}}\right]\right),$$
$$\sqrt{N_2}\left[\frac{\sum_{i,j:i\neq j}\langle\eta(\nu_i),\eta(\nu_j)\rangle_{\mathcal{H}}}{N_2(N_2-1)} - \mathbb{E}_{\nu,\nu'\sim Q}\langle\eta(\nu),\eta(\nu')\rangle_{\mathcal{H}}\right] \rightsquigarrow N\left(0, 4\mathbb{V}_{\nu\sim Q}\left[\mathbb{E}_{\nu'\sim Q}\langle\eta(\nu),\eta(\nu')\rangle_{\mathcal{H}}\right]\right).$$

For the third summand, using an equivalent CLT for two-sample U-statistic (Dehling & Fried, 2012, Theorem 2.1),

$$\sqrt{N}\left[\frac{\sum_{i,j}\langle\eta(\mu_i),\eta(\nu_j)\rangle_{\mathcal{H}}}{N_1N_2} - \mathbb{E}_{\mu\sim P,\nu\sim Q}\langle\eta(\mu),\eta(\nu)\rangle_{\mathcal{H}}\right] \rightsquigarrow$$
$$N\left(0, \frac{1}{\rho_1}\mathbb{V}_{\mu\in P}\left[\mathbb{E}_{\nu\sim Q}\langle\eta(\mu),\eta(\nu)\rangle_{\mathcal{H}}\right] + \frac{1}{\rho_2}\mathbb{V}_{\nu\in Q}\left[\mathbb{E}_{\mu\sim P}\langle\eta(\mu),\eta(\nu)\rangle_{\mathcal{H}}\right]\right).$$

We obtain (A.4) by combining the above three results. □

The following result now ensures that approximations of $\hat{\mathbb{T}}$ using the top few eigenfunctions and a finite number of CDF embeddings can be constructed with small approximation errors, provided the manifold eigenvalues are declining suitably fast and the finite dimensional $\eta_D(\cdot)$ is suitably smooth.

Proposition 10. *Suppose that (i), (ii) and (iii) hold. Then we have* $\sqrt{N}(\hat{\mathbb{T}} - \hat{\mathbb{T}}_{L_N}) = o_p(1)$ *and* $\sqrt{N}(\hat{\mathbb{T}}_{L_N} - \tilde{\mathbb{T}}_{L_N,D_N}) = o_p(1)$ *for the following choices of* L_N, D_N*:*

$$L_N \geq \min_{L'}\left\{L' : \sum_{\ell=L'+1}^{\infty}\alpha_\ell\lambda_\ell^{(n+3)/2} \leq \frac{1}{N^{1+\delta}}\right\}, \quad D_N \geq kc^2N^{1+\delta}\sum_{l=1}^{L_N}\alpha_\ell\lambda_\ell^{(n-1)/2},$$

where $\delta, k > 0$ *are constants depending only on* $\mathcal{X}$*.*

As we mention in the discussion after condition (i), for the heat kernel with tuning parameter t: $\alpha(\lambda) = \exp(-t\lambda)$, the assumption (i) that $\sum_{\ell=1}^{\infty}\alpha_l\lambda_\ell^{(n+3)/2} < \infty$ holds. The bound on D_N is a consequence of classical bounds on Riemann sum approximation errors in terms of $\|\eta'\|_\infty$. Absolute continuity of $\mu \sim P, \nu \sim Q$ ensures the existence of $(F^{-1}_{\phi_\ell\sharp\mu})'(s), (F^{-1}_{\phi_\ell\sharp\nu})'(s)$ (where prime denotes the derivative) for Lebesgue-almost every $s \in [0,1]$ (Gavish et al., 2019, Lemma 2.3).

Proof. Notice that given L_N, summands in the expression $\hat{\mathbb{T}} - \hat{\mathbb{T}}_{L_N}$ are the tail sums $\sum_{\ell=L_N+1}^{\infty}\alpha_l\mathcal{W}_2^2(\phi_\ell\sharp\cdot,\phi_\ell\sharp\cdot)$ starting at the L_N+1^{th} term. Using a similar approach as the proof of Proposition 7, this is bounded above by a scalar multiple of the geodesic distance, specifically $c\mathcal{W}_2^{\mathcal{X}}(\cdot,\cdot)\sqrt{\sum_{\ell=L_N+1}^{\infty}\alpha_\ell\lambda_\ell^{(n+3)/2}}$. By assumption $\sum_{\ell=1}^{\infty}\alpha_\ell\lambda_\ell^{(n+3)/2} < \infty$, so that given $\epsilon > 0$ we can always choose a starting point to make the tail sum $< \epsilon$. The choice of L_N follows by taking $\epsilon = N^{-(1+\delta)}$.

To obtain the choice of D_N, we first use a similar approach to the proof of Proposition 9 to simplify $\tilde{\mathbb{T}}_{L,D'}$ for any L, D':

$$\tilde{\mathbb{T}}_{L,D'} = \sum_{\ell=1}^{L}\left[\frac{1}{N_1(N_1-1)}\sum_{i,j:i\neq j}\eta_{D'}(\phi_\ell\sharp\mu_i)^T\eta_{D'}(\phi_\ell\sharp\mu_j) + \frac{1}{N_2(N_2-1)}\sum_{i,j:i\neq j}\eta_{D'}(\phi_\ell\sharp\nu_i)^T\eta_{D'}(\phi_\ell\sharp\nu_j)\right.$$
$$\left. - \frac{2}{N_1N_2}\sum_{i,j}\eta_{D'}(\phi_\ell\sharp\mu_i)^T\eta_{D'}(\phi_\ell\sharp\nu_j)\right]. \tag{A.8}$$

Recall that the inverse CDF transformation induced by $\eta_0(\phi_\ell\sharp\mu) \equiv F^{-1}_{\phi_\ell\sharp\mu}$ maps $[0,1]$ to a bounded interval that is the range of ϕ_ℓ, and $\|\phi_\ell\|_\infty \leq c\lambda_\ell^{(n-1)/4}$ using Hörmander's bound on the supremum norm of the eigenfunctions. Using classical results on Riemann sum approximation errors (Bakhvalov, 2015; Brown & Steinerberger, 2020) we thus have for any ℓ:

$$\left|\alpha_\ell\langle\eta_0(\phi_\ell\sharp\mu),\eta_0(\phi_\ell\sharp\nu)\rangle_{\mathcal{H}} - \eta_{D'}(\phi_\ell\sharp\mu)^T\eta_{D'}(\phi_\ell\sharp\nu)\right| \leq \frac{k}{D'}\alpha_\ell\left\|(F^{-1}_{\phi_\ell\sharp\mu}F^{-1}_{\phi_\ell\sharp\nu})'\right\|_\infty \leq \frac{2kc^2}{D'}\alpha_\ell\lambda_\ell^{(n-1)/2}.$$

Given $L = L_N$, we simply choose $D' = D_N$ large enough to make the right hand side above smaller than $N^{-(1+\delta)}$. While it is possible to make the upper bound tighter using recent results (such as (Brown & Steinerberger, 2020)), the above coarser bound suffices for our purpose. □

We now state a version of Theorem 2, with specifications for $\gamma_m, \sigma_1^2, L_N, D_N$ now available through the above two results.

Theorem 2. *Assume conditions (i)-(iii) hold. Define $N = N_1 + N_2$, and suppose that as $N_1, N_2 \to \infty$, we have $N_1/N \to \rho_1, N_2/N \to \rho_2 = 1 - \rho_1$, for some fixed $0 < \rho_1 < 1$. With $L \geq L_N, D' \geq D_N$ chosen per Proposition 10, under $H_0 : C_{\eta\#P} = C_{\eta\#Q}$ we have*

$$N\tilde{\mathbb{T}}_{L,D'} \rightsquigarrow \sum_{m=1}^{\infty} \gamma_m (A_m^2 - 1),$$

where A_m, γ_m are defined as in Proposition 9. Further, under $H_1 : C_{\eta\#P} \neq C_{\eta\#Q}$ we have $\sqrt{N}\left(\tilde{\mathbb{T}}_{L,D'} - \mathbb{T}\right) \rightsquigarrow N(0, \sigma_1^2)$.

Proof. This a combination of Propositions 9 and 10, and Slutsky's theorem. □

We conclude with a proof of Theorem 3, which gives power guarantee of the test based on $\tilde{\mathbb{T}}_{L,D'}$ for contiguous alternatives.

Theorem 3. *Assume conditions (i)-(iii) hold, and let L, D' be chosen as in Theorem 2. Then for the sequence of contiguous alternatives H_{1N} such that $N\|\delta_N\|_{\mathcal{H}}^2 \to \infty$, the test based on $\tilde{\mathbb{T}}_{L,D'}$ is consistent for any $\alpha \in (0,1)$, that is as $N \to \infty$ the asymptotic power approaches 1.*

Proof. It is enough the prove consistency using $\hat{\mathbb{T}}$, as the difference between $\hat{\mathbb{T}}$ and $\tilde{\mathbb{T}}_{L,D'}$ is negligible by choice of L, D'. To do so we utilize proof techniques similar to Theorem 13 in Gretton et al. (2012). Define $c_N := N^{1/2}\|\delta_N\|_{\mathcal{H}}$, and expand the simplified centered version of the test statistic in (A.5) but under H_1 so that the centering terms do not cancel out:

$$\begin{aligned}\hat{\mathbb{T}}_c =& \frac{1}{N_1(N_1-1)} \sum_{i,j:i\neq j} \langle \eta(\mu_i) - C_{\eta\#P}, \eta(\mu_j) - C_{\eta\#P}\rangle_{\mathcal{H}} \\ &+ \frac{1}{N_2(N_2-1)} \sum_{i,j:i\neq j} \langle \eta(\nu_i) - C_{\eta\#Q}, \eta(\nu_j) - C_{\eta\#Q}\rangle_{\mathcal{H}} \\ &- \frac{2}{N_1 N_2} \sum_{i,j} \langle \eta(\mu_i) - C_{\eta\#P}, \eta(\nu_j) - C_{\eta\#Q}\rangle_{\mathcal{H}} \Bigg].\end{aligned} \tag{A.9}$$

The centered pushforwards have the same Hilbert centroids, thus as $N \to \infty$ by Proposition 9,

$$N\hat{\mathbb{T}}_c \rightsquigarrow \sum_{m=1}^{\infty} \gamma_m (A_m^2 - 1) := S.$$

Subtracting $\hat{\mathbb{T}}_c$ from $\hat{\mathbb{T}}$ and its expansion in Eq. (A.2) on the left and right hand respectively, then simplifying we have

$$\begin{aligned} N(\hat{\mathbb{T}} - \hat{\mathbb{T}}_c) &= N \left[-\frac{1}{N_1}\sum_{i=1}^{N_1} \langle \delta_N, \eta(\mu_i) - C_{\eta\#P}\rangle_{\mathcal{H}} + \frac{1}{N_2}\sum_{i=1}^{N_2} \langle \delta_N, \eta(\nu_i) - C_{\eta\#Q}\rangle_{\mathcal{H}} + \frac{\langle \delta_N, \delta_N\rangle_{\mathcal{H}}}{2} \right] \\ &= N \left[\frac{\|\delta_N\|_{\mathcal{H}}}{N_1} \sum_{i=1}^{N_1} \left\langle \frac{\delta_N}{\|\delta_N\|_{\mathcal{H}}}, \eta(\mu_i) - C_{\eta\#P} \right\rangle_{\mathcal{H}} \right. \\ &\quad \left. - \frac{\|\delta_N\|_{\mathcal{H}}}{N_2} \sum_{i=1}^{N_2} \left\langle \frac{\delta_N}{\|\delta_N\|_{\mathcal{H}}}, \eta(\nu_i) - C_{\eta\#Q} \right\rangle_{\mathcal{H}} + \frac{\|\delta_N\|_{\mathcal{H}}^2}{2} \right]. \end{aligned} \tag{A.10}$$

Given N the inner products $\langle \delta_N/\|\delta_N\|_{\mathcal{H}}, \eta(\mu_i) - C_{\eta\#P}\rangle_{\mathcal{H}}$ are i.i.d. random variables with mean 0, so by CLT then using $\|\delta_N\|_{\mathcal{H}} = c_N N^{-1/2}$ we get

$$\frac{1}{\sqrt{N_1}} \sum_{i=1}^{N_1} \left\langle \frac{\delta_N}{\|\delta_N\|_{\mathcal{H}}}, \eta(\mu_i) - C_{\eta\#P} \right\rangle_{\mathcal{H}} \rightsquigarrow U \quad \Rightarrow \quad \frac{N\|\delta_N\|_{\mathcal{H}}}{N_1} \sum_{i=1}^{N_2} \left\langle \frac{\delta_N}{\|\delta_N\|_{\mathcal{H}}}, \eta(\nu_i) - C_{\eta\#Q} \right\rangle_{\mathcal{H}} \rightsquigarrow \frac{c_N}{\sqrt{\rho_1}} U,$$

where U is the zero mean Gaussian random variable that is the limiting distribution of the above inner product sum. Similarly we have

$$\frac{N\|\delta_N\|_{\mathcal{H}}}{N_2}\sum_{i=1}^{N_2}\left\langle \frac{\delta_N}{\|\delta_N\|_{\mathcal{H}}}, \eta(\nu_i) - C_{\eta\#Q}\right\rangle_{\mathcal{H}} \rightsquigarrow \frac{c_N}{\sqrt{\rho_2}}V,$$

where V is also Gaussian, zero mean, and independent of U. Putting everything together in the right hand side of (A.10), and using $\|\delta_N\|_{\mathcal{H}} = c_N N^{-1/2}$, given the threshold t_α for a level-α test

$$P_{H_N}\left(N\hat{\mathbb{T}} > t_\alpha\right) \to P\left[S + c_N\left(\frac{U}{\sqrt{\rho_1}} - \frac{V}{\sqrt{\rho_2}}\right) + \frac{c_N^2}{2} > t_\alpha\right].$$

By assumption $c_N^2 \to \infty$, so the asymptotic power approaches 1 as $N \to \infty$. □

A.5. Proofs and Notes for Section 5.2

To guarantee size control when using the the harmonic mean p-value we establish a version of Theorem 1 from Liu & Xie (2020). Assume that a test statistic $Z \in \mathbb{R}^D$ has null distribution with zero mean and every pair of coordinates of Z follows bivariate Gaussian distribution. Compute the coordinate-wise two-sided p-values $p_k = 2(1 - \Phi(|Z_k|))$ where Φ is the standard Gaussian CDF.

Theorem 4. *Let $p_k, k = 1, ..., D$ be the null p-values as above and p^H computed via harmonic mean approach, then*

$$\lim_{\alpha\to 0}\frac{\text{Prob}\{p^H \le \alpha\}}{\alpha} = 1.$$

Proof. The proof of Theorem 1 from (Liu & Xie, 2020) hinges on Lemma 3 in their supplemental material. We show that Lemma 3 holds for the harmonic mean combination method. Note that the multiplication by π present in Lemma 3 cancels out when inverse cotangent with a multiplier of $1/\pi$ is applied later on; so it is not relevant to the flow of the proof.

To this end, consider the functions $p(x) = 2(1 - \Phi(|x|))$ and $h(x) = 1/p(x)$. We need to prove the following three statements:

(1) for any $|x| > \Phi^{-1}(3/4)$,

$$\frac{\cos[p(x)\pi]}{p(x)} \le h(x) \le \frac{1}{p(x)}$$

(2) For any constant $0 < |a| < 1$, we have

$$\lim_{x\to+\infty}\frac{h(x)}{x^2 h(ax)} > c_a > 0,$$

where c_a is some constant only dependent on a.

(3) Suppose that X_0 has standard normal distribution, then we have

$$P\{h(X_0) \ge t\} = \frac{1}{t} + O(1/t^3).$$

Statement (1) is trivial, as $h(x) = 1/p(x)$ by definition and the cosine function is upper bounded by one. Statement (2) holds by the same argument as in the supplement of (Liu & Xie, 2020). Statement (3) follows from the fact that when X_0 is standard normal, then $p(x)$ is a null p-value, and so

$$P\{h(X_0) \ge t\} = P\{p(X_0) \le 1/t\} = \frac{1}{t}.$$

Note that there is no $O(1/t^3)$ term at all, but we kept the form of the statement the same as in (Liu & Xie, 2020).

Now, the proof of Theorem 1 from (Liu & Xie, 2020) with weights $\omega_k = 1/D, k = 1, 2, ..., D$ goes through to give

$$P\left\{\frac{1}{D}\sum\frac{1}{p_k} \ge t\right\} = \frac{1}{t} + o(1/t).$$

Note that $p^H = H\left(D/(\frac{1}{p_1}+\frac{1}{p_2}+\cdots+\frac{1}{p_D})\right)$, where the function H has a known form described in (Wilson, 2019) and satisfies $H(x)/x \to 1$ as $x \to 0$. Thus, as $\alpha \to 0$, we have

$$P\left\{p^H \le \alpha\right\} \asymp P\left\{\frac{1}{D}\sum\frac{1}{p_k} \ge 1/\alpha\right\} \asymp \frac{1}{1/\alpha} + o(\frac{1}{1/\alpha}) \asymp \alpha.$$

□

B. Details of numerical experiments

B.1. Synthetic data

We compare the performance of our tests on data from a number of domains with several existing methods, and settings of the embedding parameters L, D'. For evaluation, we use empirical power at different degrees of departure from the null hypothesis, calculated by averaging the proportion of rejections at level $\alpha = 0.05$ over 1000 independent datasets with samples divided into two groups of sizes $n_1 = 60, n_2 = 40$. To ensure the tests are well-calibrated, we also calculate nominal sizes assuming the two sample groups are drawn from the same random meta-distribution. We calculate eigenvalues and eigenfunctions using analytical expressions provided in Table 4. We fix $\alpha(\lambda) = e^{-\lambda}$ (i.e. heat kernel with $t = 1$) for all experiments, in order to avoid unfair advantage from tuning this parameter when comparing to baselines. Also, when using the p-value combination test each t-test scales the mean difference by standard deviation, so the weight functions cancel out anyway. In general, when t is small, the $\hat{\mathbb{T}}$ statistic is more democratic between low and high frequency eigenfunctions, allowing to capture finer details of the underlying domain (assuming large enough L). When t is large, $\hat{\mathbb{T}}$ focuses on low frequency eigenfunctions so more global differences dominate the computation of $\hat{\mathbb{T}}$.

Finite intervals To obtain our base measures μ_i, ν_i, we generate bin probabilities as (shifted and normalized) values of the function $f(t_j) = \mu(t_j) + \alpha(t_j)$ at $m = 30$ fixed design points $t_j = j/(m+1), j = \{1, 2, \ldots, m\}$, and

$$\begin{aligned}\mu(t_j) &= 1.2 + 2.3\cos(2\pi t_j) + 4.2\sin(2\pi t_j),\\ \alpha(t_j) &= \epsilon_0 + \sqrt{2}\epsilon_1\cos(2\pi t_j) + \sqrt{3}\epsilon_2\sin(2\pi t_j),\end{aligned}$$

where $\epsilon_0, \epsilon_1, \epsilon_2 \sim N(0,1)$ clipped between $[-3, 3]$. Group 1 and 2 samples are obtained as $\mu_i(\cdot) \equiv f(\cdot)$ and $\nu_i(\cdot) \equiv f(\cdot) + \delta$ respectively, where $\delta \in [0, 4]$ is a constant. To make the sample functions non-negative, we shift all functions by $M = 3(1 + \sqrt{2} + \sqrt{3})$. Finally, as the m-length vector of bin counts for a sample, we generate a random vector from the Multinomial distribution with 1000 trials, m outcomes and the outcome probabilities proportional to the shifted functional observations corresponding to that sample.

We use embedding dimensions $L = 3, D' = 10$ to compare our method against 11 functional ANOVA tests—for brevity we report results for 3 of them which use different methodological approaches (see Appendix for complete results). All methods maintain nominal size for $\delta = 0$ (Figure 3 a). While the combination test (ISD comb) based on our proposal outperformed all the other tests across all values of δ, the bootstrap test that uses the overall $\mathbb{T}$ statistic (ISD T boot) performs better than Fmaxb but worse than others. Table 5 shows the outputs for the other 8 competing methods from the R package `fdANOVA` for the finite intervals synthetic data setting[1].

We also compare the p-value combination test based on an *unsliced* 24-dimensional inverse CDF embedding with sliced ISW_2-based tests (Figure 3 b). We use multiple pairs of (L, D') values, all of them giving overall embeddings of dimension $D = LD' = 24$. The performance of an ISW_2-based test that uses slicing over only the first eigenfunction is almost as good as the unsliced version. With more eigenfunctions, the powers first improve considerably, then become similar to the unsliced version again.

Manifold domains We consider data from distributions on circles and cylinders. For circular data, we take von Mises distributions with randomly chosen parameters as our samples. For an angle x (measured in radians), the von Mises probability density function is given by $f(x|\mu,\kappa) = \exp[\kappa\cos(x-\mu)](2\pi I_0(\kappa))^{-1}$, where $I_0(\kappa)$ is the modified Bessel function of order 0. We fix $\kappa = 2$, and use $\mu \equiv \mu_i \sim N(0, 0.1^2)$, $\mu \equiv \nu_i \sim N(\delta, 0.1^2)$ for samples from group 1 and 2

[1] See https://www.rdocumentation.org/packages/fdANOVA/versions/0.1.2/topics/fanova.tests for full names of all methods.

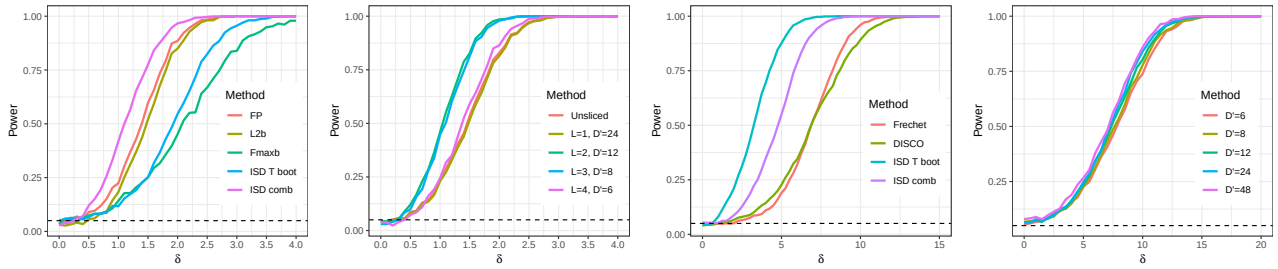


Figure 4. Performance on synthetic finite interval and manifold data. Finite interval: (a) comparison with existing methods—a test based on basis function representation (FP) (Gorecki & Smaga, 2015), a sum-type ℓ_2 norm-based test (L2b) (Zhang, 2013), and a max-type test (Zhang et al., 2019) that uses the maximum of coordinate-wise F statistic (Fmaxb); (b) unsliced vs. different settings of (L, D'). Manifold data: (c) circular data, comparing with Fréchet ANOVA (Dubey & Müller, 2019), and the DISCO nonparametric test (Rizzo & Székely, 2010); (d) harmonic combination tests on cylindrical data for $L = 4$. Dotted lines indicates nominal size of all tests ($\alpha = 0.05$).

δ	CH	CS	L2N	L2b	FN	FB	Fb	GPF
0	0.031	0.03	0.033	0.024	0.031	0.028	0.033	0.026
0.1	0.025	0.024	0.03	0.044	0.027	0.03	0.041	0.021
0.2	0.026	0.029	0.037	0.06	0.033	0.034	0.058	0.025
0.3	0.036	0.041	0.044	0.067	0.041	0.04	0.067	0.033
0.4	0.034	0.035	0.036	0.057	0.034	0.035	0.056	0.032
0.5	0.051	0.052	0.058	0.091	0.056	0.057	0.088	0.044
0.6	0.056	0.066	0.066	0.089	0.061	0.066	0.088	0.051
0.7	0.07	0.083	0.083	0.121	0.084	0.081	0.119	0.064
0.8	0.085	0.097	0.095	0.151	0.093	0.094	0.144	0.081
0.9	0.118	0.142	0.14	0.2	0.144	0.137	0.194	0.118
1	0.158	0.182	0.176	0.232	0.183	0.173	0.228	0.154
1.1	0.215	0.247	0.246	0.303	0.251	0.242	0.301	0.212
1.2	0.27	0.31	0.303	0.375	0.311	0.3	0.368	0.27
1.3	0.328	0.363	0.357	0.438	0.37	0.353	0.43	0.324
1.4	0.395	0.432	0.432	0.504	0.436	0.423	0.499	0.394
1.5	0.488	0.52	0.514	0.592	0.521	0.511	0.586	0.483
1.6	0.534	0.595	0.576	0.652	0.593	0.566	0.647	0.544
1.7	0.628	0.677	0.669	0.723	0.678	0.661	0.719	0.631
1.8	0.704	0.737	0.727	0.789	0.748	0.725	0.785	0.707
1.9	0.785	0.823	0.812	0.869	0.827	0.806	0.867	0.793
2	0.83	0.849	0.844	0.88	0.85	0.841	0.875	0.832
2.1	0.865	0.888	0.881	0.916	0.887	0.878	0.915	0.872
2.2	0.903	0.922	0.916	0.946	0.928	0.912	0.946	0.907
2.3	0.938	0.95	0.944	0.964	0.951	0.944	0.963	0.944
2.4	0.958	0.973	0.967	0.977	0.972	0.966	0.976	0.964
2.5	0.974	0.98	0.976	0.985	0.981	0.975	0.985	0.974
2.6	0.977	0.981	0.979	0.987	0.981	0.978	0.986	0.977
2.7	0.989	0.996	0.992	0.997	0.996	0.992	0.997	0.991
2.8	0.997	0.998	0.997	0.998	0.998	0.997	0.998	0.996
2.9	0.996	0.997	0.996	0.999	0.997	0.996	0.999	0.997
3	0.998	1	0.999	1	1	0.999	1	0.999

Table 5. Outputs for other methods in the functional curves synthetic data setting.

respectively—with $\delta \in [0, 15] \times \pi/180$ (i.e. 0 to 15 degrees converted to radians). As each observation vector, we take 100 random draws from each sample-specific distribution. For our embeddings, we use $L = 10, D' = 20$, and so our final embedding dimension is $10 \times 20 \times 2 = 400$. Since the competing methods cannot handle circular geometry directly, to

implement them we cut the circle into an interval. Figure 3 (c) shows that all methods maintain nominal size, but both our tests maintain considerably higher power than existing methods for all δ.

We generate cylindrical data in the form of samples of a bivariate random vector (Θ, X), using the cylindrical density function proposed by (Mardia & Sutton, 1978):

$$f(\theta, x) = \frac{e^{\kappa \cos(\theta - \mu)}}{2\pi I_0(\kappa)} \frac{1}{\sqrt{2\pi}\sigma_c} e^{-\frac{(x-\mu_c)^2}{2\sigma_c^2}},$$

clipping values of the X-coordinate between the bounded interval $[0, 2\pi]$. This distribution has the parameters $\mu \in [-\pi, \pi], \mu_0 \in \mathbb{R}, \kappa \geq 0, \rho_1 \in [0, 1), \rho_2 \in [0, 1), \sigma > 0$, where μ, κ denote parameters for the (circular) marginal along the Θ-coordinate. and given $\Theta = \theta$, X is sampled from $N(\mu_c, \sigma_c^2)$, with

$$\begin{aligned} \mu_c &= \mu + \sqrt{\kappa}\sigma \left\{\rho_1(\cos\theta - \cos\mu) + \rho_2(\sin\theta - \sin\mu)\right\}, \\ \sigma_c &= \sigma^2(1 - \rho^2), \rho = (\rho_1^2 + \rho_2^2)^{1/2}. \end{aligned}$$

In our experiments, we fix $\rho_1 = \rho_2 = 0.5, \sigma = 1, \kappa = 2$ across both populations. As random samples of distributions, we draw $\mu, \mu_0 \sim \text{Unif}(0, 1)$ and $\mu, \mu_0 \sim \text{Unif}(\delta, \delta + 1)$ for samples of group 1 and 2 respectively, with $n_1 = 60, n_2 = 40$. We repeat the above for $\delta \in [0, 30]$ degrees converted to radians, and obtain bivariate histograms corresponding to each sample distribution from 500 random draws from that distribution. To evaluate the effects of choosing L, D' we calculate our embeddings for $L \in \{2, 3, 4, 5\}, D' \in \{6, 8, 12, 24, 48\}$. The choice of L has small effect on performance, so we report results for $L = 4$ in Figure 3 (d). Higher values of D' result in some increase in power.

Discussion Our ISW_2-based method is able to exploit the non-euclidean nature of the problems and and their generality beyond mean comparison more effectively than competing methods, which are based on mean comparison on functional data/densities (frechet ANOVA, all functional ANOVA methods), and/or L2 distance-based comparisons (all functional ANOVA methods, DISCO). Regarding the optimal choice of embedding dimensions, while proving theorem 2 we show that (Proposition 10 therein) choosing both L and D above certain thresholds ensures close approximation to the population test statistic. For the combination test, adding more dimensions to the embedding can have a two-fold effect: a) probing more dimensions can help with finding differences, but b) every dimension adds another test and so potentially leads to loss of power. Thus, for the combination test, there must be an optimal data dependent choice of the embedding dimension, which can potentially be found via split testing procedures. We leave this to future work.

Computational Complexity Assume an underlying graph $G = (V, E)$, and we use L slices, D' quantiles to calculate ISW_2. Each distribution consists of V atoms (distribution lives on graph vertices). Then, computation of our embeddings includes the following steps:

1. The computation of eigenvectors/values of symmetric sparse matrices is a well-studied problem with stable and efficient algorithms available, e.g. ARPACK providing an implementation of the Arnoldi method in MATLAB, Python, or R. These methods incur linear time complexity in the size of graph: $O(L(|V| + |E|)$. This computation is done only once per domain, and the overhead is negligible ($<<$ 1 second) for our graph experiments.

2. Hilbert embedding computations require computing L pushforwards and D' quantiles, which need sorting. This complexity is $O(L(|V| \log |V| + |D'|)$.

3. Testing using p-value combination requires computing t-test p-values for LD' dimensions, with overall complexity of $O(LD'N)$ where N is total sample size.

In contrast, Sobolev Transport (ST) requires computing the Sliced ST Distance, based on computing shortest paths from randomly selected nodes in a graph. This has complexity $O(k(|V| \log |V| + |E|)$ (Le et al., 2022, pg. 4) Implementing permutation tests means repeating the above computation P times, with $P = 1000$ being a typical choice.

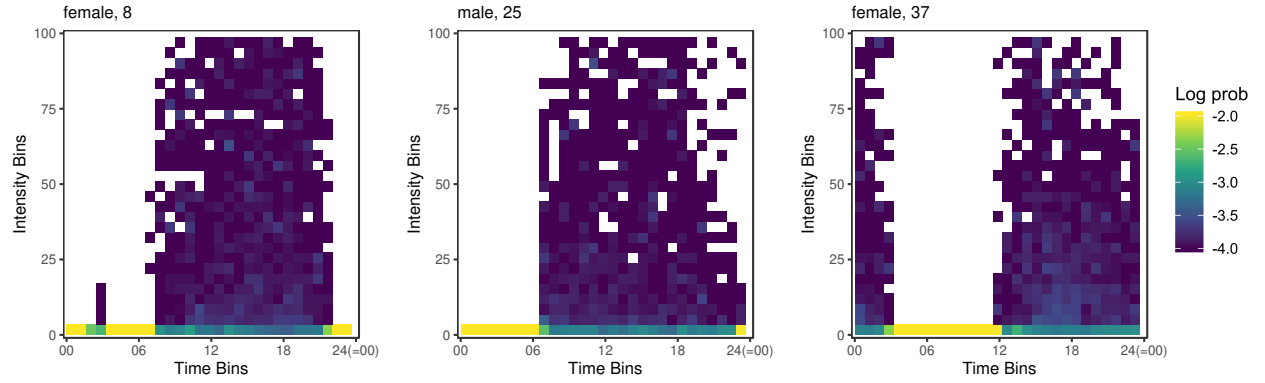


Figure 5. Activity histograms for three individuals from NHANES dataset. There are 100 bins in the intensity and 96 in the time dimension; we show hour of day on the time axis. The time dimension is periodic where 00:00 is identified with 24:00, giving rise to a cylindrical histogram domain.

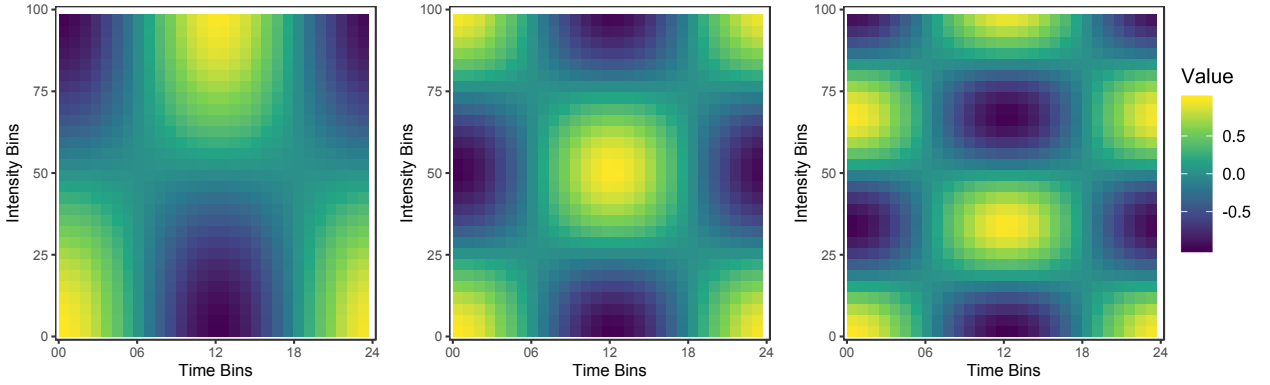


Figure 6. Three eigenfunctions for the NHANES histogram domain normalized by the maximum absolute value. Note that the eigenfunctions are periodic in the time direction (i.e. match when glued over the side cut) but not in the intensity direction, reflecting the cylindrical geometry of the underlying domain.

B.2. NHANES data on physical activity monitoring

As our first real data application, we analyze the Physical Activity Monitor (PAM) data from the 2005-2006 National Health and Nutrition Examination Survey (NHANES)[2]. This contains physical activity pattern readings for a large number of people collected over 1 week period on a per-minute granularity. After basic pre-processing steps to ensure no missing entries, as well as data reliability and well-calibrated activity monitors, we use data from 6839 individuals. The data for each individual corresponds to device intensity value from the PAM for $24 \times 60 = 1440$ minutes throughout the day, for 7 days.

For each individual we can capture their activity patterns into a cylindrical histogram with time and intensity dimensions. For each observation, its time during the day is discretized into 15-minute intervals giving 96 bins for the time dimension; its intensity value (capped at 1,000) is discretized into a 100 equidistant bins. Since the time dimension is periodic, we obtain a histogram over the cylinder $S^1(T_1) \times [0, T_2)$, with $T_1 = 96, T_2 = 100$. Normalized counts can thus be considered as person-specific probability distributions; several examples are shown in Figure 5. Note that flattening the domain by cutting the cylinder will arbitrarily split activity patterns (see especially Figure 5, Female 37) and will lead to inefficiencies due to horizontal variability.

We apply the proposed methodology to check if the activity patterns vary across different groups of individuals obtained as follows. We first split the overall dataset based on the individual's age using the following inclusive ranges: 6–15, 16–25, ...,76–85; this covers all the ages in the dataset. From each split we sample 100 males and 100 females to avoid gender imbalance driving the results. Thus, we end up with 8 age groups with 200 individuals per group. Our goal is to compare

[2] https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/PAXRAW_D.htm

Age Groups	6–15	16–25	26–35	36–45	46–55	56–65	66–75	76–85
6–15		0.979	0.31	0.383	0.297	0.905	0.921	0.326
16–25	**3.7e-11**		0.998	0.963	0.443	0.872	0.442	0.529
26–35	**4.6e-20**	**1.0e-05**		0.987	0.818	0.93	0.731	0.992
36–45	**3.2e-26**	**3.5e-11**	**0.01**		0.945	0.984	0.974	0.327
46–55	**6.6e-27**	**8.4e-16**	**0.002**	0.377		0.832	0.618	0.844
56–65	**2.4e-32**	**7.5e-20**	**3.1e-04**	**0.042**	0.977		0.509	0.98
66–75	**5.4e-45**	**1.6e-16**	**7.7e-06**	**1.6e-04**	**0.001**	**0.011**		0.557
76–85	**3.4e-52**	**1.4e-23**	**1.4e-15**	**2.7e-12**	**9.7e-16**	**1.4e-09**	**2.1e-06**	

Table 6. Comparing the activity intensity of different age groups based on the NHANES dataset. Below diagonal: p-values corresponding to the actual data comparisons. Above diagonal: null p-values obtained by combining and randomly splitting the two involved groups. The entries in boldface correspond to the rejected hypotheses with the BH procedure at the FDR level of 0.1.

these 8 groups' activity patterns by conducting pair-wise tests.

To perform our analysis we compute the eigenvalues and eigenfunctions as per the 4th row of Table 4 using $\ell_1 = 1, 2, 3$ and $\ell_2 = 1, 2, 3$, giving a total of $L = 2 \times 3 \times 3 = 18$ eigenfunctions; three of the resulting eigenfunctions are shown in Figure 6. We consider a $D' = 5$ dimensional embedding for the inverse CDF transformation, hence the final embedding dimension after the slicing construction is $D = LD' = 18 \times 5 = 90$.

We summarize the results in Table 6, *below the diagonal*. The p-values are obtained via the harmonic mean combination approach. We run the Benjamini-Hochberg (Benjamini & Hochberg, 1995) procedure on the resulting p-values at the false discovery rate of 0.1, and the rejected hypotheses are indicated by the p-values in bold. Our method detects statistically significant differences between all pairs of groups, except 46–55 versus 36–45 and 56-65 groups. As a control experiment, we provide our method with null cases and display the p-values in Table 6, *above the diagonal*. The null cases are obtained by combining the individuals from the two comparison groups and splitting it arbitrarily (i.e. mixing the two age groups). As expected, the p-values of the control comparisons do not concentrate near zero.

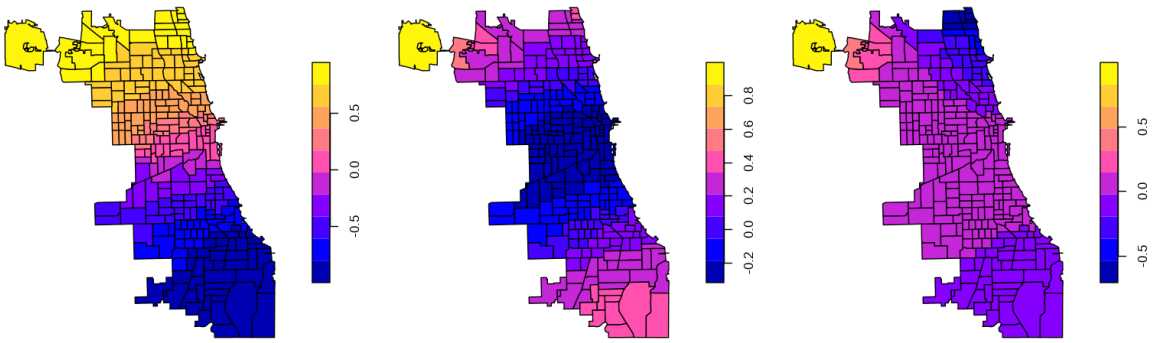
Curiously, our method can be used "off-label" to conduct *functional data analyses* over different dimensions of the NHANES dataset. For example, one can concentrate on a single day of activity intensity data which gives a curve over the 24-hour circle. Since activity intensity is a non-negative number, these curves can be normalized so as to obtain probability distributions. Now we can use our methodology to detect pair-wise differences across groups. While this has the benefit of accounting for underlying geometry of data, it loses the absolute magnitude information due to the normalization. Clearly the appropriateness of such an analysis would depend on the goal of the exercise and the particular research question attached to that goal; our proposal provides a framework that is flexible enough to handle data of different modalities.

B.3. Chicago Crime

We demonstrate the use of our methodology on histograms over graphs. In this experiment, we use the Chicago Crimes 2018 dataset[3] which captures incidents of crime in the City of Chicago. We base our analysis on the type of crime, the beat (geographic area subdivision used by police, see Figure 7) where the incident took place, and the date of the incident. To capture the spatial aspect of the data we build a graph with one vertex per beat; two vertices are connected by an edge if the corresponding beats share a geographic boundary. For each crime type and day, we capture the total count of that crime type for each beat; after normalizing this gives a daily probability distribution over the graph. Our goal is to compare the collection of distributions of, say, theft occurring on Tuesday to those of Thursday and Saturday. The Tuesday versus Thursday comparison is intended as a null case, as we do not expect to see any differences between them (Rustamov & Klosowski, 2020).

We build the un-normalized Laplacian of the beat adjacency graph, and compute its lowest frequency $L = 20$ eigenvalues and eigenvectors. The first three eigenvectors are plotted in Figure 7. The number of inverse CDF values used in the embedding is $D' = 5$, which gives rise to $D = 100$ dimensional embedding. The results of comparisons are shown in the last two columns of Table 7; the p-values are obtained via the harmonic mean combination approach. We run the Benjamini-Hochberg (Benjamini & Hochberg, 1995) procedure on the 20 resulting p-values at the false discovery rate of 0.1, and the rejected hypotheses are indicated by the p-values in bold. As expected, no differences were detected between

[3] data.cityofchicago.org



(a) $\phi_1(\cdot)$ (b) $\phi_2(\cdot)$ (c) $\phi_3(\cdot)$

Figure 7. First three eigenvectors of the Laplacian are shown for the beat adjacency graph, mapped back to the geographic locations. All of the eigenvectors are normalized by the maximum absolute value. The spatial smoothness of the eigenvectors—somewhat masked here due to the discrete colormap—is crucial to efficiently capturing horizontal variability of the data (i.e. distribution shifts over the graph). The boundaries of beats are shown based on the shape file from Chicago Data portal.

Crime Type	Tuesday		Thursday		Saturday		Tue vs Thu	Tue vs Sat
	N	$\overline{\text{count}}$	N	$\overline{\text{count}}$	N	$\overline{\text{count}}$	p-value	p-value
Theft	52	178.7	52	182.9	52	180.2	0.452	**4.7e-06**
Deceptive Practice	51	55.8	52	54.9	52	44.4	0.255	**4.2e-04**
Battery	52	125.8	52	123.0	52	154.9	0.374	**0.001**
Robbery	50	25.2	50	25.1	52	28.1	0.130	**0.002**
Narcotics	51	36.0	51	34.6	50	36.9	0.890	**0.008**
Criminal Damage	52	70.0	52	73.7	52	83.0	0.901	**0.03**
Other Offense	52	49.5	52	48.4	52	44.1	0.670	0.037
Burglary	52	34.0	52	33.1	52	29.1	0.157	0.183
Motor Vehicle Theft	52	27.9	52	26.2	51	28.1	0.923	0.365
Assault	52	57.2	52	59.3	52	52.4	0.996	0.617

Table 7. Results on Chicago Crime 2018 dataset. The entries in bold correspond to the rejected hypotheses with the BH procedure at the FDR level of 0.1. The N column captures the number of days passing the filtering criteria, and the $\overline{\text{count}}$ column shows the average per-day crime count.

Tuesday and Thursday patterns. On the other hand, we see that there are statistically significant differences between Tuesday and Saturday patterns in the following categories of crime: theft, deceptive practice, battery, robbery, narcotics, and criminal damage.

B.4. Brain Connectomics

In this example, we consider two publicly available brain connectomics datasets (Aine et al., 2017; Arroyo Relión et al., 2019) distributed as a part of the R package `graphclass`[4]. Both are based on resting state functional magnetic resonance imaging (fMRI): COBRE has data on 54 schizophrenics and 70 controls, and UMich with 39 schizophrenics and 40 controls. The datasets capture the pairwise correlations between 264 regions of interest (ROI) of Power parcellation (Power et al., 2011) and can be considered as a 264 node graph (263 nodes for COBRE as ROI 75 is missing) with positive and negative edge weights.

We define three probability measures supported on the nodes of the graph. For each ROI we take the sum of absolute values of all its correlations with the remaining ROIs. Now we have a positive number assigned to each node capturing its overall connectivity to the rest of the graph and we normalize to obtain a measure; this construction will be referred to as "all correlations". Note that each scanned subject gives rise to a separate "all correlations" probability measure on the same

[4] http://github.com/jesusdaniel/graphclass

Dataset	All correlations	Positive correlations	Negative correlations
COBRE	0.0084	0.00019	0.0019
UMich	0.609	0.116	0.022

Table 8. Comparison results between the schizophrenic and control groups for brain connectomics datasets.

underlying node set. The "positive correlations" and "negative correlations" constructions are based on keeping respectively only positive or only negative correlations and aggregating as above.

We also need a fixed base graph for the computation of the Laplacian eigen-decomposition; this graph should capture the spatial connectivity of the ROIs which is relevant due to the smooth nature of the blood oxygenation level dependent (BOLD) signal that is used for computing the correlations. To this end, we obtain the coordinates for the centers of the 264 ROIs[5] and build the base graph by connecting each ROI to its nearest 8 ROIs.We compute the lowest frequency $L = 20$ eigenvalues and eigenvectors of the corresponding un-normalized Laplacian. The number of inverse CDF values used in the embedding is $D' = 5$, which gives rise to $D = 100$ dimensional embedding.

Table 8 shows the result of comparing the schizophrenic group to the control group for both of the datasets; the p-values are obtained via the harmonic mean combination approach. We can see that our approach detects statistically significant differences between the two groups in COBRE dataset in all of the three types of measures on graphs. In contrast, for UMich dataset, the difference is detected only in the negative correlations and loses significance when corrected for multiple testing. This is potentially caused by the higher inhomogeneity of the UMich dataset that was pooled across five different experiments spanning seven years (Arroyo Relión et al., 2019). An interesting aspect of our analysis is that due to normalization (to obtain probability measures) the total sum of connectivity is factored out by the proposed method. As a result, the detected differences are not related to the well-known change in the overall connectivities between the two groups, but rather to distributional changes in marginal connectivity strengths.

[5] www.jonathanpower.net/2011-neuron-bigbrain.html