

A NONPARAMETRIC TWO-SAMPLE HYPOTHESIS TESTING PROBLEM FOR RANDOM DOT PRODUCT GRAPHS

BY MINH TANG, AVANTI ATHREYA, VINCE LYZINSKI
DANIEL L. SUSSMAN AND CAREY E. PRIEBE

Johns Hopkins University and Harvard University

We consider the problem of testing whether two finite-dimensional random dot product graphs have generating latent positions that are independently drawn from the same distribution, or distributions that are related via scaling or projection. We propose a test statistic that is a kernel-based function of the adjacency spectral embedding for each graph. We obtain a limiting distribution for our test statistic under the null and we show that our test procedure is consistent across a broad range of alternatives.

1. Introduction. The nonparametric two-sample hypothesis testing problem involves

$$\{X_i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} F, \quad \{Y_k\}_{k=1}^m \stackrel{\text{i.i.d.}}{\sim} G; \quad \mathbb{H}_0: F = G \quad \text{against} \quad \mathbb{H}_A: F \neq G$$

where F and G are two distributions taking values in \mathbb{R}^d . This is a classical problem and there exist a large number of test statistics $T(\{X_i\}_{i=1}^n, \{Y_k\}_{k=1}^m)$ that are consistent for any arbitrary distributions F and G .

In this paper, we consider a related problem that arises naturally in the context of inference on random graphs. That is, suppose that the $\{X_i\}_{i=1}^n$ and $\{Y_k\}_{k=1}^m$ are *unobserved*, and we observe instead adjacency matrices \mathbf{A} and \mathbf{B} corresponding to random dot product graphs on n and m vertices with latent positions $\{X_i\}_{i=1}^n$ and $\{Y_k\}_{k=1}^m$, respectively. Denoting by $\{\hat{X}_i\}_{i=1}^n$ and $\{\hat{Y}_k\}_{k=1}^m$ the adjacency spectral embedding of \mathbf{A} and \mathbf{B} , we construct $T(\{\hat{X}_i\}_{i=1}^n, \{\hat{Y}_k\}_{k=1}^m)$ for testing $F = G$ (and related hypothesis) that is consistent for a broad collection of distributions.

In other words, we construct a test for the hypothesis that two random dot product graphs have the same underlying distribution of latent positions, or underlying distributions that are related via scaling or projection. This formulation includes, as a special case, a test for whether two graphs come from the same stochastic blockmodel or from the same degree-corrected stochastic blockmodel. This problem may be viewed as the nonparametric analogue of the semiparametric inference problem considered in [30], in which a valid test is given for the hypothesis that two random dot product graphs have the same fixed latent positions.

The test statistic we construct is an empirical estimate of the maximum mean discrepancy of [13]. The maximum mean discrepancy in this context is equivalent to an L_2 -distance between kernel density estimates of distributions of the latent positions (see e.g. [2]). The test statistic can

*Supported in part by National Security Science and Engineering Faculty Fellowship (NSSEFF), Johns Hopkins University Human Language Technology Center of Excellence (JHU HLT COE), and the XDATA program of the Defense Advanced Research Projects Agency (DARPA) administered through Air Force Research Laboratory contract FA8750-12-2-0303.

AMS 2000 subject classifications: Primary 62G10; secondary 62H12, 05C80, 60B20

Keywords and phrases: nonparametric graph inference, two-sample hypothesis testing, random dot product graph

also be framed as a weighted L_2 -distance between empirical estimates of characteristic functions similar to those of [3, 10, 14]. Indeed, techniques for the estimation and comparison of densities or characteristic functions given i.i.d data are well-known. We strongly emphasize, however, that in our case, the observed data are *not the true latent positions*—which are themselves random and drawn from the unknown distributions whose equality we wish to test—but rather the adjacency matrices of the resulting random dot product graphs. Thus one of our main technical contributions is the demonstration that functions of the true latent positions are well-approximated by functions of the adjacency spectral embeddings. We remark that our perspective on kernel-based testing for graphs differs significantly from [5, 13]. In [5, 13], the kernel statistics are computed on the graphs, using, for example, the random walk graph kernels [32]. Unfortunately, there is no known computationally tractable kernel that is also characteristic for general distributions on graphs. Indeed, the existence of such a kernel implies a polynomial time algorithm for subgraph isomorphism, a NP-complete problem [11]. However, in this paper we restrict ourselves to the family of distributions of latent positions for random dot product graphs, for which the computation of kernel statistics on the adjacency spectral embedding yields a characteristic kernel.

We organize the paper as follows. In Section 2, we recall the definition of a random dot product graph and the adjacency spectral embedding; we review the relevant background in kernel-based hypothesis testing; and we formulate three specific nonparametric two-sample tests for random dot product graphs. In Section 3, we propose test procedures in which the test statistics are a function of the adjacency spectral embedding. Our test procedures are valid and consistent, and we derive the asymptotic distributions of our test statistics in each case. In Section 4, we illustrate our test procedures with experimental results on simulated and real data. All proofs are presented in the appendix of the paper.

2. Background and Setting. We first recall the notion of a random dot product graph [35].

DEFINITION 1. Let Ω be a subset of \mathbb{R}^d such that, for all $\omega, \omega' \in \Omega$, $0 \leq \langle \omega, \omega' \rangle \leq 1$. For any given $n \geq 1$, let $\mathbf{X} = [X_1, X_2, \dots, X_n]^T$ be a $n \times d$ matrix whose rows are arbitrary elements of Ω . Given \mathbf{X} , suppose \mathbf{A} is a random $n \times n$ adjacency matrix with probability

$$\mathbb{P}[\mathbf{A} | \{X_i\}_{i=1}^n] = \prod_{i \leq j} (X_i^T X_j)^{\mathbf{A}_{ij}} (1 - X_i^T X_j)^{1 - \mathbf{A}_{ij}}.$$

\mathbf{A} is then said to be the adjacency matrix of a *random dot product graph* (RDPG) with *latent positions* \mathbf{X} and we denote this by $\mathbf{A} \sim \text{RDPG}(\mathbf{X})$. Now suppose that the rows of \mathbf{X} are not fixed, but are instead independent random variables sampled according to some distribution F on Ω . Then \mathbf{A} is said to be the adjacency matrix of a *random dot product graph* with *latent positions* \mathbf{X} *sampled according to* F and we denote this by writing $(\mathbf{X}, \mathbf{A}) \sim \text{RDPG}(F)$. We shall also write $\mathbf{A} \sim \text{RDPG}(F)$ when the dependency of \mathbf{A} on \mathbf{X} is integrated out.

Given a matrix of latent positions \mathbf{X} , the random dot product model generates a symmetric adjacency matrix \mathbf{A} whose edges $\{\mathbf{A}_{ij}\}_{i < j}$ are independent Bernoulli random variables with parameters $\{\mathbf{P}_{ij}\}_{i < j}$, where $\mathbf{P} = \mathbf{X}\mathbf{X}^T$. Random dot product graphs are a specific example of *latent position graphs* [16], in which each vertex is associated with a latent position and, conditioned on the latent positions, the presence or absence of the edges in the graph are independent. The edge presence probability between two vertices is given by a symmetric link function of the latent positions of the associated vertices. A random dot product graph on n vertices is also, when viewed as an induced subgraph of an infinite graph, an example of an *exchangeable random graph* [8]. The notion

of random dot product graphs is related to the notion of stochastic block model graphs [17] and degree-corrected stochastic block model graphs [18], as well as mixed membership block models [1]; for example, a stochastic block model graph with K blocks and a positive semidefinite block probability matrix \mathbf{B} corresponds to a random dot product graph whose latent positions are drawn from a mixture of K point masses.

REMARK. We note that non-identifiability is an intrinsic property of random dot product graphs. Indeed, for any matrix \mathbf{X} and any orthogonal matrix \mathbf{W} , the inner product between any rows i, j of \mathbf{X} is identical to that between the rows i, j of \mathbf{XW} . Hence, for any probability distribution F on Ω and unitary operator U , the adjacency matrices $\mathbf{A} \sim \text{RDPG}(F)$ and $\mathbf{B} \sim \text{RDPG}(F \circ U)$ are identically distributed.

ASSUMPTION 1. We assume that the dimension d of the latent positions X is known and that the distribution F for the latent positions X is such that the second moment matrix $\mathbb{E}[XX^T]$ has d distinct eigenvalues.

We now define the notion of adjacency spectral embedding.

DEFINITION 2. The *adjacency spectral embedding* of \mathbf{A} into \mathbb{R}^d is given by $\hat{\mathbf{X}} = \mathbf{U}_\mathbf{A} \mathbf{S}_\mathbf{A}^{1/2}$ where

$$|\mathbf{A}| = [\mathbf{U}_\mathbf{A} | \tilde{\mathbf{U}}_\mathbf{A}] [\mathbf{S}_\mathbf{A} \oplus \tilde{\mathbf{S}}_\mathbf{A}] [\mathbf{U}_\mathbf{A} | \tilde{\mathbf{U}}_\mathbf{A}]$$

is the spectral decomposition of $|\mathbf{A}| = (\mathbf{A}^T \mathbf{A})^{1/2}$ and $\mathbf{S}_\mathbf{A}$ is the matrix of the d largest eigenvalues of $|\mathbf{A}|$ and $\mathbf{U}_\mathbf{A}$ is the matrix whose columns are the corresponding eigenvectors.

2.1. *Hypothesis tests.* In this paper we propose nonparametric versions of the two-sample hypothesis tests examined in [30]. To wit, [30] presents a two-sample random dot product graph hypothesis test as follows. Let \mathbf{X}_n and \mathbf{Y}_n be $n \times d$ matrices of fixed (non-random) latent positions, and $\mathcal{O}(d)$ the collection of orthogonal matrices in $\mathbb{R}^{d \times d}$. Suppose $\mathbf{A} \sim \text{RDPG}(\mathbf{X}_n)$ and $\mathbf{B} \sim \text{RDPG}(\mathbf{Y}_n)$ are the adjacency matrices of random dot product graphs with latent positions \mathbf{X}_n and \mathbf{Y}_n , respectively. Consider the sequence of hypothesis tests

$$\begin{aligned} & H_0^n: \mathbf{X}_n \perp \mathbf{Y}_n \\ \text{against } & H_a^n: \mathbf{X}_n \not\perp \mathbf{Y}_n \end{aligned}$$

where \perp denotes that there exists an $\mathbf{W} \in \mathcal{O}(d)$ such that $\mathbf{X}_n = \mathbf{Y}_n \mathbf{W}$. In [30], it is shown that rejecting for large values of the test statistic T_n defined by

$$T_n = \min_{\mathbf{W} \in \mathcal{O}(d)} \|\hat{\mathbf{X}}_n \mathbf{W} - \hat{\mathbf{Y}}_n\|_F,$$

yields a consistent test procedure for any sequence of latent positions that satisfy $\{\mathbf{X}_n\}, \{\mathbf{Y}_n\}$ for which $\min_{\mathbf{W} \in \mathcal{O}(d)} \|\mathbf{X}_n - \mathbf{Y}_n \mathbf{W}\|$ diverges as $n \rightarrow \infty$.

Our main point of departure in this work, then, is the assumption that, for each n , the latent positions \mathbf{X}_n and \mathbf{Y}_n are independent samples from fixed distributions F and G , respectively. The corresponding tests are, therefore, tests of equality between F and G , up to unitary transformation, scaling, and projection.

Formally, we consider the following two-sample nonparametric testing problems for random dot product graphs. Let F and G be probability distributions with compact support on \mathbb{R}^d for some d ;

we shall assume throughout this paper that d is known. Given $\mathbf{A} \sim \text{RDPG}(F)$ and $\mathbf{B} \sim \text{RDPG}(G)$, we consider the following tests:

(a) (*Equality, up to a unitary transformation*)

$$\begin{array}{c} H_0: F \perp G \\ \text{against } H_A: F \not\perp G \end{array}$$

where $F \perp G$ denotes that there exists a unitary operator U on \mathbb{R}^d such that $F = G \circ U$ and $F \not\perp G$ denotes that $F \neq G \circ U$ for any unitary operator U on \mathbb{R}^d .

(b) (*Equality, up to scaling*)

$$\begin{array}{c} H_0: F \perp G \circ c \quad \text{for some } c > 0 \\ \text{against } H_A: F \not\perp G \circ c \quad \text{for any } c > 0. \end{array}$$

(c) (*Equality of projection*)

$$\begin{array}{c} H_0: F \circ \pi^{-1} \perp G \circ \pi^{-1} \\ \text{against } H_A: F \circ \pi^{-1} \not\perp G \circ \pi^{-1} \end{array}$$

where $\pi: \mathbb{R}^d \mapsto \mathcal{S}$ is the projection $x \mapsto x/\|x\|$ that maps x onto the unit sphere \mathcal{S} in \mathbb{R}^d

We note that the above null hypothesis are nested; $F \perp G$ implies $F \perp G \circ c$ for some $c > 0$ while $F \perp G \circ c$ for some $c > 0$ implies $F \circ \pi^{-1} \perp G \circ \pi^{-1}$.

2.2. Kernel-based two-sample tests. We now introduce the kernel two-sample test of [13]. The test of [13] is just one of several examples of kernel-based testing procedures; see [15] for a recent survey of the literature. Let Ω be a compact metric space and $\kappa: \Omega \times \Omega \mapsto \mathbb{R}$ a continuous, symmetric, and positive definite kernel on Ω . Denote by \mathcal{H} the reproducing kernel Hilbert space associated with κ . Now let F be a probability distribution on Ω . Under mild conditions on κ , the map $\mu[F]$ defined by

$$\mu[F] := \int_{\Omega} \kappa(\omega, \cdot) dF(\omega).$$

belongs to \mathcal{H} . Now, for given probability distributions F and G on Ω , the *maximum mean discrepancy* between F and G with respect to \mathcal{H} is the measure

$$\text{MMD}(F, G; \mathcal{H}) := \|\mu[F] - \mu[G]\|_{\mathcal{H}}.$$

We summarize some important properties of the maximum mean discrepancy from [13]. In particular, if κ is chosen so that μ is an injective map, then $\|\mu[F] - \mu[G]\|_{\mathcal{H}}$ yields a consistent test for testing the hypothesis $\mathbb{H}_0: F = G$ against the hypothesis $\mathbb{H}_A: F \neq G$ for any two arbitrary but fixed distributions F and G on Ω .

LEMMA 1. *Let \mathcal{F} be the unit ball of \mathcal{H} , i.e., $\mathcal{F} = \{h \in \mathcal{H}: \|h\|_{\mathcal{H}} \leq 1\}$. Let F and G be probability distributions on Ω ; X and X' independent random variables with distribution F ; and Y and Y' independent random variables with distribution G . Suppose that X independent of Y . Then*

$$\begin{aligned} (1) \quad \|\mu[F] - \mu[G]\|_{\mathcal{H}}^2 &= \sup_{h \in \mathcal{F}} |\mathbb{E}_F[h] - \mathbb{E}_G[h]|^2 \\ &= \mathbb{E}[\kappa(X, X')] - 2\mathbb{E}[\kappa(X, Y)] + \mathbb{E}[\kappa(Y, Y')]. \end{aligned}$$

Denote by $\Phi: \Omega \mapsto \mathcal{H}$ the canonical feature map $\Phi(X) = \kappa(\cdot, X)$ of κ . Given $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$ and $\mathbf{Y} = \{Y_1, Y_2, \dots, Y_m\}$ with $\{X_i\} \stackrel{\text{i.i.d.}}{\sim} F$ and $\{Y_i\} \stackrel{\text{i.i.d.}}{\sim} G$, the quantity $V_{n,m}(\mathbf{X}, \mathbf{Y})$ defined by

$$(2) \quad \begin{aligned} V_{n,m}(\mathbf{X}, \mathbf{Y}) &= \left\| \frac{1}{n} \sum_{i=1}^n \Phi(X_i) - \frac{1}{m} \sum_{k=1}^m \Phi(Y_k) \right\|_{\mathcal{H}}^2 \\ &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \kappa(X_i, X_j) - \frac{2}{mn} \sum_{i=1}^n \sum_{k=1}^m \kappa(X_i, Y_k) + \frac{1}{m^2} \sum_{k=1}^m \sum_{l=1}^m \kappa(Y_k, Y_l). \end{aligned}$$

is a biased consistent estimate of $\|\mu[F] - \mu[G]\|_{\mathcal{H}}^2$ while the quantity $U_{n,m}(\mathbf{X}, \mathbf{Y})$ defined by

$$(3) \quad U_{n,m}(\mathbf{X}, \mathbf{Y}) = \frac{1}{n(n-1)} \sum_{j \neq i} \kappa(X_i, X_j) - \frac{2}{mn} \sum_{i=1}^n \sum_{k=1}^m \kappa(X_i, Y_k) + \frac{1}{m(m-1)} \sum_{l \neq k} \kappa(Y_k, Y_l).$$

is an unbiased consistent estimate of $\|\mu[F] - \mu[G]\|_{\mathcal{H}}^2$. Finally, if κ is a universal or characteristic kernel, as defined below and in [26, 27], then μ is an injective map, i.e., $\mu[F] = \mu[G]$ if and only if $F = G$.

REMARK. A kernel $\kappa: \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is universal if κ is a continuous function of both its arguments and if the reproducing kernel Hilbert space \mathcal{H} induced by κ is dense in the space of continuous functions on \mathcal{X} with respect to the supremum norm. Let \mathcal{M} be a family of Borel probability measures on \mathcal{X} . A kernel κ is characteristic for \mathcal{M} if the map $\mu \in \mathcal{M} \mapsto \int \kappa(\cdot, z) \mu(dz)$ is injective. If κ is universal, then κ is characteristic for any \mathcal{M} [26]. As an example, let \mathcal{X} be a finite dimensional Euclidean space and define, for any $q \in (0, 2]$, $k_q(x, y) = \frac{1}{2}(\|x\|^q + \|y\|^q - \|x - y\|^q)$. The kernels k_q are then characteristic for the collection of probability distributions with finite second moments [19, 24]. In addition, by Eq. (1), the maximum mean discrepancy with reproducing kernel k_q can be written as

$$\text{MMD}^2(F, G; k_q) = 2\mathbb{E}\|X - Y\|^q - \mathbb{E}\|X - X'\|^q - \mathbb{E}\|Y - Y'\|^q.$$

where X, X' are independent with distribution F , Y, Y' are independent with distribution G , and X, Y are independent. This coincides with the notion of the energy distances of [29], or, when $q = 1$, a special case of the one-dimensional interpoint comparisons of [21].

REMARK. Maximum mean discrepancy can be related to test procedures based on the L_2 -distance between kernel density estimates [2] or test procedures based on the weighted L_2 -distance of characteristic functions [10, 14]. Suppose that f and g are probability density function for F and G respectively, and consider the kernel density estimate \hat{f} and \hat{g} of f and g defined as follows

$$\hat{f} = \frac{1}{n} \sum_{i=1}^n \kappa(\cdot - X_i); \quad \hat{g} = \frac{1}{m} \sum_{k=1}^m \kappa(\cdot - Y_k).$$

Then one has

$$\begin{aligned} \|\hat{f} - \hat{g}\|_{L_2} &= \int_{\Omega} \left(\frac{1}{n} \sum_{i=1}^n \kappa(\omega - X_i) - \frac{1}{m} \sum_{k=1}^m \kappa(\omega - Y_k) \right)^2 d\omega \\ &= \int_{\Omega} \frac{1}{n^2} \sum_{i,j=1}^n \kappa(\cdot - X_i) \kappa(\cdot - X_j) - \frac{2}{mn} \sum_{i,k=1}^n \kappa(\cdot - X_i) \kappa(\cdot - Y_k) + \frac{1}{m^2} \sum_{k,l=1}^m \kappa(\cdot - Y_k) \kappa(\cdot - Y_l) d\omega \\ &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \kappa^*(X_i, X_j) - \frac{2}{mn} \sum_{i=1}^n \sum_{k=1}^m \kappa^*(X_i, Y_k) + \frac{1}{m^2} \sum_{k=1}^m \sum_{l=1}^m \kappa^*(Y_k, Y_l) \end{aligned}$$

where κ^* is the convolution kernel $\kappa^*(x-y) = \int_{\Omega} \kappa(\omega-x)\kappa(\omega-y)d\omega$ [13]. In addition, by Plancherel's identity, one has

$$\|\hat{f} - \hat{g}\|_{L_2} = \int_{\Omega} (\hat{f}(\omega) - \hat{g}(\omega))^2 d\omega = C \int_{\Omega} |\hat{\psi}_F(t) - \hat{\psi}_G(t)|^2 |\tilde{\kappa}(t)|^2 dt$$

for some constant C , where $\hat{\psi}_F$ and $\hat{\psi}_G$ are the *empirical* characteristic functions of F and G , respectively, and $\tilde{\kappa}(t)$ is the characteristic function of κ [2, 14]. If κ is the Gaussian kernel with bandwidth h , then κ^* is a Gaussian kernel with bandwidth $\sqrt{2}h$ and $\tilde{\kappa}$ is also a Gaussian kernel. That is to say, if density estimation is done with respect to the Gaussian kernel, then the L_2 -distance between the kernel density estimates, the weighted L_2 -distance between empirical characteristic functions, and the maximum mean discrepancy with the Gaussian kernel are equivalent. Finally, we note that the Gaussian kernel is universal for any $h > 0$, and hence the test procedures based on L^2 -distance, as well as the test procedure based on the maximum mean discrepancy, are consistent for testing the equality of any pair of distributions F and G . One notable consequence of this is that consistent density estimation is not a prerequisite for consistent nonparametric hypothesis testing; that is, “density estimation is harder than testing”. In particular, in [2], it is established that if the bandwidth h is chosen to converge to 0 as $n, m \rightarrow \infty$, so that \hat{f} and \hat{g} are consistent estimates of f and g , then the test statistic $\|\hat{f} - \hat{g}\|_{L_2}$ yields a consistent test procedure against local alternatives only when $\|f - g\|_{L_2}$ is of order at least $n^{-1/2}h^{-d/2}$. Meanwhile, if h is fixed, $\|\hat{f} - \hat{g}\|_{L_2}$ is consistent against local alternatives for which $\|f - g\|_{L_2}$ is of order $n^{-1/2}$.

We end this section with a result from [13] for the asymptotic distribution of the unbiased estimate $U_{n,m}(\mathbf{X}, \mathbf{Y})$ of $\text{MMD}(F, G; \mathcal{H})$ under the null hypothesis $F = G$.

THEOREM 1. *Let F be a probability distribution on \mathcal{X} and $\kappa: \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ a positive definite kernel. Denote by \mathcal{H} the reproducing kernel Hilbert space associated with κ and by Φ the canonical feature map $\Phi(x) = k(\cdot, x)$. Define the kernel $\tilde{\kappa}$ by*

$$\tilde{\kappa}(x, y) = \kappa(x, y) - \mathbb{E}_{z \sim F} \kappa(x, z) - \mathbb{E}_{z' \sim F} \kappa(z', y) + \mathbb{E}_{z, z' \sim F} \kappa(z, z') = \langle \Phi(x) - \mu[F], \Phi(y) - \mu[F] \rangle_{\mathcal{H}}.$$

We refer to $\tilde{\kappa}$ as the centered version of κ . Let $\mathbf{X} = \{X_1, \dots, X_n\}$ and $\mathbf{Y} = \{Y_1, \dots, Y_m\}$ be two independent collection of independent random variables drawn from F . Suppose that $\lim_{m, n \rightarrow \infty} m/(m+n) \rightarrow \rho \in (0, 1)$. Then

$$(4) \quad (m+n)U_{n,m}(\mathbf{X}, \mathbf{Y}) \xrightarrow{d} \frac{1}{\rho(1-\rho)} \sum_{l=1}^{\infty} \lambda_l (\chi_{1l}^2 - 1)$$

where $\{\chi_{1l}^2\}$ is a sequence of independent χ^2 random variables with one degrees of freedom, and $\{\lambda_l\}$ are the eigenvalues of the integral operator $\mathcal{I}_{F, \tilde{\kappa}}(\phi)$ defined with respect to F and $\tilde{\kappa}$:

$$(5) \quad \mathcal{I}_{F, \tilde{\kappa}}(\phi) = \int_{\Omega} \phi(y) \tilde{\kappa}(x, y) dF(y)$$

REMARK. We observe that the limiting distribution of $(m+n)U_{n,m}(\mathbf{X}, \mathbf{Y})$ above depends on the $\{\lambda_l\}$ which, in turn, depend on the distribution F ; thus the limiting distribution is not distribution-free. Moreover the eigenvalues $\{\lambda_l\}$ can, at best, be estimated; for finite n , they cannot be explicitly determined when F is unknown. In practice, generally the critical values are estimated through a bootstrap resampling or permutation test.

3. Main Results.

3.1. *Equality case.* We state below our first result, an analogue of Theorem 1 for random dot product graphs. Due to the non-identifiability of the random dot product graph under unitary transformations, we shall assume henceforth that the kernel κ is a radial basis kernel; furthermore, since our goal is to exhibit a nonparametric consistent test procedure, we shall assume that κ is universal. Finally, we also assume that κ is twice differentiable. Examples of such kernels are the Gaussian kernels and the inverse multiquadric kernels $\kappa(x, y) = (c^2 + \|x - y\|^2)^{-\beta}$ for $c, \beta > 0$.

THEOREM 2. *Assume the setting of Theorem 1. Let $(\mathbf{X}, \mathbf{A}) \sim \text{RDPG}(F)$ and $(\mathbf{Y}, \mathbf{B}) \sim \text{RDPG}(G)$ be independent random dot product graphs with latent position distributions F and G satisfying Assumption 1. Consider the hypothesis test*

$$H_0: F \perp G \quad \text{against} \quad H_A: F \not\perp G$$

Denote by $\hat{\mathbf{X}} = \{\hat{X}_1, \dots, \hat{X}_n\}$ and $\hat{\mathbf{Y}} = \{\hat{Y}_1, \dots, \hat{Y}_m\}$ the adjacency spectral embedding of \mathbf{A} and \mathbf{B} , respectively. Define the test statistic $U_{n,m} = U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ as follows:

$$(6) \quad U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) = \frac{1}{n(n-1)} \sum_{j \neq i} \kappa(\hat{X}_i, \hat{X}_j) - \frac{2}{mn} \sum_{i=1}^n \sum_{k=1}^m \kappa(\hat{X}_i, \hat{Y}_k) + \frac{1}{m(m-1)} \sum_{l \neq k} \kappa(\hat{Y}_k, \hat{Y}_l)$$

The statistic $U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ is the counterpart of the statistic $U_{n,m}(\mathbf{X}, \mathbf{Y})$ defined in Eq. (3). Suppose that $m, n \rightarrow \infty$ and $m/(m+n) \rightarrow \rho \in (0, 1)$. Then under the null hypothesis of $F \perp G$,

$$(7) \quad (m+n)(U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) - U_{n,m}(\mathbf{X}, \mathbf{Y}\mathbf{W})) \xrightarrow{\text{a.s.}} 0$$

where \mathbf{W} is any orthogonal matrix such that $F = G \circ \mathbf{W}$. Hence, under the null hypothesis of $F \perp G$,

$$(8) \quad (m+n)U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) \xrightarrow{d} \frac{1}{\rho(1-\rho)} \sum_{l=1}^{\infty} \lambda_l (\chi_{1l}^2 - 1)$$

where $\{\chi_{1l}^2\}$ is a sequence of independent χ^2 random variables with one degree of freedom and $\{\lambda_l\}$ are the eigenvalues of the integral operator $\mathcal{I}_{F, \tilde{\kappa}}$ defined in Eq. (5). Under the alternative hypothesis of $F \not\perp G$, there exists a $d \times d$ orthogonal matrix \mathbf{W} , depending on F and G but independent of m and n , such that

$$(9) \quad \frac{m+n}{\log^2(m+n)} (U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) - U_{n,m}(\mathbf{X}, \mathbf{Y}\mathbf{W})) \xrightarrow{\text{a.s.}} 0.$$

REMARK. Eq.(7) and Eq.(9) state that the test statistic $U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ using the *estimated* latent positions is almost identical to the statistic $U_{n,m}(\mathbf{X}, \mathbf{Y})$ defined in Eq. (3) using the true latent positions, under both the null and alternative hypothesis. As κ is a universal kernel, $U_{n,m}(\mathbf{X}, \mathbf{Y})$ converges to 0 under the null and converges to a positive number under the alternative. The test statistic $U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ therefore yields a test procedure that is consistent against any alternative, provided that both F and G satisfy Assumption 1, namely that the second moment matrices have d distinct eigenvalues. Much of the difficulty in proving this theorem is a consequence of the normalization, namely $m+n$, of the test statistic $U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$; this normalization is chosen so that we can derive a limiting distribution similar to that in Theorem 1. Alternatively, one can determine simple but somewhat loose upper bounds on critical values based on concentration inequalities for $U_{m,n}(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$. We emphasize that it follows trivially from this theorem that the test statistic $(m+n)U_{m,n}(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ converges to a non-degenerate distribution under the null and diverges under the alternative.

REMARK. The proof of Theorem 2 can also be adapted to show that under the null hypothesis, $\Delta_\theta = (m+n)(U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) - U_{n,m}(\mathbf{X}, \mathbf{Y}))$ converges to 0 uniformly over some family of kernels $\{\kappa_\theta: \theta \in \Theta\}$. For example, $\{\kappa_\theta: \theta \in \Theta\}$ could be the set of Gaussian kernels with bandwidth $\theta \in \Theta$ for some bounded set $\Theta \subset \mathbb{R}_+$. The uniform convergence of Δ_θ then implies that data-adaptive bandwidth selections behave similarly for $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$ as for \mathbf{X} and \mathbf{Y} .

3.2. *Scaling case.* We now consider the case of testing the hypothesis that the distributions F and G are equal up to scaling. The test statistic is now a simple modification of the one in Theorem 2, i.e., we first scale the adjacency spectral embeddings by the norm of the empirical means before computing the kernel test statistic. The limiting distribution of the test statistic depend on the eigenvalues of integral operators induced by kernels that are transformations of the original kernel. One of these is the scaled and centered version $\tilde{\kappa}^{(s)}$ of κ where

$$(10) \quad \tilde{\kappa}^{(s)}(x, x') = \langle \Phi(x/s) - \mu[F \circ s], \Phi(x'/s) - \mu[F \circ s] \rangle_{\mathcal{H}}.$$

For a given distribution F on Ω , let $\mathcal{C} = \mathcal{C}(F)$ denotes the class of all positive constants c for which $c^2 \|X\|^2 \leq 1$ for F -almost all X . We consider the case of testing the null hypothesis $H_0: F \perp G \circ c$ for some $c \in \mathcal{C}(F)$ against the alternative that $H_A: F \not\perp G \circ c$ for any $c \in \mathcal{C}(F)$. In what follows, we will write $c > 0$, but will always assume that $c \in \mathcal{C}(F)$; otherwise the problem is ill-posed, since X and Y are the latent positions for a random dot product graph. We then have

THEOREM 3. *Let $(\mathbf{X}, \mathbf{A}) \sim \text{RDGP}(F)$ and $(\mathbf{Y}, \mathbf{B}) \sim \text{RDGP}(G)$ be two random dot product graphs with latent positions distribution F and G satisfying Assumption 1. Consider the hypothesis test*

$$\begin{aligned} &H_0: F \perp G \circ c \quad \text{for some } c > 0 \\ \text{against} \quad &H_A: F \not\perp G \circ c \quad \text{for all } c > 0. \end{aligned}$$

Denote by \hat{s}_X, \hat{s}_Y the quantities $n^{-1/2} \|\hat{\mathbf{X}}\|_F$ and $m^{-1/2} \|\hat{\mathbf{Y}}\|_F$. The test statistic $U_{n,m}(\hat{\mathbf{X}}/\hat{s}_X, \hat{\mathbf{Y}}/\hat{s}_Y)$ is defined as follows:

$$(11) \quad U_{n,m}(\hat{\mathbf{X}}/\hat{s}_X, \hat{\mathbf{Y}}/\hat{s}_Y) = \frac{1}{n(n-1)} \sum_{j \neq i} \kappa\left(\frac{\hat{X}_i}{\hat{s}_X}, \frac{\hat{X}_j}{\hat{s}_X}\right) - \frac{2}{mn} \sum_{i=1}^n \sum_{k=1}^m \kappa\left(\frac{\hat{X}_i}{\hat{s}_X}, \frac{\hat{Y}_k}{\hat{s}_Y}\right) + \frac{1}{m(m-1)} \sum_{l \neq k} \kappa\left(\frac{\hat{Y}_k}{\hat{s}_Y}, \frac{\hat{Y}_l}{\hat{s}_Y}\right)$$

Suppose that $m, n \rightarrow \infty$ and $m/(m+n) \rightarrow \rho \in (0, 1)$. Then under the null hypothesis of $F \perp G \circ c$,

$$(m+n)(U_{n,m}(\hat{\mathbf{X}}/\hat{s}_X, \hat{\mathbf{Y}}/\hat{s}_Y) - U_{n,m}(\mathbf{X}/s_X, \mathbf{Y}/s_Y)) \xrightarrow{\text{a.s.}} 0$$

where $s_X = n^{-1/2} \|\mathbf{X}\|_F$ and $s_Y = m^{-1/2} \|\mathbf{Y}\|_F$. Let $g(\omega, X; u)$ for $\omega, X \in \mathbb{R}^d$, $u \in \mathbb{R}$ be the function

$$g(\omega, X; u) = \cos(\omega^\top X/u) + \sin(\omega^\top X/u).$$

By Bochner's Theorem, whenever κ is a positive definite radial kernel, there exists a finite Borel measure M on \mathbb{R}^d such that $\kappa(x, x') = \int g(\omega, x; 1)g(\omega, x'; 1)dM(\omega)$.

Define the function $\tilde{g}(\omega, X; u)$ and $h(\omega; u)$ by

$$\tilde{g}(\omega, X; u) = g(\omega, X; u) - \mathbb{E}[g(\omega, X; u)]; \quad h(\omega; u) = \frac{d}{du} [\mathbb{E}[g(\omega, X; u)]]$$

where both of the above expectations are taken with respect to X only. Denote by σ_X the quantity $(\mathbb{E}[\|X\|^2])^{1/2}$. Finally, define the kernel $\kappa^*(x, x')$ by

$$\kappa^*(x, x') = \int \left(\tilde{g}(\omega, x; \sigma_X) - h(\omega; \sigma_X) \frac{\|x\|^2 - \sigma_X^2}{2\sigma_X} \right) \left(\tilde{g}(\omega, x'; \sigma_X) - h(\omega; \sigma_X) \frac{\|x'\|^2 - \sigma_X^2}{2\sigma_X} \right) dM(\omega).$$

Then under the null hypothesis,

$$(12) \quad (m+n)U_{n,m}(\hat{\mathbf{X}}/\hat{s}_X, \hat{\mathbf{Y}}/\hat{s}_Y) \xrightarrow{d} \frac{1}{\rho(1-\rho)} \sum_{l=1}^{\infty} \lambda_l^* \chi_{1l}^2 - \frac{1}{\rho(1-\rho)} \sum_{l=1}^{\infty} \lambda_l^{(\sigma_X)}$$

where $\{\chi_{1l}^2\}$ is a sequence of independent χ^2 random variables with one degree of freedom, $\{\lambda_l^{(\sigma_X)}\}$ are the eigenvalues of the integral operator $\mathcal{I}_{F, \tilde{\kappa}(\sigma_X)}$ and λ_l^* are the eigenvalues of the integral operator $\mathcal{I}_{F, \kappa^*}$.

REMARK. Theorem 3 states that the limiting distribution of $U_{n,m}(\hat{\mathbf{X}}/\hat{s}_X, \hat{\mathbf{Y}}/\hat{s}_Y)$ coincides with that of $U_{n,m}(\mathbf{X}/s_X, \mathbf{Y}/s_Y)$ and both depend on the eigenvalues of the integral operators $\mathcal{I}_{F, \kappa^*}$ and $\mathcal{I}_{F, \tilde{\kappa}(\sigma_X)}$. As we will demonstrate in the proof of Theorem 3, $U_{n,m}(\mathbf{X}/s_X, \mathbf{Y}/s_Y)$ can be expressed in terms of degenerate U -statistics using estimated parameters; that is, using s_X in place of σ_X . The integral operator κ^* can thus be viewed as the “limit” of $\tilde{\kappa}^{(s_n)}$, where s_n is a \sqrt{n} -consistent estimator for σ_X . The effect of parameter estimation in degenerate U and V statistics has been previously studied: see, for example, [7, 12, 22]; general results regarding the resulting limiting distributions are available. These results, however, assume several regularity conditions that are tedious to verify. We thus choose to present a slightly more self-contained derivation of the limiting distribution for $U_{n,m}(\mathbf{X}/s_X, \mathbf{Y}/s_Y)$ in our proof of Theorem 3. Finally we remark that, in general, the eigenvalues of κ^* are not the same as those of $\kappa^{(\sigma_X)}$. However, in the case where F corresponds to a stochastic blockmodel with all entries on the diagonal being equal, then the eigenvalues of κ^* and those of $\kappa^{(\sigma_X)}$ coincide, so that replacing σ_X by s_X has no effect on the limiting distribution. Indeed, F in this case is a mixture of point masses on a sphere and $\|X_i\| = s_X = \sigma_X$ for all X_i .

3.3. *Projection case.* Finally we consider the case of testing $H_0: F \circ \pi^{-1} \perp G \circ \pi^{-1}$. We shall assume that 0 is not an atom of either F or G , i.e., $F(0) = G(0) = 0$, for otherwise the problem is possibly ill-posed: specifically, $\pi(0)$ is undefined. In addition, for simplicity in the proof, we shall also assume that the support of F and G is bounded away from 0, i.e., there exists some $\epsilon > 0$ such that $F(\{x: \|x\| \leq \epsilon\}) = G(\{x: \|x\| \leq \epsilon\}) = 0$. A truncation argument with $\epsilon \rightarrow 0$ allows us to handle the general case of distributions on Ω where 0 is not an atom. The limiting distribution of the test statistic depends on the eigenvalues of the kernel $\tilde{\kappa}^{(\pi)}$ where

$$(13) \quad \tilde{\kappa}^{(\pi)}(x, x') = \langle \Phi(\pi(x)) - \mu[F \circ \pi^{-1}], \Phi(\pi(x')) - \mu[F \circ \pi^{-1}] \rangle_{\mathcal{H}}.$$

We refer to $\tilde{\kappa}^{(\pi)}$ as the projected and centered version of κ . We then have

THEOREM 4. *Let $(\mathbf{X}, \mathbf{A}) \sim \text{RDGP}(F)$ and $(\mathbf{Y}, \mathbf{B}) \sim \text{RDGP}(G)$ be two random dot product graphs with latent positions distribution F and G satisfying Assumption 1. Suppose also that the support of F and G are both bounded away from 0. Consider the hypothesis test*

$$\begin{aligned} & H_0: F \circ \pi^{-1} \perp G \circ \pi^{-1} \\ \text{against} \quad & H_A: F \circ \pi^{-1} \not\perp G \circ \pi^{-1} \end{aligned}$$

For any matrix $\mathbf{Z} \in \mathbb{R}^{n \times d}$, let $\mathcal{D}(\mathbf{Z})$ be the diagonal matrix whose diagonal entries are the Euclidean norm of the rows of \mathbf{Z} , i.e.,

$$\mathcal{D}(\mathbf{Z}) = (\text{diag}(\mathbf{Z}\mathbf{Z}^T))^{1/2}.$$

Denote by $\pi(\mathbf{Z})$ the projection of the rows of \mathbf{Z} onto the unit sphere, i.e.,

$$\pi(\mathbf{Z}) = \mathcal{D}^{-1}(\mathbf{Z})\mathbf{Z}$$

where, for simplicity, we write $\mathcal{D}^{-1}(\mathbf{Z})$ for $(\mathcal{D}(\mathbf{Z}))^{-1}$. Define the test statistic $U_{n,m}$ as follows:

$$\begin{aligned} U_{n,m}(\pi(\hat{\mathbf{X}}), \pi(\hat{\mathbf{Y}})) &= \frac{1}{n(n-1)} \sum_{j \neq i} \kappa(\pi(\hat{X}_i), \pi(\hat{X}_j)) - \frac{2}{mn} \sum_{i=1}^n \sum_{k=1}^m \kappa(\pi(\hat{X}_i), \pi(\hat{Y}_k)) \\ &\quad + \frac{1}{m(m-1)} \sum_{l \neq k} \kappa(\pi(\hat{Y}_k), \pi(\hat{Y}_l)). \end{aligned}$$

Suppose that $m, n \rightarrow \infty$ and $m/(m+n) \rightarrow \rho \in (0, 1)$. Then under the hypothesis of $F \circ \pi^{-1} \perp G \circ \pi^{-1}$

$$(m+n) \left(U_{n,m}(\pi(\hat{\mathbf{X}}), \pi(\hat{\mathbf{Y}})) - U_{n,m}(\pi(\mathbf{X}), \pi(\mathbf{Y})) \right) \xrightarrow{\text{a.s.}} 0$$

and hence, under the null hypothesis of $F \circ \pi^{-1} \perp G \circ \pi^{-1}$,

$$(14) \quad (m+n) U_{n,m}(\pi(\hat{\mathbf{X}}), \pi(\hat{\mathbf{Y}})) \xrightarrow{d} \frac{1}{\rho(1-\rho)} \sum_{l=1}^{\infty} \lambda_l^{(\pi)} (\chi_{1l}^2 - 1)$$

where $\{\chi_{1l}^2\}$ is a sequence of independent χ^2 random variables with one degree of freedom and $\{\lambda_l^{(\pi)}\}$ are the eigenvalues of $\tilde{\kappa}^{(\pi)}$.

4. Experimental Results. We illustrate the hypothesis tests through several simulated and real data examples. For our first example, let F_ϵ for a given $\epsilon > 0$ be mixture of point masses corresponding to a two-block stochastic block model with block membership probabilities $(0.4, 0.6)$ and block probabilities $\mathbf{B}_\epsilon = \begin{bmatrix} 0.5+\epsilon & 0.2 \\ 0.2 & 0.5+\epsilon \end{bmatrix}$. We then test, for a given $\epsilon > 0$, the hypothesis $H_0: F_0 \perp F_\epsilon$ against the alternative $H_A: F_0 \not\perp F_\epsilon$ using the kernel-based testing procedure of § 3. The kernel is chosen to be the Gaussian kernel with bandwidth $\sigma = 0.5$. We first evaluate the performance through simulation using 1000 Monte Carlo replicates; in each replicate we sample two graphs on n vertices from $\text{RDPG}(F_0)$ and one graph on n vertices from $\text{RDPG}(F_\epsilon)$. We then perform an adjacency spectral embedding on the graphs, in which we embed the graphs into a \mathbb{R}^2 , and we proceed to compute the kernel-based test statistic. The embeddings of the graphs sampled according to F_0 are used to compute the empirical distribution of the test statistic under the null hypothesis. The results for $\epsilon = 0.2$ and $n = 1000$ are presented in Fig. 1. For the purposes of comparison, we also include the empirical distribution of the test statistics as computed using the true sampled positions. We note that the distributions of the test statistics computed from the embeddings $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$ well approximate those computed from the true sampled positions \mathbf{X} and \mathbf{Y} under both the null and alternative. We also evaluate the performance of the test procedures for both $U_{n,m}(\mathbf{X}, \mathbf{Y})$ and $U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ by estimating the size and power of the test statistic for various choices of $n \in \{100, 200, 500, 1000\}$ and $\epsilon \in \{0, 0.05, 0.1, 0.2\}$ through Monte Carlo simulation. The significance level is set to $\alpha = 0.05$ and the rejection regions are specified via $B = 200$ bootstrap permutation using either the true latent positions \mathbf{X} and \mathbf{Y} or the estimated latent positions $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$. These estimates are given in Table 1. For our second example, let F_ϵ be the uniform distribution

n	$\epsilon = 0.02$		$\epsilon = 0.05$		$\epsilon = 0.1$	
	$\{\mathbf{X}, \mathbf{Y}\}$	$\{\hat{\mathbf{X}}, \hat{\mathbf{Y}}\}$	$\{\mathbf{X}, \mathbf{Y}\}$	$\{\hat{\mathbf{X}}, \hat{\mathbf{Y}}\}$	$\{\mathbf{X}, \mathbf{Y}\}$	$\{\hat{\mathbf{X}}, \hat{\mathbf{Y}}\}$
100	0.07	0.06	0.07	0.09	0.21	0.27
200	0.06	0.09	0.11	0.17	0.89	0.83
500	0.08	0.1	0.37	0.43	1	1
1000	0.1	0.14	1	1	1	1

TABLE 1

Power estimates for testing the null hypothesis $F \perp G$ at a significance level of $\alpha = 0.05$ using bootstrap permutation tests for the U -statistics $U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ and $U_{n,m}(\mathbf{X}, \mathbf{Y})$. In each bootstrap test, $B = 200$ bootstrap samples were generated. Each estimate of power is based on 1000 Monte Carlo replicates of the corresponding bootstrap test.

on $[\epsilon, 1/\sqrt{2}]^2$ where $\epsilon \geq 0$. In addition, let G be the uniform distribution on $[0, 1/\sqrt{3}]^2$. For a given ϵ , we test the hypothesis $H_0: F_\epsilon \perp G \circ c$ for some constant $c > 0$ against the alternative $H_A: F_\epsilon \not\perp G \circ c$ for any constant $c > 0$. The testing procedure is based on the test statistic $(m+n)U_{n,m}(\hat{\mathbf{X}}/\hat{s}_X, \hat{\mathbf{Y}}/\hat{s}_Y)$ using a Gaussian kernel with bandwidth $\sigma = 0.5$. Once again, we evaluate the performance through numerical simulation using 1000 Monte Carlo replicates. The results are presented in Fig. 2. We also include the empirical distribution of the test statistics computed using the true parameter $\sigma_X = \mathbb{E}[\|X\|^2]^{1/2}$. Once gain, the distributions of the test statistics computed from the embeddings well approximate those computed from the true latent positions. Table 2 is the analogue of Table 1 and presents estimates of the size and power for $U_{n,m}(\mathbf{X}/s_X, \mathbf{Y}/s_Y)$ and $U_{n,m}(\hat{\mathbf{X}}/\hat{s}_X, \hat{\mathbf{Y}}/\hat{s}_Y)$ for various choices of n and ϵ .

n	$\epsilon = 0$		$\epsilon = 0.05$		$\epsilon = 0.1$		$\epsilon = 0.2$	
	$\{\mathbf{X}, \mathbf{Y}\}$	$\{\hat{\mathbf{X}}, \hat{\mathbf{Y}}\}$	$\{\mathbf{X}, \mathbf{Y}\}$	$\{\hat{\mathbf{X}}, \hat{\mathbf{Y}}\}$	$\{\mathbf{X}, \mathbf{Y}\}$	$\{\hat{\mathbf{X}}, \hat{\mathbf{Y}}\}$	$\{\mathbf{X}, \mathbf{Y}\}$	$\{\hat{\mathbf{X}}, \hat{\mathbf{Y}}\}$
100	0.05	0.04	0.184	0.02	0.79	0.16	1	0.91
200	0.06	0.1	0.39	0.11	0.98	0.7	1	1
500	0.07	0.07	0.83	0.66	1	1	1	1
1000	0.06	0.03	1	0.98	1	1	1	1

TABLE 2

Power estimates for testing the null hypothesis $F \perp G \circ c$ at a significance level of $\alpha = 0.05$ using bootstrap permutation tests for the U -statistics $U_{n,m}(\hat{\mathbf{X}}/\hat{s}_X, \hat{\mathbf{Y}}/\hat{s}_Y)$ and $U_{n,m}(\mathbf{X}/s_X, \mathbf{Y}/s_Y)$. In each bootstrap test, $B = 200$ bootstrap samples were generated. Each estimate of power is based on 1000 Monte Carlo replicates of the corresponding bootstrap test. The entries for $\epsilon = 0$ coincides with bootstrap estimate for the size of the test.

For our last example, we consider the problem of classifying proteins into enzyme versus non-enzymes. We use the dataset of [9], which consist of 1178 protein networks labeled as enzymes (691 networks) and non-enzymes (487 networks). Our classification procedure is as follows. We first embed each of the protein networks into \mathbb{R}^5 using adjacency spectral embedding. We then compute a 1178×1178 matrix \mathbf{S} of pairwise dissimilarity between the adjacency spectral embedding of the protein networks using a Gaussian kernel with bandwidth $h = 1$. The classifier is a k -NN classifier using the dissimilarities in \mathbf{S} in place of the Euclidean distance. We evaluate the classification accuracy using a 10-fold cross validation. The results are presented in Table 3. For the purpose of comparison, we also include the accuracy of several other classifiers that were previously applied on this data set, see [6, 9]. The results of [9] are based on modeling the proteins using various features such as secondary-structure content, surface properties, ligands, and amino acid propensities, and then training a SVM using a radial basis kernel on these feature vectors. The results of [6] are based on representing the proteins as graphs, using their secondary-structure content, and then training a SVM classifier using a random walk kernel on the result graphs. We note that the classifier accuracy is comparable among the different classifiers.

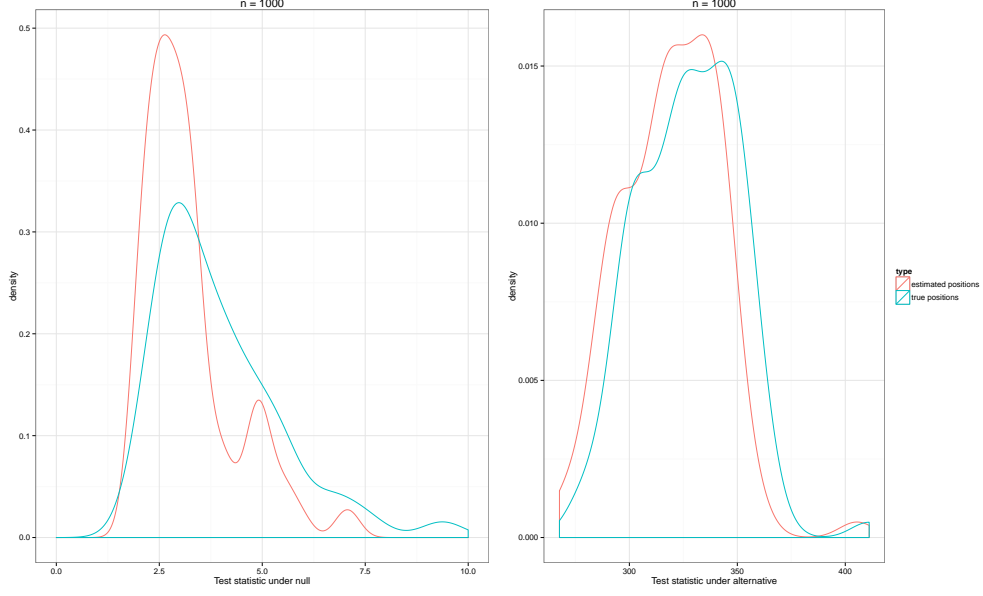


Fig 1: Distribution of test statistics under null and alternative as computed from the latent positions and those estimated from adjacency spectral embedding for testing the null hypothesis $F \perp G$.

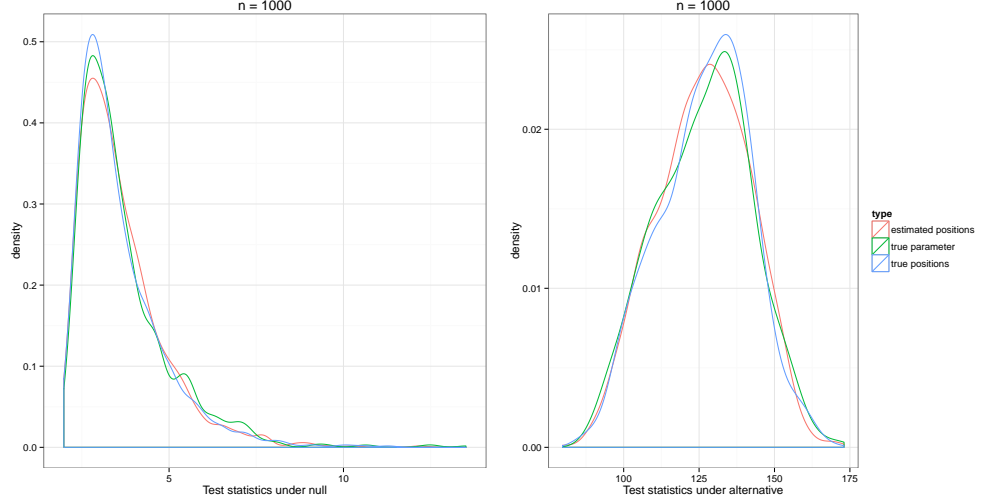


Fig 2: Distribution of test statistics under null and alternative as computed from the latent positions and those estimated from adjacency spectral embedding for testing the null hypothesis $F_{0.2} \perp G \circ c$ for some constant $c > 0$.

Classifier	Accuracy (%)	standard deviation
SVM with feature vector kernel [9]	76.86	1.23
SVM with optimized feature vector kernel [9]	80.17	1.24
SVM with random walk kernel [6]	77.30	1.2
SVM with random walk kernel without structure [6]	72.33	5.32
k -NN with dissimilarities based on $U_{n,m}$	78.20	2.10

TABLE 3

Classification accuracy on the enzyme dataset.

5. Discussion. In summary, we show in this paper that the adjacency spectral embedding can be used to generate simple and intuitive test statistics for the nonparametric inference problem of testing whether two random dot product graphs have the same or related distribution of latent positions. The two-sample formulations presented here and the corresponding test statistics are intimately related. Indeed, for random dot product graphs, the adjacency spectral embedding yields a consistent estimate of the latent positions as points in \mathbb{R}^d ; there then exists a wide variety of classical and well-studied testing procedures for data in Euclidean spaces.

New results on stochastic blockmodels suggest that they can be regarded as a universal approximation to graphons in exchangeable random graphs, see e.g., [33, 34]. There is thus potential theoretical value in the formulation of a two-sample hypothesis testing for latent position models in terms of a random dot product graph model on \mathbb{R}^d with possibly varying d . However, because the link function and the distribution of latent positions are intertwined in the context of latent position graphs, any proposed test procedure that is sufficiently general might also possess little to no power.

The two-sample hypothesis testing we consider here is also closely related to the problem of testing goodness-of-fit; the results in this paper can be easily adapted to address the latter question. In particular, we can test, for a given graph, whether the graph is generated from some specified stochastic blockmodel. A more general problem is that of testing whether a given graph is generated according to a latent position model with a specific link function. This problem has been recently studied; see [34] for a brief discussion, but much remains to be investigated. For example, the limiting distribution of the test statistic in [34] is not known.

Finally, two-sample hypothesis testing is also closely related to testing for independence; given a random sample $\{(X_i, Y_i)\}$ with joint distribution F_{XY} and marginal distributions F_X and F_Y , X and Y are independent if the F_{XY} differs from the product $F_X F_Y$. For example, the Hilbert-Schmidt independence criterion is a measure for statistical dependence in terms of the Hilbert-Schmidt norm of a cross-covariance operator. It is based on the maximum mean discrepancy between F_{XY} and $F_X F_Y$. Another example is Brownian distance covariance [28], a measure of dependence based on the energy distance between F_{XY} and $F_X F_Y$. In particular, consider the test of whether two given two random dot product graphs $(\mathbf{X}, \mathbf{A}) \sim \text{RDPG}(F_X)$ and $(\mathbf{Y}, \mathbf{B}) \sim \text{RDPG}(F_Y)$ on the same vertex set have independent latent position distributions F_X and F_Y . While we surmise that it may be possible to adapt our present results to this question, we stress that the conditional independence of \mathbf{A} given \mathbf{X} and of \mathbf{B} given \mathbf{Y} suggests that independence testing may merit a more intricate approach.

REFERENCES

- [1] AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E. and XING, E. P. (2008). Mixed membership stochastic blockmodels. *The Journal of Machine Learning Research* **9** 1981–2014.
- [2] ANDERSON, N., HALL, P. and TITTERINGTON, D. (1994). Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis* **50** 41–54.
- [3] BARINGHAUS, L. and HENZE, N. (1988). A consistent test for multivariate normality based on the empirical characteristic function. *Metrika* **35** 339–348.
- [4] BILLINGSLEY, P. (1999). *Convergence of probability measures*, Second ed. Wiley.
- [5] BORGWARDT, K. M., GRETTON, A., RASCH, M. J., KRIEGEL, H. P., SCHÖLKOPF, B. and SMOLA, A. J. (2006). Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics* **22** 49–57.
- [6] BORGWARDT, K. M., ONG, C. S., SCHONAUER, S., VISHWANATHAN, S. V. N., SMOLA, A. J. and KRIEGEL, H. P. (2005). Protein function prediction via graph kernels. *Bioinformatics* **21** 47–56.

- [7] DE WET, T. and RANDLES, R. H. (1987). On the effect of substituting parameter estimators in limiting χ^2 U and V statistics. *Annals of Statistics* **15** 398–412.
- [8] DIACONIS, P. and JANSON, S. (2008). Graph Limits and Exchangeable Random Graphs. *Rendiconti di Matematica, Serie VII* **28** 33–61.
- [9] DOBSON, P. D. and DOIG, A. J. (2003). Distinguishing enzyme structures from non-enzymes without alignments. *Journal of Molecular Biology* **330** 771–781.
- [10] FERNÁNDEZ, V. A., GAMERO, M. D. J. and GARCÍA, J. M. (2008). A test for the two-sample problem based on empirical characteristic functions. *Computational Statistics and Data Analysis* **52** 3730–3748.
- [11] GÄRTNER, T., FLACH, P. and WROBEL, S. (2003). On graph kernels: hardness results and efficient alternatives. In *Proceedings of the 16th Annual Conference on Learning Theory*.
- [12] GREGORY, G. G. (1977). Large sample theory for U -statistics and tests of fit. *Annals of Statistics* **5** 110–123.
- [13] GRETTON, A., BORGWADT, K. M., RASCH, M. J., SCHÖLKOPF, B. and SMOLA, A. (2012). A Kernel Two-Sample Test. *Journal of Machine Learning Research* **13** 723–773.
- [14] HALL, P., LOMBARD, F. and POTGIETER, C. J. (2013). A new approach to function-based hypothesis testing in location-scale families. *Technometrics* **55** 215–223.
- [15] HARCHAoui, Z., BACH, F., CAPPÉ, O. and MOULINES, E. (2013). Kernel-based methods for hypothesis testing: A Unified View. *IEEE Signal Processing Magazine* **30** 87–97.
- [16] HOFF, P. D., RAFTERY, A. E. and HANDCOCK, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association* **97** 1090–1098.
- [17] HOLLAND, P. W., LASKEY, K. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Social Networks* **5** 109–137.
- [18] KARRER, B. and NEWMAN, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **83**.
- [19] LYONS, R. (2013). Distance covariance in metric spaces. *Annals of Probability* **41** 3284–3305.
- [20] LYZINSKI, V., SUSSMAN, D. L., TANG, M., ATHREYA, A. and PRIEBE, C. E. (2013). Perfect Clustering for Stochastic Blockmodel Graphs via Adjacency Spectral Embedding. Arxiv preprint. <http://arxiv.org/abs/1310.0532>.
- [21] MAA, J. F., PEARL, D. K. and BARTOSZYŃSKI, R. (1996). Reducing multidimensional two-sample data to one-dimensional interpoint comparisons. *Annals of Statistics* **24** 1067–1074.
- [22] MOORE, D. S. and SPRUILL, M. C. (1975). Unified large-sample theory of general chi-squared statistics for tests of fit. *Annals of Statistics* **3** 599–616.
- [23] PINELIS, I. (1994). Optimum bounds for the distribution of martingales in Banach spaces. *Annals of Probability* **22** 1679–1706.
- [24] SEJDINOVIC, D., SRIPERUMBUDUR, B., GRETTON, A. and FUKUMIZU, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics* **41** 2263–2291.
- [25] SERFLING, R. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.
- [26] SRIPERUMBUDUR, B. K., FUKUMIZU, K. and LANCKRIET, G. R. G. (2011). Universality, characteristic kernels and RKHS embeddings of measures. *Journal of Machine Learning Research* **12** 2389–2410.
- [27] STEINWART, I. (2001). On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research* **2** 67–93.
- [28] SZÉKELY, G. J. and RIZZO, M. L. (2009). Brownian distance covariance. *Annals of Applied Statistics* **3** 1236–1265.
- [29] SZÉKELY, G. J. and RIZZO, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference* **143** 1249–1272.
- [30] TANG, M., ATHREYA, A., SUSSMAN, D. L., LYZINSKI, V. and PRIEBE, C. E. (2014). Two-sample hypothesis testing for random dot product graphs via adjacency spectral embedding. arXiv preprint. <http://arxiv.org/abs/1403.7249>.
- [31] VAN DE GEER, S. (2000). *Empirical processes in M-estimation*. Cambridge University Press.
- [32] VISHWANATHAN, S. V. N., SCHRAUDOLPH, N. N., KONDOR, R. and BORGWADT, K. M. (2010). Graph kernels. *Journal of Machine Learning Research* **11** 1201–1242.
- [33] WOLFE, P. J. and OLHEDE, S. C. (2013). Nonparametric graphon estimation. arXiv preprint at <http://arxiv.org/abs/1309.5936>.
- [34] YANG, J. J., HAN, Q. and AIROLDI, E. M. (2014). Nonparametric estimation and testing of exchangeable graph models. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics* 1060–1067.
- [35] YOUNG, S. and SCHEINERMAN, E. (2007). Random dot product graph models for social networks. In *Proceedings of the 5th international conference on algorithms and models for the web-graph* 138–149.

APPENDIX A: TWO TECHNICAL LEMMAS

The proofs of Theorem 2 through Theorem 4 depend on the following two lemmas. The first lemma bounds, *simultaneously*, the difference between $\hat{\mathbf{X}}$ and \mathbf{X} and the difference between $\hat{\mathbf{Y}}$ and \mathbf{Y} for two choices of matrix norms. We consider the Frobenius norm $\|\cdot\|_F$ and the maximum of the l_2 norms of the rows $\|\cdot\|_{2 \rightarrow \infty}$. The norm $\|\cdot\|_{2 \rightarrow \infty}$ is induced by the vector norms $\|\cdot\|_2$ and $\|\cdot\|_\infty$ via $\|\mathbf{A}\|_{2 \rightarrow \infty} = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_\infty$. The second lemma provides uniform control, for all functions f belonging to a particular class \mathcal{F} , of a scaled sum of terms of the form $f(X_i)^\top (\hat{X}_i - X_i)$. We state the lemmas below; their proofs are given in section B of the appendix.

LEMMA 2. *Let $(\mathbf{X}, \mathbf{A}) \sim \text{RDPG}(F)$ and $(\mathbf{Y}, \mathbf{B}) \sim \text{RDPG}(G)$ be random dot product graphs whose latent position distributions F and G both satisfy the conditions in Assumption 1. Suppose, without loss of generality, that $n \geq m$. Let $c > 0$ be arbitrary but fixed. There exists $m_0(c)$ such that if $m \geq m_0$ and η satisfies $m^{-c} < \eta < 1/4$, then there exist deterministic matrices \mathbf{W}_1 and \mathbf{W}_2 such that, with probability at least $1 - 4\eta$,*

$$(15) \quad \|\hat{\mathbf{X}} - \mathbf{X}\mathbf{W}_1\|_F \leq C(F) + C_1 \sqrt{\frac{\log(m/\eta)}{m}} + C_2 \sqrt{\log(1/\eta)};$$

$$(16) \quad \|\hat{\mathbf{X}} - \mathbf{X}\mathbf{W}_1\|_{2 \rightarrow \infty} \leq C_3 \sqrt{\frac{\log(m/\eta)}{m}} \quad \text{and}$$

$$(17) \quad \|\hat{\mathbf{Y}} - \mathbf{Y}\mathbf{W}_2\|_F \leq C(F) + C_1 \sqrt{\frac{\log(m/\eta)}{m}} + C_2 \sqrt{\log(1/\eta)};$$

$$(18) \quad \|\hat{\mathbf{Y}} - \mathbf{Y}\mathbf{W}_2\|_{2 \rightarrow \infty} \leq C_3 \sqrt{\frac{\log(m/\eta)}{m}}$$

Furthermore, if $F = G$ then there exist $\mathbf{W}_1 = \mathbf{W}_2$ satisfying the above inequalities.

LEMMA 3. *Denote by $(\partial\Phi(Z)) \in \{f: \mathbb{R}^d \mapsto \mathbb{R}^d\}$ the gradient of the feature map Φ at Z , i.e.,*

$$(\partial\Phi(Z))(\cdot) = \left. \frac{\partial \kappa(X, \cdot)}{\partial X} \right|_{X=Z}.$$

Let \mathcal{F} be the following class of functions, indexed by values $Z \in \Omega$:

$$\mathcal{F} = \{f : f = (\partial\Phi(\cdot))(Z) : Z \in \Omega\}$$

Then there exists an orthogonal $\mathbf{W} \in \mathbb{R}^{d \times d}$ such that

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n f(X_i)^T (\mathbf{W} \hat{X}_i - X_i) \right| \rightarrow 0$$

almost surely.

APPENDIX B: PROOFS

Proof of Lemma 2. The bounds given in Eq. (15) through Eq. (18) were previously proved in [20] for arbitrary distributions F and G . We now show that in the case when $F = G$, one can choose \mathbf{W}_1 and \mathbf{W}_2 such that $\mathbf{W}_1 = \mathbf{W}_2$. Denote by Σ_F the second moment matrix of F and let $\Sigma_{\mathbf{X}}$ and

$\Sigma_{\mathbf{Y}}$ be the empirical estimates of Σ_F obtained from \mathbf{X} and \mathbf{Y} , respectively. We then have, for some constant $C > 0$,

$$\begin{aligned}\|\Sigma_{\mathbf{X}} - \Sigma_F\| &= \left\| \frac{1}{n} \mathbf{X}^T \mathbf{X} - \Sigma_F \right\| \leq C \sqrt{\frac{\log(1/\eta)}{n}} \quad \text{and} \\ \|\Sigma_{\mathbf{Y}} - \Sigma_F\| &= \left\| \frac{1}{m} \mathbf{Y}^T \mathbf{Y} - \Sigma_F \right\| \leq C \sqrt{\frac{\log(1/\eta)}{m}}\end{aligned}$$

with probability at least $1 - 2\eta$. Now, we can find orthogonal matrices \mathbf{W}_1 and \mathbf{W}_2 such that $\mathbf{X}\mathbf{W}_1 = \mathbf{U}_{\mathbf{P}}\mathbf{S}_{\mathbf{P}}^{1/2}$ and similarly $\mathbf{Y}\mathbf{W}_2 = \mathbf{U}_{\mathbf{Q}}\mathbf{S}_{\mathbf{Q}}$. The triangle inequality implies that

$$\left\| \frac{1}{n} \mathbf{X}^T \mathbf{X} - \frac{1}{m} \mathbf{Y}^T \mathbf{Y} \right\| = \left\| \frac{1}{n} \mathbf{W}_1^T \mathbf{S}_{\mathbf{P}} \mathbf{W}_1 - \frac{1}{m} \mathbf{W}_2^T \mathbf{S}_{\mathbf{Q}} \mathbf{W}_2 \right\| \leq C \sqrt{\frac{\log(1/\eta)}{m}}$$

with probability at least $1 - 4\eta$. In addition, as a consequence of Assumption 1, the eigenvalues of $\mathbf{S}_{\mathbf{P}}/n$ and $\mathbf{S}_{\mathbf{Q}}/m$ are also distinct, provided that the preceding events occur. Therefore, by the Davis-Kahan theorem applied to the individual columns of \mathbf{W}_1 and \mathbf{W}_2 , we deduce that

$$\|\mathbf{W}_1 - \mathbf{W}_2\| \leq C \sqrt{\frac{\log(1/\eta)}{m}}$$

with probability at least $1 - 4\eta$. Henceforth

$$\|\hat{\mathbf{Y}} - \mathbf{Y}\mathbf{W}_1\|_F \leq \|\hat{\mathbf{Y}} - \mathbf{Y}\mathbf{W}_2\|_F + \|\mathbf{Y}(\mathbf{W}_1 - \mathbf{W}_2)\|_F \leq C(F) + C_1 \sqrt{\frac{\log(1/\eta)}{m}} + C_2 \sqrt{\log(1/\eta)}.$$

Similarly,

$$\|\hat{\mathbf{Y}} - \mathbf{Y}\mathbf{W}_1\|_{2 \rightarrow \infty} \leq \|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}\mathbf{W}_2\|_{2 \rightarrow \infty} + \|\mathbf{Y}\|_{2 \rightarrow \infty} \|\mathbf{W}_1 - \mathbf{W}_2\| \leq C_3 \sqrt{\frac{\log(m/\eta)}{m}},$$

and hence one can take $\mathbf{W}_1 = \mathbf{W}_2$, as desired.

Proof of Lemma 3. For any $f \in \mathcal{F}$, and any X_1, \dots, X_n , let $\mathbf{M}(f) = \mathbf{M}(f; X_1, \dots, X_n) \in \mathbb{R}^{n \times d}$ be the matrix whose rows are the vectors $f(X_i)$. We deduce that

$$\begin{aligned}\zeta(f) &:= \frac{1}{\sqrt{n}} \sum_{i=1}^n [f(X_i)]^T (\mathbf{W} \hat{X}_i - X_i) \\ &= \frac{1}{\sqrt{n}} \text{tr}(\hat{\mathbf{X}}\mathbf{W} - \mathbf{X})[\mathbf{M}(f)]^T \\ &= \frac{1}{\sqrt{n}} \text{tr}(\mathbf{U}_{\mathbf{A}}\mathbf{S}_{\mathbf{A}}^{1/2} - \mathbf{U}_{\mathbf{P}}\mathbf{S}_{\mathbf{P}}^{1/2})\widetilde{\mathbf{W}}[\mathbf{M}(f)]^T \\ &= \frac{1}{\sqrt{n}} \text{tr}\left(\mathbf{A}(\mathbf{U}_{\mathbf{A}} - \mathbf{U}_{\mathbf{P}})\mathbf{S}_{\mathbf{A}}^{-1/2} + \mathbf{A}\mathbf{U}_{\mathbf{P}}(\mathbf{S}_{\mathbf{A}}^{-1/2} - \mathbf{S}_{\mathbf{P}}^{-1/2}) + (\mathbf{A} - \mathbf{P})\mathbf{U}_{\mathbf{P}}\mathbf{S}_{\mathbf{P}}^{-1/2}\right)\widetilde{\mathbf{W}}[\mathbf{M}(f)]^T\end{aligned}$$

where $\widetilde{\mathbf{W}}$ is some orthogonal matrix derivable from \mathbf{W} and \mathbf{X} . We therefore derive that

$$\begin{aligned}\sup_{f \in \mathcal{F}} |\zeta(f)| &\leq \frac{\sup_{f \in \mathcal{F}} \|\mathbf{M}(f)\|_F}{\sqrt{n}} \left(\|\mathbf{A}(\mathbf{U}_{\mathbf{A}} - \mathbf{U}_{\mathbf{P}})\mathbf{S}_{\mathbf{A}}^{-1/2}\widetilde{\mathbf{W}}\|_F + \|\mathbf{A}\mathbf{U}_{\mathbf{P}}(\mathbf{S}_{\mathbf{A}}^{-1/2} - \mathbf{S}_{\mathbf{P}}^{-1/2})\widetilde{\mathbf{W}}\|_F \right) \\ &\quad + \frac{1}{\sqrt{n}} \sup_{f \in \mathcal{F}} |\text{tr}[\mathbf{M}(f)]^T (\mathbf{A} - \mathbf{P})\mathbf{U}_{\mathbf{P}}\mathbf{S}_{\mathbf{P}}^{-1/2}\widetilde{\mathbf{W}}|\end{aligned}$$

We bound the first two terms on the right hand side of the above expression using the following lemma from [20].

LEMMA 4. Let $(\mathbf{X}, \mathbf{A}) \sim \text{RDPG}(F)$ where F satisfy Assumption 1. Let $c > 0$ be arbitrary but fixed. There exists $n_0(c)$ such that if $n > n_0$ and η satisfies $n^{-c} < \eta < 1/2$, then with probability at least $1 - 2\eta$, the following bounds hold simultaneously

$$(19) \quad \|\mathbf{A}(\mathbf{U}_{\mathbf{A}} - \mathbf{U}_{\mathbf{P}})\mathbf{S}_{\mathbf{A}}^{-1/2}\|_F \leq \frac{24\sqrt{2}d \log(n/\eta)}{\sqrt{\gamma^5(F)n}}$$

$$(20) \quad \|\mathbf{A}\mathbf{U}_{\mathbf{P}}(\mathbf{S}_{\mathbf{A}}^{-1/2} - \mathbf{S}_{\mathbf{P}}^{-1/2})\|_F \leq \frac{18d^{3/2} \log(n/\eta)}{\sqrt{\gamma^7(F)n}}$$

Applying the bounds in Lemma 4, we have

$$\sup_f |\zeta(f)| \leq \frac{C(F) \log n}{\sqrt{n}} + \frac{1}{\sqrt{n}} \sup_{f \in \mathcal{F}} |\text{tr}[\mathbf{M}(f)]^T (\mathbf{A} - \mathbf{P}) \mathbf{U}_{\mathbf{P}} \mathbf{S}_{\mathbf{P}}^{-1/2} \widetilde{\mathbf{W}}|$$

with probability at least $1 - n^{-2}$, where $C(F)$ is a constant depending only on F .

We next show that the last term on the right hand side of the above display is also of order $n^{-1/2}(\log n)$ with probability at least $1 - n^{-2}$. To control this supremum, we use a chaining argument. For a given $f \in \mathcal{F}$ let $\|f\|_{\infty}$ denote the quantity $\sup_{Z \in \Omega} \|f(Z)\|_2$, where $\|\cdot\|_2$ is the Euclidean norm in \mathbb{R}^d . Similarly, for given $f, g \in \mathcal{F}$, let $\|f - g\|_{\infty}$ denote the quantity $\sup_{Z \in \Omega} \|f(Z) - g(Z)\|_2$. By the smoothness of Φ and the compactness of Ω , we derive that \mathcal{F} is compact with respect to this norm. Put $\delta = \sup_{f \in \mathcal{F}} \|f\|_{\infty}$. By the compactness of \mathcal{F} , for any $j \in \mathbb{N}$, we can find a finite subset $S_j = \{f_1, f_2, \dots, f_{n_j}\}$ of \mathcal{F} such that for all $f \in \mathcal{F}$, there exists a $f_l \in S_j$ with $\|f - f_l\|_{\infty} \leq \delta_j := 2^{-j}\delta$. We shall assume that S_j is *minimal* among all sets with the above property. In addition, $|S_0| = 1$.

Given S_j , define Π_j as the mapping that maps any $f \in \mathcal{F}$ to an (arbitrary) $f_l \in S_j$ satisfying the condition $\|f_l - f\|_{\infty} \leq \delta_j$. Denote by $\tilde{X}_1, \dots, \tilde{X}_n$ the rows of the matrix $\mathbf{A}\mathbf{U}_{\mathbf{P}}\mathbf{S}_{\mathbf{P}}^{-1/2}\widetilde{\mathbf{W}}$. Then by the separability of Ω , we have

$$\begin{aligned} \frac{1}{\sqrt{n}} \sup_{f \in \mathcal{F}} |\text{tr}[\mathbf{M}(f)]^T (\mathbf{A} - \mathbf{P}) \mathbf{U}_{\mathbf{P}} \mathbf{S}_{\mathbf{P}}^{-1/2} \widetilde{\mathbf{W}}| &= \sup_{f \in \mathcal{F}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n f(X_i)^T (\tilde{X}_i - X_i) \right| \\ &= \sup_{f \in \mathcal{F}} \left| \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \sum_{j=0}^{\infty} (\Pi_{j+1}f - \Pi_j f)(X_i)^T (\tilde{X}_i - X_i) \right) + \frac{c_0}{\sqrt{n}} \right| \\ &= \sup_{f \in \mathcal{F}} \left| \left(\frac{1}{\sqrt{n}} \sum_{j=0}^{\infty} \sum_{i=1}^n (\Pi_{j+1}f - \Pi_j f)(X_i)^T (\tilde{X}_i - X_i) \right) + \frac{c_0}{\sqrt{n}} \right| \\ &\leq \sum_{j=0}^{\infty} \sup_{f \in \mathcal{F}} \left| \frac{1}{\sqrt{n}} \left(\sum_{i=1}^n (\Pi_{j+1}f - \Pi_j f)(X_i)^T (\tilde{X}_i - X_i) \right) \right| + \left| \frac{c_0}{\sqrt{n}} \right| \end{aligned}$$

where $c_0 = \sum_{i=1}^n (\Pi_0 f)(X_i)^T (\tilde{X}_i - X_i)$.

The term $n^{-1/2} \sum_{i=1}^n (\Pi_{j+1}f - \Pi_j f)(X_i)^T (\tilde{X}_i - X_i)$ can be written as sum of quadratic form, i.e.,

$$(21) \quad \frac{1}{\sqrt{n}} \sum_{i=1}^n (\Pi_{j+1}f - \Pi_j f)(X_i)^T (\tilde{X}_i - X_i) = \frac{1}{\sqrt{n}} \sum_{s=1}^d (\pi_s^{(j,j+1)}(f))^T (\mathbf{A} - \mathbf{P}) \mathbf{u}_s \lambda_s^{-1/2}$$

where $\pi_s^{(j,j+1)}(f)$ for $s = 1, 2, \dots, d$ are the columns of the $n \times d$ matrix with rows $(\Pi_{j+1}f - \Pi_j f)(X_i)$ for $i = 1, \dots, n$ and \mathbf{u}_s and λ_s are the eigenvectors and corresponding eigenvalues of \mathbf{P} .

Now, for any vectors $\mathbf{b} = (b_1, b_2, \dots, b_n)$ and $\mathbf{c} = (c_1, c_2, \dots, c_n)$,

$$\mathbf{b}^T (\mathbf{A} - \mathbf{P}) \mathbf{c} = \sum_{i < j} b_i (\mathbf{A} - \mathbf{P})_{ij} c_j + \sum_i \mathbf{P}_{ii} b_i c_i$$

The sum over the indices $i < j$ in the above display is a sum of independent random variables. Therefore, by Hoeffding's inequality ensures that

$$\mathbb{P} \left[\left| \sum_{i < j} b_i (\mathbf{A} - \mathbf{P})_{ij} c_j \right| \geq t \right] \leq 2 \exp \left(-\frac{t^2}{2 \sum_{i < j} b_i^2 c_j^2} \right) \leq 2 \exp \left(-\frac{t^2}{2 \|\mathbf{b}\|^2 \|\mathbf{c}\|^2} \right)$$

In addition, $\sum_i \mathbf{P}_{ii} b_i c_i \leq \|\mathbf{b}\| \|\mathbf{c}\|$. We apply the above argument to Eq. (21). First, $\|\pi_s^{(j,j+1)}(f)\|_2 \leq 3/2\delta_j\sqrt{n}$ for all $f \in \mathcal{F}$. In addition, $\|\mathbf{u}_s\| = 1$ for all s . Hence, for all $t \geq 2\delta_j\lambda_d^{-1/2}$,

$$\mathbb{P} \left[\frac{1}{\sqrt{n}} \left| \sum_{s=1}^d (\pi_s^{(j,j+1)}(f))^T (\mathbf{A} - \mathbf{P}) \mathbf{u}_s \lambda_s^{-1/2} \right| \geq dt \right] \leq 2d \exp \left(-\frac{t^2}{K \delta_j^2 \lambda_d^{-1}} \right)$$

for some universal constant $K > 0$. We therefore have

$$(22) \quad \mathbb{P} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n (\Pi_{j+1} f - \Pi_j f)(X_i)^T (\tilde{X}_i - X_i) \right| \geq dt \right] \leq 2d |\{\Pi_{j+1} f - \Pi_j f : f \in \mathcal{F}\}| \exp \left(-\frac{t^2}{K \delta_j^2 \lambda_d^{-1}} \right)$$

The set $\{\Pi_{j+1} f - \Pi_j f : f \in \mathcal{F}\}$ is of cardinality at most $|S_{j+1}|^2$; hence for any $t_j > 0$,

$$(23) \quad \mathbb{P} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n (\Pi_{j+1} f - \Pi_j f)(X_i)^T (\tilde{X}_i - X_i) \right| \geq d \sqrt{K \delta_j^2 \lambda_d^{-1} (t_j^2 + \log |S_{j+1}|^2)} \right] \leq 2d \exp(-t_j^2)$$

Summing Eq. (23) over all $j \geq 0$, we arrive at

$$\mathbb{P} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{\sqrt{n}} \text{tr}[\mathbf{M}(f)]^T (\mathbf{A} - \mathbf{P}) \mathbf{U}_\mathbf{P} \mathbf{S}_\mathbf{P}^{-1/2} \widetilde{\mathbf{W}} \right| \geq \sum_{j=0}^{\infty} d \sqrt{K \delta_j^2 \lambda_d^{-1} (t_j^2 + \log |S_{j+1}|^2)} \right] \leq 2d \sum_{j=0}^{\infty} \exp(-t_j^2)$$

We now bound the sum $\sum_{j=0}^{\infty} d \sqrt{K \delta_j^2 \lambda_d^{-1} (t_j^2 + \log |S_{j+1}|^2)}$. By compactness of Ω and smoothness of Φ , $|S_j|$ can be bounded in terms of the covering number for compact subsets of \mathbb{R}^d using balls of radius δ_j , i.e., $|S_j| \leq (L/\delta_j)^d$ for some constant L independent of δ_j (see e.g., Lemma 2.5 in [31]). Then by taking $t_j^2 = 2(\log(1/j) + \log n)$,

$$(24) \quad \mathbb{P} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{\sqrt{n}} \text{tr}[\mathbf{M}(f)]^T (\mathbf{A} - \mathbf{P}) \mathbf{U}_\mathbf{P} \mathbf{S}_\mathbf{P}^{-1/2} \widetilde{\mathbf{W}} \right| \geq d \lambda_d^{-1/2} (C_1 \log n + C_2) \right] \leq \frac{2dC_3}{n^2}$$

for some constants C_1 , C_2 and C_3 . Since there exists some constant $c > 0$ for which $\lambda_d/(cn) \rightarrow 1$ almost surely, an application of the Borel-Cantelli lemma to Eq. (24) yields

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{\sqrt{n}} \text{tr}[\mathbf{M}(f)]^T (\mathbf{A} - \mathbf{P}) \mathbf{U}_\mathbf{P} \mathbf{S}_\mathbf{P}^{-1/2} \widetilde{\mathbf{W}} \right| \xrightarrow{\text{a.s.}} 0$$

as $n \rightarrow \infty$. The term $n^{-1/2}c_0$ can be controlled by another application of Hoeffding's inequality. We thus have $\sup_{f \in \mathcal{F}} |\zeta(f)| \rightarrow 0$ as $n \rightarrow \infty$, thereby establishing Lemma 3.

Proof of Theorem 2. We shall prove that the difference

$$(25) \quad \Delta = (m+n)(V_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) - V_{n,m}(\mathbf{X}, \mathbf{Y})) \xrightarrow{\text{a.s.}} 0$$

converges to zero almost surely under the hypothesis $F = G$ (note the distinction with $F \perp G$), where $V_{n,m}(\mathbf{X}, \mathbf{Y})$ is as defined in Eq. (2) is a biased estimate of $\|\mu[F] - \mu[G]\|_{\mathcal{H}}^2$ and $V_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ is defined similarly to $U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$, i.e.,

$$V_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \kappa(\hat{X}_i, \hat{X}_j) - \frac{2}{mn} \sum_{i=1}^n \sum_{k=1}^m \kappa(\hat{X}_i, \hat{Y}_k) + \frac{1}{m^2} \sum_{k=1}^m \sum_{l=1}^m \kappa(\hat{Y}_k, \hat{Y}_l).$$

The claim in Eq. (7) follows from Eq. (25) and the following expression

$$(m+n)(V_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) - V_{n,m}(\mathbf{X}, \mathbf{Y})) = (m+n)(U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) - U_{n,m}(\mathbf{X}, \mathbf{Y})) + r_1 + r_2$$

where r_1 and r_2 are defined as

$$\begin{aligned} r_1 &= \frac{m+n}{n(n-1)} \sum_{i=1}^n \left(\kappa(X_i, X_i) - \kappa(\hat{X}_i, \hat{X}_i) \right) + \frac{m+n}{m(m-1)} \sum_{k=1}^m \left(\kappa(Y_k, Y_k) - \kappa(\hat{Y}_k, \hat{Y}_k) \right) \\ r_2 &= \frac{m+n}{n^2(n-1)} \sum_{i=1}^n \sum_{j=1}^n \left(\kappa(X_i, X_j) - \kappa(\hat{X}_i, \hat{X}_j) \right) + \frac{m+n}{m^2(m-1)} \sum_{k=1}^m \sum_{l=1}^m \left(\kappa(Y_k, Y_l) - \kappa(\hat{Y}_k, \hat{Y}_l) \right). \end{aligned}$$

As κ is twice continuously differentiable, we can show, by the compactness of Ω and the bounds in Lemma 2 that both r_1 and r_2 converges to 0 almost surely. In particular, there exists an L depending only on κ such that both $|r_1|$ and $|r_2|$ is bounded from above by

$$L(m+n) \left\{ \frac{\|\hat{\mathbf{X}}\mathbf{W} - \mathbf{X}\|_{2 \rightarrow \infty}}{n-1} + \frac{\|\hat{\mathbf{Y}}\mathbf{W} - \mathbf{Y}\|_{2 \rightarrow \infty}}{m-1} \right\}$$

We thus proceed to establishing Eq. (25). For a $d \times d$ orthogonal matrix \mathbf{W} , define $\xi, \hat{\xi}_W \in \mathcal{H}$ by

$$\begin{aligned} \xi &= \frac{\sqrt{m+n}}{n} \sum_{i=1}^n \kappa(X_i, \cdot) - \frac{\sqrt{m+n}}{m} \sum_{k=1}^m \kappa(Y_k, \cdot); \\ \hat{\xi}_W &= \frac{\sqrt{m+n}}{n} \sum_{i=1}^n \kappa(\mathbf{W}\hat{X}_i, \cdot) - \frac{\sqrt{m+n}}{m} \sum_{k=1}^m \kappa(\mathbf{W}\hat{Y}_k, \cdot). \end{aligned}$$

Note that

$$\left| (m+n)(V_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) - V_{n,m}(\mathbf{X}, \mathbf{Y})) \right| = \left| \|\xi\|_{\mathcal{H}}^2 - \|\hat{\xi}_W\|_{\mathcal{H}}^2 \right| \leq \|\xi - \hat{\xi}_W\|_{\mathcal{H}} (2\|\xi\|_{\mathcal{H}} + \|\xi - \hat{\xi}_W\|_{\mathcal{H}})$$

We now bound the terms $\|\xi - \hat{\xi}_W\|_{\mathcal{H}}$ and $\|\xi\|_{\mathcal{H}}$. We first bound $\|\xi\|_{\mathcal{H}}$. Under the hypothesis $F = G$,

$$\xi = \sqrt{\frac{m+n}{n}} \sum_{i=1}^n \frac{\kappa(X_i, \cdot) - \mu[F]}{\sqrt{n}} - \sqrt{\frac{m+n}{m}} \sum_{k=1}^m \frac{\kappa(Y_k, \cdot) - \mu[F]}{\sqrt{m}}.$$

That is, ξ is a sum of independent mean zero random elements of \mathcal{H} . In addition $\|\kappa(Z, \cdot) - \mu[F]\|_{\mathcal{H}} \leq 2$ for any $Z \in \mathbb{R}^d$. Using a Hilbert space concentration inequality [23], we obtain that

$$\mathbb{P}[\|\xi\|_{\mathcal{H}} \geq \sqrt{m+n}(s/\sqrt{n} + t/\sqrt{m})] \leq 2 \left(\exp(-(1+m/n)s^2/8) + \exp(-(1+n/m)t^2/8) \right),$$

which implies that $\|\xi\|_{\mathcal{H}}$ is bounded in probability. We now bound $\|\xi - \hat{\xi}_W\|_{\mathcal{H}}$. We begin by noting

$$\xi - \hat{\xi}_W = \sqrt{\frac{m+n}{n}} \sum_{i=1}^n \frac{\kappa(X_i, \cdot) - \kappa(\mathbf{W}\hat{X}_i, \cdot)}{\sqrt{n}} + \sqrt{\frac{m+n}{n}} \sum_{k=1}^m \frac{\kappa(Y_k, \cdot) - \kappa(\mathbf{W}\hat{Y}_k, \cdot)}{\sqrt{m}}$$

Denote by $(\partial^2 \Phi(Z)) \in \{f: \mathbb{R}^d \mapsto \mathbb{R}^{d^2}\}$ the Hessian of the feature map Φ at Z , i.e.,

$$(\partial^2 \Phi(Z))(\cdot) = \frac{\partial^2 \kappa(X, \cdot)}{\partial X^2} \Big|_{X=Z}$$

A Taylor expansion of κ allows us to conclude that

$$\sum_{i=1}^n \frac{(\Phi(X_i) - \Phi(\mathbf{W}\hat{X}_i))(\cdot)}{\sqrt{n}} = \sum_{i=1}^n \frac{(\partial \Phi(X_i))(\cdot)^\top (\mathbf{W}\hat{X}_i - X_i) + (\mathbf{W}\hat{X}_i - X_i)^\top (\partial^2 \Phi(X_i^*))(\cdot) (\mathbf{W}\hat{X}_i - X_i)}{\sqrt{n}}$$

where, for any i , $X_i^* \in \mathbb{R}^d$ is such that $\|X_i^* - X_i\| \leq \|\mathbf{W}\hat{X}_i - X_i\|$. The convergence to zero of the terms involving the gradient $\partial \Phi$ is a consequence of Lemma 3. To bound the quadratic terms, i.e. those depending on $\partial^2 \Phi$, we note that since $\kappa \in \mathcal{C}^2(\mathcal{X} \times \mathcal{X})$, $\sup_{Z \in \mathcal{X}} \|\partial^2 \Phi(Z)\|$ is bounded (the norm under consideration is the spectral norm on matrices). Therefore,

$$\sup_{Z \in \Omega} \left| \sum_{i=1}^n \frac{(\mathbf{W}\hat{X}_i - X_i)^\top (\partial^2 \Phi(X_i^*))(\mathbf{W}\hat{X}_i - X_i)}{\sqrt{n}} \right| \leq \frac{\sup_{Z \in \mathcal{X}} \|\partial^2 \Phi(Z)\|}{\sqrt{n}} \|\hat{\mathbf{X}}\mathbf{W} - \mathbf{X}\|_F^2 \xrightarrow{\text{a.s.}} 0$$

as $n \rightarrow \infty$. Similar derivations hold for $\kappa(Y_i, \cdot) - \kappa(\mathbf{W}\hat{Y}_i, \cdot)$. Therefore $\|\xi - \hat{\xi}_W\|_{\mathcal{H}} \rightarrow 0$ as required. Eq. (25) is thus established, under the hypothesis $F = G$. The general null hypothesis $F \perp G$ follows by transforming G into F through an appropriate unitary transformation.

Proof of Theorem 3. The first part of this proof parallels that of Theorem 2. We sketch here the requisite modifications. We first show

$$(26) \quad (m+n)(V_{n,m}(\hat{\mathbf{X}}/\hat{s}_X, \hat{\mathbf{Y}}/\hat{s}_Y) - V_{n,m}(\mathbf{X}/s_X, \mathbf{Y}/s_Y)) \xrightarrow{\text{a.s.}} 0.$$

Define $\xi, \hat{\xi}_W \in \mathcal{H}$ by

$$\begin{aligned} \xi &= \frac{\sqrt{m+n}}{n} \sum_{i=1}^n \kappa(X_i/s_X, \cdot) - \frac{\sqrt{m+n}}{m} \sum_{k=1}^m \kappa(Y_k/s_Y, \cdot) \\ \hat{\xi} &= \frac{\sqrt{m+n}}{n} \sum_{i=1}^n \kappa(\mathbf{W}\hat{X}_i/\hat{s}_X, \cdot) - \frac{\sqrt{m+n}}{m} \sum_{k=1}^m \kappa(\mathbf{W}\hat{Y}_k/\hat{s}_Y, \cdot). \end{aligned}$$

Defined r_1 and r_2 similar to that in the proof of Theorem 2, i.e.,

$$\begin{aligned} r_1 &= \frac{m+n}{n(n-1)} \sum_{i=1}^n \left\{ \kappa\left(\frac{\hat{X}_i}{\hat{s}_X}, \frac{\hat{X}_i}{\hat{s}_X}\right) - \kappa\left(\frac{X_i}{s_X}, \frac{X_i}{s_X}\right) \right\} + \frac{m+n}{m(m-1)} \sum_{k=1}^m \left\{ \kappa\left(\frac{\hat{Y}_k}{\hat{s}_Y}, \frac{\hat{Y}_k}{\hat{s}_Y}\right) - \kappa\left(\frac{Y_k}{s_Y}, \frac{Y_k}{s_Y}\right) \right\} \\ r_2 &= \frac{m+n}{n^2(n-1)} \sum_{i=1}^n \sum_{j=1}^n \left\{ \kappa\left(\frac{\hat{X}_i}{\hat{s}_X}, \frac{\hat{X}_j}{\hat{s}_X}\right) - \kappa\left(\frac{X_i}{s_X}, \frac{X_j}{s_X}\right) \right\} + \frac{m+n}{m^2(m-1)} \sum_{k=1}^m \sum_{l=1}^m \left\{ \kappa\left(\frac{\hat{Y}_k}{\hat{s}_Y}, \frac{\hat{Y}_l}{\hat{s}_Y}\right) - \kappa\left(\frac{Y_k}{s_Y}, \frac{Y_l}{s_Y}\right) \right\}. \end{aligned}$$

There exists an L depending only on κ such that both $|r_1|$ and $|r_2|$ is bounded from above by

$$L(m+n) \left\{ \frac{\|\mathbf{X} - \hat{\mathbf{X}}\mathbf{W}\|_{2 \rightarrow \infty}}{(n-1)\hat{s}_X} + \frac{|s_X - \hat{s}_X|}{(n-1)s_X \hat{s}_X} + \frac{\|\mathbf{Y} - \hat{\mathbf{Y}}\mathbf{W}\|_{2 \rightarrow \infty}}{(m-1)\hat{s}_Y} + \frac{|s_Y - \hat{s}_Y|}{(m-1)s_Y \hat{s}_Y} \right\}.$$

Lemma 2 implies $|r_1 + r_2| \rightarrow 0$ almost surely. When $F = G \circ c$, $\mu[F \circ s_X] = \mu[G \circ s_Y]$ and

$$\xi = \frac{\sqrt{m+n}}{\sqrt{n}} \sum_{i=1}^n \frac{\kappa(X_i/s_X, \cdot) - \mu[F \circ s_X]}{\sqrt{n}} + \frac{\sqrt{m+n}}{\sqrt{m}} \sum_{k=1}^m \frac{\kappa(Y_k/s_Y, \cdot) - \mu[G \circ s_Y]}{\sqrt{m}}$$

is a sum of mean zero elements of \mathcal{H} . The summands, however, are no longer independent due to the appearance of the factors s_X and s_Y . On the other hand, the random variable

$$\frac{\sqrt{m+n}}{\sqrt{n}} \sum_{i=1}^n \frac{\kappa(X_i/\sigma_X, \cdot) - \mu[F \circ \sigma_X]}{\sqrt{n}} + \frac{\sqrt{m+n}}{\sqrt{m}} \sum_{k=1}^m \frac{\kappa(Y_k/\sigma_Y, \cdot) - \mu[G \circ \sigma_Y]}{\sqrt{m}},$$

where $\sigma_X = (\mathbb{E}[\|X\|^2])^{1/2}$ and $\sigma_Y = (\mathbb{E}[\|Y\|^2])^{1/2}$, is a sum of independent mean zero elements of \mathcal{H} and is bounded in probability. In addition, s_X and s_Y are \sqrt{n} -consistent estimator of σ_X and σ_Y , respectively. One can thus show that $\|\xi\|_{\mathcal{H}}$ is also bounded in probability.

We next bound $\|\xi - \hat{\xi}_W\|_{\mathcal{H}}$. A Taylor expansion of κ yields

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (\Phi(\frac{X_i}{s_X}) - \Phi(\frac{\mathbf{W}\hat{X}_i}{\hat{s}_X}))(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \partial\Phi\left(\frac{X_i}{s_X}\right)(\cdot)^\top \left(\frac{\mathbf{W}\hat{X}_i}{\hat{s}_X} - \frac{X_i}{s_X}\right) + \left(\frac{\mathbf{W}\hat{X}_i}{\hat{s}_X} - \frac{X_i}{s_X}\right)^\top \partial^2\Phi\left(\frac{X_i^*}{s_X}\right)(\cdot) \left(\frac{\mathbf{W}\hat{X}_i}{\hat{s}_X} - \frac{X_i}{s_X}\right)$$

The terms depending on $\partial^2\Phi$ in the above display is bounded as

$$\begin{aligned} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\frac{\mathbf{W}\hat{X}_i}{\hat{s}_X} - \frac{X_i}{s_X}\right)^\top \partial^2\Phi\left(\frac{X_i^*}{s_X}\right)(\cdot) \left(\frac{\mathbf{W}\hat{X}_i}{\hat{s}_X} - \frac{X_i}{s_X}\right) \right| &\leq \frac{\sup_{Z \in \Omega} \|\partial^2\Phi(Z)\|}{\sqrt{n}} \sum_{i=1}^n \left\| \frac{\mathbf{W}\hat{X}_i}{\hat{s}_X} - \frac{X_i}{s_X} \right\|^2 \\ &\leq \frac{2 \sup_{Z \in \Omega} \|\partial^2\Phi(Z)\| \|\mathbf{W}\hat{\mathbf{X}} - \mathbf{X}\|_F^2}{\sqrt{n}(\hat{s}_X)^2} \end{aligned}$$

which converges to 0 almost surely. For the terms depending on $\partial\Phi$, we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \partial\Phi\left(\frac{X_i}{s_X}\right)(\cdot)^\top \left(\frac{\mathbf{W}\hat{X}_i}{\hat{s}_X} - \frac{X_i}{s_X}\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \partial\Phi\left(\frac{X_i}{s_X}\right)(\cdot)^\top \frac{\mathbf{W}\hat{X}_i - X_i}{\hat{s}_X} + \frac{1}{\sqrt{n}} \sum_{i=1}^n \partial\Phi\left(\frac{X_i}{s_X}\right)(\cdot)^\top X_i \left(\frac{\hat{s}_X - s_X}{\hat{s}_X s_X}\right)$$

The first sum on the right hand side of the above display can be bounded using a chaining argument identical to that in the proof of Theorem 2 and an application of Slutsky's theorem (for $\hat{s}_X \rightarrow (\mathbb{E}[\|X\|^2])^{1/2}$ almost surely). For the second sum on the right hand side, we have

$$\begin{aligned} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \partial\Phi\left(\frac{X_i}{s_X}\right)(\cdot)^\top X_i \left(\frac{\hat{s}_X - s_X}{\hat{s}_X s_X}\right) \right| &= \left| \frac{1}{n} \sum_{i=1}^n \Phi\left(\frac{X_i}{s_X}\right)(\cdot)^\top X_i \sqrt{n} \left(\frac{\hat{s}_X^2 - s_X^2}{(\hat{s}_X + s_X)\hat{s}_X s_X}\right) \right| \\ &\leq \frac{\sup_{Z, Z' \in \Omega} |(\partial\Phi(Z))(Z')^\top|}{\sqrt{n}} \frac{\|\hat{\mathbf{X}}\|_F^2 - \|\mathbf{X}\|_F^2}{(\hat{s}_X + s_X)\hat{s}_X s_X}. \end{aligned}$$

Hence by Lemma 2, compactness of Ω , smoothness of Φ and Slutsky's theorem, the second sum also converges to 0 almost surely, thereby establishing Eq. (26).

We now derive the limiting distribution of $(m+n)V_{n,m}(\mathbf{X}/s_X, \mathbf{Y}/s_Y)$. Under the null hypothesis, there exists some $\mathbf{X}' = \{X'_1, \dots, X'_m\}$ with the $\{X'_i\}$ independent of the $\{X_i\}$ such that $V_{n,m}(\mathbf{X}/s_X, \mathbf{Y}/s_Y) = V_{n,m}(\mathbf{X}/s_X, \mathbf{X}'/s_{X'})$. Hence we shall be working with two independent collection \mathbf{X} and \mathbf{X}' of i.i.d. random variables from F . Let $\xi_{n,m} = V_{n,m}(\mathbf{X}/s_X, \mathbf{X}'/s_{X'})$. Then

$$\xi_{n,m} = \int_{\mathbb{R}^d} \left(\frac{1}{n} \sum_{i=1}^n g(\omega, X_i; s_X) - \frac{1}{m} \sum_{k=1}^m g(\omega, X'_k; s_{X'}) \right)^2 dM(\omega).$$

For a given $A > 0$, denote by $\xi_{n,m,A}$ the quantity

$$\int_{\|\omega\|>A} \left(\frac{1}{n} \sum_{i=1}^n g(\omega, X_i; s_X) - \frac{1}{m} \sum_{k=1}^m g(\omega, X'_k; s_{X'}) \right)^2 dM(\omega).$$

Let $\eta_A := \lim_{n,m \rightarrow \infty} (m+n)\xi_{n,m,A}$. By Theorem 3.2 in [4], $(m+n)\xi_{n,m}$ has the same limiting distribution as η_A when $A \rightarrow \infty$ if

$$(27) \quad \lim_{A \rightarrow \infty} \limsup_{n,m \rightarrow \infty} \mathbb{P}[|(m+n)(\xi_{n,m,A} - \xi_{n,m})| \geq \epsilon] = 0$$

for all ϵ . We show Eq. (27) using an argument almost identical to that given in [14]. We have

$$\begin{aligned} \mathbb{E}[(\xi_{n,m} - \xi_{n,m,A})^2] &\leq \frac{2}{n} \int_{\|\omega\|>A} \text{Var}[g(\omega, X_1; s_X)] dM(\omega) + \frac{2}{m} \int_{\|\omega\|>A} \text{Var}[g(\omega, X'_1; s_{X'})] dM(\omega) \\ &\leq \frac{2(m+n)}{mn} \int_{\|\omega\|>A} dM(\omega). \end{aligned}$$

and hence, by Chebyshev's inequality,

$$\limsup_{n,m \rightarrow \infty} \mathbb{P}[|(m+n)(\xi_{n,m} - \xi_{n,m,A})| \geq \epsilon] \leq \limsup_{n,m \rightarrow \infty} \frac{2(m+n)^2}{\epsilon^2 mn} \int_{\|\omega\|>A} dM(\omega) \rightarrow 0$$

as $A \rightarrow \infty$. We now derive the limiting distribution of η_A . Uniformly for all ω such that $\|\omega\| \leq A$,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n g(\omega, X_i; s_X) &= \frac{1}{n} \sum_{i=1}^n g(\omega, X_i; \sigma_X) - \frac{1}{n} \sum_{i=1}^n \frac{d}{du} [g(\omega, X_i; u)]_{u=\sigma_X} (s_X - \sigma_X) + O((s_X - \sigma_X)^2) \\ &= \frac{1}{n} \sum_{i=1}^n g(\omega, X_i; \sigma_X) - \frac{d}{du} [\mathbb{E}[g(\omega, X_1; u)]]_{u=\sigma_X} \frac{s_X^2 - \sigma_X^2}{2\sigma_X} + o_{\mathbb{P}}(n^{-1/2}) \end{aligned}$$

where the expectation $\mathbb{E}[g(\omega, X_1; u)]$ is taken with respect to X_1 . A similar derivation holds for $m^{-1} \sum_{k=1}^m g(\omega, X'_k; s_{X'})$. Let $\kappa_A^*(x, x')$ be the kernel defined by

$$\kappa_A^*(x, x') = \int_{\|w\| \leq A} \left(\tilde{g}(\omega, x; \sigma_X) - h(\omega; \sigma_X) \frac{\|x\|^2 - \sigma_X^2}{2\sigma_X} \right) \left(\tilde{g}(\omega, x'; \sigma_X) - h(\omega; \sigma_X) \frac{\|x'\|^2 - \sigma_X^2}{2\sigma_X} \right) dM(\omega).$$

Now s_X^2 and $s_{X'}^2$ are \sqrt{n} -consistent estimators for σ_X^2 . Therefore, under the null hypothesis,

$$\xi_{n,m,A} = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \kappa_A^*(X_i, X_j) - \frac{2}{mn} \sum_{i=1}^n \sum_{k=1}^m \kappa_A^*(X_i, Y_k) + \frac{1}{m^2} \sum_{k=1}^m \sum_{l=1}^m \kappa_A^*(Y_k, Y_l) + o_{\mathbb{P}}(1/n)$$

Thus for any fixed but arbitrary A , distributional convergence of degenerate V -statistics yields

$$\lim_{n,m \rightarrow \infty} (n+m)\xi_{n,m,A} \xrightarrow{d} \frac{1}{\rho(1-\rho)} \sum_{l=1}^{\infty} \lambda_{l,A}^* \chi_{1l}^2$$

where $\lambda_{l,A}^*$ are the eigenvalues of the integral operator $\mathcal{I}_{F, \kappa_A^*}$ (see [25, §6.4]). Letting $A \rightarrow \infty$ along with Eq. (27), we deduce that

$$(28) \quad \lim_{n,m \rightarrow \infty} (n+m)\xi_{n,m} \xrightarrow{d} \frac{1}{\rho(1-\rho)} \sum_{l=1}^{\infty} \lambda_l^* \chi_{1l}^2$$

where the $\{\lambda_l^*\}$ are the eigenvalues of the integral operator $\mathcal{I}_{F, \kappa^*}$. Eq. (12) follows from Eq. (28) and straightforward calculation.

Proof of Theorem 4. The proof of this result is almost identical to that of Theorem 2. We note here the requisite modifications. Define $\xi, \hat{\xi}_W \in \mathcal{H}$ by

$$\begin{aligned}\xi &= \frac{\sqrt{m+n}}{n} \sum_{i=1}^n \kappa(\pi(X_i), \cdot) - \frac{\sqrt{m+n}}{m} \sum_{k=1}^m \kappa(\pi(Y_k), \cdot) \\ \hat{\xi}_W &= \frac{\sqrt{m+n}}{n} \sum_{i=1}^n \kappa(\pi(\mathbf{W}\hat{X}_i), \cdot) - \frac{\sqrt{m+n}}{m} \sum_{k=1}^m \kappa(\pi(\mathbf{W}\hat{Y}_k), \cdot).\end{aligned}$$

In addition, define r_1 and r_2 by

$$\begin{aligned}r_1 &= \frac{m+n}{n(n-1)} \sum_{i=1}^n \left(\kappa(\pi(X_i), \pi(X_i)) - \kappa(\pi(\hat{X}_i), \pi(\hat{X}_i)) \right) + \frac{m+n}{m(m-1)} \sum_{k=1}^m \left(\kappa(\pi(Y_k), \pi(Y_k)) - \kappa(\pi(\hat{Y}_k), \pi(\hat{Y}_k)) \right) \\ r_2 &= \frac{m+n}{n^2(n-1)} \sum_{i,j} \left(\kappa(\pi(X_i), \pi(X_j)) - \kappa(\pi(\hat{X}_i), \pi(\hat{X}_j)) \right) + \frac{m+n}{m^2(m-1)} \sum_{k,l} \left(\kappa(\pi(Y_k), \pi(Y_l)) - \kappa(\pi(\hat{Y}_k), \pi(\hat{Y}_l)) \right)\end{aligned}$$

Using the assumption $\|Z\| \geq c_0$ F -almost everywhere for some constant $c_0 > 0$, both $|r_1|$ and $|r_2|$ can be bounded from above by

$$L(m+n) \left\{ \frac{2\|\mathbf{X} - \hat{\mathbf{X}}\mathbf{W}\|_{2 \rightarrow \infty}}{(n-1)c_0} + \frac{2\|\mathbf{Y} - \hat{\mathbf{Y}}\mathbf{W}\|_{2 \rightarrow \infty}}{(m-1)c_0} \right\}$$

for some constant L depending only on κ . We bound $\|\xi - \hat{\xi}_W\|_{\mathcal{H}}$ by applying the chaining argument in the proof of Theorem 2 to the family of functions

$$\mathcal{F} = \{f = (\partial(\Phi \circ \pi)(\cdot))(Z) : Z \in \Omega\}.$$

The limiting distribution of $(m+n)U_{n,m}(\pi(\mathbf{X}), \pi(\mathbf{Y}))$ follows immediately from the definition of $\tilde{\kappa}^{(\pi)}$ and limiting distribution of degenerate U -statistics (see [12] or [25, Section 5.5]).

DEPARTMENT OF APPLIED MATHEMATICS AND STATISTICS,
JOHNS HOPKINS UNIVERSITY,
3400 N. CHARLES ST, BALTIMORE, MD 21218, USA.
E-MAIL: mtang10@jhu.edu; dathrey1@jhu.edu; vlyzins1@jhu.edu; cep@jhu.edu

DEPARTMENT OF STATISTICS,
HARVARD UNIVERSITY,
ONE OXFORD STREET, CAMBRIDGE, MA, 02138, USA
E-MAIL: daniellsussman@fas.harvard.edu