Bernoulli **23**(3), 2017, 1599–1630
DOI: 10.3150/15-BEJ789

A nonparametric two-sample hypothesis testing problem for random graphs

MINH TANG[1,*], AVANTI ATHREYA[1,**], DANIEL L. SUSSMAN[2], VINCE LYZINSKI[3] and CAREY E. PRIEBE[1,†]

[1]*Department of Applied Mathematics and Statistics, Johns Hopkins University.*
E-mail: [*]*mtang10@jhu.edu;* [**]*dathrey1@jhu.edu;* [†]*cep@jhu.edu*
[2]*Department of Statistics, Harvard University. E-mail: daniellsussman@fas.harvard.edu*
[3]*Human Language Technology Center of Excellence, Johns Hopkins University. E-mail: vlyzins1@jhu.edu*

We consider the problem of testing whether two independent finite-dimensional random dot product graphs have generating latent positions that are 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 estimated latent positions obtained from the adjacency spectral embedding for each graph. We show that our test statistic using the estimated latent positions converges to the test statistic obtained using the true but unknown latent positions and hence that our proposed test procedure is consistent across a broad range of alternatives. Our proof of consistency hinges upon a novel concentration inequality for the suprema of an empirical process in the estimated latent positions setting.

Keywords: empirical process; nonparametric graph inference; random dot product graph

1. Introduction

The nonparametric two-sample hypothesis testing problem involves

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

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 $\mathbf{B}$ (see Definition 2), we construct test statistics $T(\{\hat{X}\}_{i=1}^n, \{\hat{Y}_k\}_{k=1}^m)$ for testing $F = G$ (and related hypotheses) that are 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 problem may be viewed as the nonparametric analogue of the semiparametric inference problem considered in [32], in which a valid test is given for the hypothesis

that two random dot product graphs have the same fixed latent positions. This formulation also includes, as a special case, a test for the parametric problem of whether two graphs come from the same stochastic blockmodel (where the block probability matrix is positive semidefinite) or from the same degree-corrected stochastic blockmodel. Determining whether two random graphs are "similar" in an appropriate sense is a problem that arises naturally in neuroscience, network analysis, and machine learning. Examples include the comparison of graphs in a time series, such as email correspondence among a group over time, the comparison of neuroimaging scans of patients under varying conditions, or the comparison of user behavior on different social media platforms.

While it might seem like there are only minor differences between the nonparametric setting of the current paper and the semiparametric setting of [32], the implications with regard to inference are quite significant. Indeed, in the semiparametric setting, the graphs are on the same vertex set with known vertex alignment; in the nonparametric setting we consider herein, the graphs need not be on the same vertex set or even have the same number of vertices. This difference implies that the nonparametric testing procedure of the current paper is applicable in more general and diverse settings; on the other hand, when the vertex correspondences exist and are known, the semiparametric testing procedure has more power. Second, in the semiparametric setting, the dimensionality of the hypotheses (the number of parameters) increases with n, the number of vertices, while in the current setup the hypotheses are fixed for all n. As such, the notion of a consistent test procedure in [32] is considerably more subtle. Finally, while rejection regions can be theoretically derived for the test procedures in both the nonparametric setting and the semiparametric setting, in practice they are usually estimated via some bootstrap resampling procedure. For the nonparametric setting wherein the null hypothesis is fixed as the size of the graphs changes, bootstrap resampling is straightforward. A feasible bootstrapping procedure in the semiparametric setting is much more involved.

The test statistic we construct is an empirical estimate of the maximum mean discrepancy of [11]. 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., [3]). The test statistic can also be framed as a weighted L_2-distance between empirical estimates of characteristic functions similar to those of [2,4,12]. 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.

The results of this paper are mainly for dense graphs, that is, those graphs for which the average degree scale linearly with the number of vertices. Analogous results for non-dense graphs, for example, those for which the average degree of the vertices grows at order $\Omega(\log^4 n)$ – n being the number of vertices in the graph – are more subtle and we touch upon this briefly in Section 5.

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 a nonparametric two-sample test of equality of distributions for the latent positions of a pair of random dot product graphs. In Section 3, we propose

a test procedure for the two-sample test of equality up to orthogonal transformation in which the test statistics are a function of the adjacency spectral embedding. We note that our hypotheses of equality are purely a function of the non-identifiability of the random dot product graph model. This non-identifiability also restricts our consideration of kernel-based hypothesis testing to radial kernels. We establish the consistency of our test procedure by deriving a novel concentration inequality for the suprema of an empirical process using the estimated latent positions. In Section 4, we illustrate our test procedure with experimental results on simulated and real data. Section 5 extends the test procedure in Section 3 to consider looser notions of equality between the two distributions as well as sparsity in the underlying graphs model.

2. Background and setting

We first recall the notion of a random dot product graph [37].

Definition 1. *Let Ω be a subset of $\mathbb{R}^d$ such that, for all $\omega_1, \omega_2 \in \Omega$, the inner product $\langle \omega_1, \omega_2 \rangle = \omega_1^\top \omega_2$ is contained in the interval $[0, 1]$. For any given $n \geq 1$, let $\mathbf{X} = [X_1, X_2, \ldots, X_n]^\top$ 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*

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

$\mathbf{A}$ *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* $\mathbf{A} \sim \mathrm{RDPG}(\mathbf{X})$. *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* $(\mathbf{X}, \mathbf{A}) \sim \mathrm{RDPG}(F)$. *We shall also write* $\mathbf{A} \sim \mathrm{RDPG}(F)$ *when the dependency of* $\mathbf{A}$ *on* $\mathbf{X}$ *is integrated out*.

As an example of random dot product graphs, one could take Ω to be the unit simplex in $\mathbb{R}^d$ and let F be a mixture of Dirichlet distributions. 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* [14], 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 with i.i.d. latent positions on n vertices is also, when viewed as an induced subgraph of an infinite graph, an example of an *exchangeable random graph* [7]. Random dot product graphs are related to stochastic block model graphs [15] and degree-corrected stochastic block model graphs [16], 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 a property of nearly all exchangeable random graph models, and specifically, it 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 (here, for a random variable $X \sim F$, we write $F \circ U$ to denote the distribution of $Y = U^\top X$).

We now define the notion of adjacency spectral embedding; this is the key intermediate step in our subsequent two-sample hypothesis testing procedures.

Definition 2. *Let* $\mathbf{A}$ *be a* $n \times n$ *adjacency matrix. Suppose the eigendecomposition of* $|\mathbf{A}| = (\mathbf{A}^\top \mathbf{A})^{1/2}$ *is given by*

$$|\mathbf{A}| = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top$$

with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ *being the eigenvalues of* $|\mathbf{A}|$ *and* $\mathbf{u}_1, \ldots, \mathbf{u}_n$ *the corresponding eigenvectors. Given a positive integer* $d \leq n$*, denote by* $\mathbf{S}_\mathbf{A} = \text{diag}(\lambda_1, \ldots, \lambda_d)$ *the diagonal matrix whose diagonal entries are* $\lambda_1, \ldots, \lambda_d$ *and denote by* $\mathbf{U}_\mathbf{A}$ *the* $n \times d$ *matrix whose columns are the corresponding eigenvectors* $\mathbf{u}_1, \ldots, \mathbf{u}_d$*. The* adjacency spectral embedding $\mathbf{A}$ *into* $\mathbb{R}^d$ *is then the* $n \times d$ *matrix* $\hat{\mathbf{X}} = \mathbf{U}_\mathbf{A} \mathbf{S}_\mathbf{A}^{1/2}$.

Remark. The intuition behind the notion of adjacency spectral embedding is as follows. We note that if $(\mathbf{A}, \mathbf{X}) \sim \text{RDPG}(F)$, then the upper triangular entries of $\mathbf{A} - \mathbf{XX}^\top$ are independent random variables. Let $\|\cdot\|$ denote the spectral norm of a matrix. Then one can show that $\|\mathbf{A} - \mathbf{XX}^\top\| = O(\|\mathbf{X}\|) = o(\|\mathbf{XX}^\top\|)$ with high probability [22]. That is to say, $\mathbf{A}$ can be viewed as a "small" perturbation of $\mathbf{XX}^\top$. If we now assume that $\mathbf{X}$ is of rank d for some d – an assumption that is justified in the random dot product graphs model – then the Davis–Kahan theorem [6] implies that the subspace spanned by the top d eigenvectors of $\mathbf{XX}^\top$ is well-approximated by the subspace spanned by the top d eigenvectors of $\mathbf{A}$. In particular, the eigendecomposition of $\mathbf{XX}^\top$ recovers the matrix $\mathbf{X}$ up to an orthogonal transformation; hence the adjacency spectral embedding of $\mathbf{A}$ is expected to yield a consistent estimate of $\mathbf{X}$ up to an orthogonal transformation (see Lemma 2).

2.1. Two-sample hypothesis testing

In this paper, we propose a nonparametric version of the two-sample hypothesis test examined in [32]. To wit, [32] 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

$$H_0^n : \mathbf{X}_n \doteq \mathbf{Y}_n \quad \text{against} \quad H_A^n : \mathbf{X}_n \not\doteq \mathbf{Y}_n,$$

where $\doteq$ denotes that there exists an $\mathbf{W} \in \mathcal{O}(d)$ such that $\mathbf{X}_n = \mathbf{Y}_n\mathbf{W}$. In [32], 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 $\{\mathbf{X}_n\}$, $\{\mathbf{Y}_n\}$ for which $\min_{\mathbf{W}\in\mathcal{O}(d)} \|\mathbf{X}_n - \mathbf{Y}_n\mathbf{W}\|$ diverges as $n \to \infty$.

Our main point of departure in this work is the assumption that, for each n, the rows of the latent positions $\mathbf{X}_n$ and $\mathbf{Y}_n$ are independent samples from some fixed distributions F and G, respectively. The corresponding tests are therefore tests of equality between F and G. More formally, we consider the following two-sample nonparametric testing problems for random dot product graphs. Let F and G be probability distributions on $\Omega \subset \mathbb{R}^d$ for some d. Given $\mathbf{A} \sim \text{RDPG}(F)$ and $\mathbf{B} \sim \text{RDPG}(G)$, we consider the tests:

1. *Equality*, *up to orthogonal transformation*

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

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

2. *Equality*, *up to scaling*

$$H_0: F \doteq G \circ c \qquad \text{for some } c > 0 \quad \text{against} \quad H_A: F \not\doteq G \circ c \qquad \text{for any } c > 0,$$

where $Y \sim F \circ c$ if $cY \sim F$.

3. *Equality*, *up to projection*

$$H_0: F \circ \pi^{-1} \doteq G \circ \pi^{-1} \quad \text{against} \quad H_A: F \circ \pi^{-1} \not\doteq G \circ \pi^{-1},$$

where π is the projection $x \mapsto x/\|x\|$; hence $Y \sim F \circ \pi^{-1}$ if $\pi^{-1}(Y) \sim F$.

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

2.2. Maximum mean discrepancy

We now introduce the notion of the maximum mean discrepancy between two distribution [11]. The maximum mean discrepancy is a distance measure for probability distributions and hence can be used to construct a non-parametric two-sample hypothesis testing procedure (see Theorem 1 below). The maximum mean discrepancy is just one of several examples of kernel-based testing procedures; see [13] for a recent survey of the literature and for a more detailed discussion.

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

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

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

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

We summarize some important properties of the maximum mean discrepancy from [11]. 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.

Theorem 1. *Let $\kappa: \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ be a positive definite kernel and denote by $\mathcal{H}$ the reproducing kernel Hilbert space associated with κ. Let F and G be probability distributions on Ω; X and X' independent random variables with distribution F, Y and Y' independent random variables with distribution G, and X is independent of Y. Then*

$$\begin{aligned}\big\|\mu[F] - \mu[G]\big\|_{\mathcal{H}}^2 &= \sup_{h \in \mathcal{H}: \|h\|_{\mathcal{H}} \leq 1} \big|\mathbb{E}_F[h] - \mathbb{E}_G[h]\big|^2 \\ &= \mathbb{E}\big[\kappa\big(X, X'\big)\big] - 2\mathbb{E}\big[\kappa(X, Y)\big] + \mathbb{E}\big[\kappa\big(Y, Y'\big)\big].\end{aligned} \tag{2.1}$$

Given $\mathbf{X} = \{X_i\}_{i=1}^n$ and $\mathbf{Y} = \{Y_k\}_{k=1}^m$ with $\{X_i\} \overset{\text{i.i.d.}}{\sim} F$ and $\{Y_i\} \overset{\text{i.i.d.}}{\sim} G$, the quantity $U_{n,m}(\mathbf{X}, \mathbf{Y})$ defined by

$$\begin{aligned}U_{n,m}(\mathbf{X}, \mathbf{Y}) = \frac{1}{n(n-1)} \sum_{j \neq i} \kappa(X_i, X_j) - \frac{2}{mn} \sum_{i=1}^{n} \sum_{k=1}^{m} \kappa(X_i, Y_k) \\ + \frac{1}{m(m-1)} \sum_{l \neq k} \kappa(Y_k, Y_l)\end{aligned} \tag{2.2}$$

is an unbiased consistent estimate of $\|\mu[F] - \mu[G]\|_{\mathcal{H}}^2$. Denote by $\tilde{\kappa}$ the kernel

$$\tilde{\kappa}(x, y) = \kappa(x, y) - \mathbb{E}_z \kappa(x, z) - \mathbb{E}_{z'} \kappa\big(z', y\big) + \mathbb{E}_{z, z'} \kappa\big(z, z'\big),$$

where the expectation is taken with respect to $z, z' \sim F$. Suppose that $\frac{m}{m+n} \to \rho \in (0, 1)$ as $m, n \to \infty$. Then under the null hypothesis of $F = G$,

$$(m+n) U_{n,m}(\mathbf{X}, \mathbf{Y}) \xrightarrow{d} \frac{1}{\rho(1-\rho)} \sum_{l=1}^{\infty} \lambda_l \big(\chi_{1l}^2 - 1\big), \tag{2.3}$$

where $\{\chi^2_{1l}\}_{l=1}^{\infty}$ *is a sequence of independent* χ^2 *random variables with one degree of freedom, and* $\{\lambda_l\}$ *are the eigenvalues of the integral operator* $\mathcal{I}_{F,\tilde{\kappa}} : \mathcal{H} \mapsto \mathcal{H}$ *defined as*

$$I_{F,\tilde{\kappa}}(\phi)(x) = \int_{\Omega} \phi(y)\tilde{\kappa}(x, y)\, dF(y).$$

Finally, if κ *is a universal or characteristic kernel* [25,26], *then* μ *is an injective map, that is,* $\mu[F] = \mu[G]$ *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 $\mathcal{M}$ [25]. 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 equation (2.1), the maximum mean discrepancy with reproducing kernel k_q can be written as

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

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 [31], or, when $q = 1$, a special case of the one-dimensional interpoint comparisons of [21].

Remark. The limiting distribution of $(m + n)U_{n,m}(\mathbf{X}, \mathbf{Y})$ under the null hypothesis of $F = G$ in Theorem 1 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

We now address the nonparametric two-sample hypothesis tests of Section 2.1 using the methodology described in Section 2.2. Throughout, we shall always assume that the distributions of the latent positions satisfy the following distinct eigenvalues assumption. The assumption implies that the estimates of the latent position obtained by the adjacency spectral embedding in Definition 2 will, in the limit, be uniquely determined.

Assumption 1. *The distribution* F *for the latent positions* $X_1, X_2, \ldots, \sim F$ *is such that the second moment matrix* $\mathbb{E}[X_1 X_1^\top]$ *has* d *distinct eigenvalues and* d *is known.*

The motivation behind this assumption is as follows: the matrix $\mathbb{E}[X_1 X_1^\top]$ is of rank d with d known so that given a graph $\mathbf{A} \sim \mathrm{RDPG}(F)$, one can construct the adjacency spectral embedding

of $\mathbf{A}$ into the "right" Euclidean space. The requirement that $\mathbb{E}[X_1 X_1^\top]$ has d distinct eigenvalues is due to the intrinsic property of non-identifiability of random dot product graphs, that is, for any random dot product graph $\mathbf{A}$, the latent position $\mathbf{X}$ associated with $\mathbf{A}$ can only be estimated up to some true but unknown orthogonal transformation. Because we are concerned with two-sample hypothesis testing, we must guard against the scenario in which we have two graphs $\mathbf{A}$ and $\mathbf{B}$ with latent positions $\mathbf{X} = \{X_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} F$ and $\mathbf{Y} = \{Y_k\}_{k=1}^m \overset{\text{i.i.d.}}{\sim} F$ but their estimates $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$ lie in different, incommensurate subspaces of $\mathbb{R}^d$. That is to say, the estimates $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$ satisfy $\hat{\mathbf{X}} \approx \mathbf{X}\mathbf{W}_1$ and $\hat{\mathbf{Y}} \approx \mathbf{Y}\mathbf{W}_2$, but $\|\mathbf{W}_1 - \mathbf{W}_2\|_F$ does not converge to 0 as $n, m \to \infty$. See also [10] for exposition of a related so-called "incommensurability phenomenon."

Indeed, we recognize that Assumption 1 is restrictive; in particular, it is not satisfied by the stochastic block model with $K > 2$ blocks of equal size and edge probabilities p within communities and q between communities. However, we are not aware of any two-sample nonparametric inference procedure in which the incommensurability problem is resolved, and Assumption 1 still permits two-sample nonparametric inference on a wide class of random graphs.

Remark. This issue of incommensurability is an intrinsic feature of many dimension reduction techniques, and is not simply an artificial complication that arises in graph estimation. Consider, for example, principal component analysis in the following setting. Let $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n \times d}$ and suppose that the rows of $\mathbf{X}$ and $\mathbf{Y}$ are i.i.d. from some distribution F. Furthermore, suppose that $\mathbf{X}$ and $\mathbf{Y}$ are unobserved, but instead $\mathbf{X}$ and $\mathbf{Y}$ are to be estimated or recovered from some higher dimension data $\mathbf{X}^* = [\mathbf{X} \mid \mathbf{Z}] \in \mathbb{R}^{n \times D}$, and $\mathbf{Y}^* = [\mathbf{Y} \mid \mathbf{Z}'] \in \mathbb{R}^{n \times D}$, say via principal component analysis, where $\mathbf{Z}$ and $\mathbf{Z}'$ are $n \times (D - d)$ matrices whose rows are i.i.d. from some other distribution H. That is to say, $\mathbf{X}$ is recovered via principal component analysis of $\mathbf{X}^*$ into $\mathbb{R}^d$ and similarly for $\mathbf{Y}$. Then depending on the covariance structure of F and H, the recovered $\mathbf{X}$ and $\mathbf{Y}$ could lie in incommensurate subspaces.

3.1. Two technical lemmas

We now state two technical lemmas. The first lemma is the culmination of results from [20] and [32]. The second lemma lays the foundation for an empirical process result and is also a central ingredient for showing the convergence to zero of a suitably scaled version of our test statistic in the two-sample setting.

Lemma 2. *Let* $(\mathbf{X}, \mathbf{A}) \sim \text{RDPG}(F)$ *be a* d*-dimensional random dot product graph on* n *vertices with latent position distributions* F *satisfying the conditions in Assumption* 1*. Let* $c > 0$ *be arbitrary but fixed. There exists* $n_0(c)$ *such that if* $n \geq n_0$ *and* η *satisfies* $n^{-c} < \eta < 1/4$*, then there exists an orthogonal matrix* $\mathbf{W}$ *dependent on* $\mathbf{X}$ *such that, with probability at least* $1 - 4\eta$*,*

$$\|\hat{\mathbf{X}} - \mathbf{X}\mathbf{W}\|_F \leq C_1, \tag{3.1}$$

$$\|\hat{\mathbf{X}} - \mathbf{X}\mathbf{W}\|_{2\to\infty} \leq C_2 \sqrt{\frac{\log(n/\eta)}{n}}, \tag{3.2}$$

where C_1 *and* C_2 *are constants depending only on* F *and* $n_0(c)$*.*

Lemma 2 bounds the difference between $\hat{\mathbf{X}}$ and $\mathbf{X}$ namely the Frobenius norm $\|\cdot\|_F$ and the maximum of the l_2 norms of the rows $\|\cdot\|_{2\to\infty}$. The norm $\|\cdot\|_{2\to\infty}$ is induced by the vector norms $\|\cdot\|_2$ and $\|\cdot\|_\infty$ via $\|\mathbf{A}\|_{2\to\infty} = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_\infty$. Equation (3.2) follows from Lemma 2.5 in [20] while equation (3.1) follows from Theorem 2.3 in [32].

As a quick application of Lemma 2, suppose $(\mathbf{X}, \mathbf{A}) \sim \mathrm{RDPG}(F)$ and $(\mathbf{Y}, \mathbf{B}) \sim \mathrm{RDPG}(G)$ where the latent position distributions F and G satisfy the distinct eigenvalues assumption and consider the hypothesis test of $\mathbb{H}_0$: $F \doteq G$. Let κ be a differentiable radial kernel and $U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ is defined as

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

Then there exists a deterministic unitary matrix $\mathbf{W}_0$ such that

$$U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) - U_{n,m}(\mathbf{X}, \mathbf{Y}\mathbf{W}_0) \to 0$$

almost surely as $n, m \to \infty$. This can be seen as follows. Let $\mathbf{W}_n$ and $\mathbf{V}_m$ be orthogonal matrices in the eigendecomposition $\mathbf{W}_n\mathbf{S}_1\mathbf{W}_n = \mathbf{X}^\top\mathbf{X}$, $\mathbf{V}_m\mathbf{S}_2\mathbf{V}_m = \mathbf{Y}^\top\mathbf{Y}$, respectively. Then

$$\begin{aligned} U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) - U_{n,m}(\mathbf{X}\mathbf{W}_n, \mathbf{Y}\mathbf{V}_m) &= \frac{1}{n(n-1)} \sum_{j\neq i} \big(\kappa(\hat{X}_i, \hat{X}_j) - \kappa(\mathbf{W}_n X_i, \mathbf{W}_n X_j)\big) \\ &\quad - \frac{2}{mn} \sum_{i=1}^{n}\sum_{k=1}^{m} \big(\kappa(\hat{X}_i, \hat{Y}_k) - \kappa(\mathbf{W}_n X_i, \mathbf{V}_m Y_k)\big) \\ &\quad + \frac{1}{m(m-1)} \sum_{l\neq k} \big(\kappa(\hat{Y}_k, \hat{Y}_l) - \kappa(\mathbf{V}_m Y_k, \mathbf{V}_m Y_l)\big). \end{aligned}$$

By differentiability of κ and compactness of Ω, we have

$$\begin{aligned} &\big|\kappa(\hat{X}_i, \hat{X}_j) - \kappa(\mathbf{W}_n X_i, \mathbf{W}_n X_j)\big| \\ &\quad \le C \max\big\{\|\hat{X}_i - \mathbf{W}_n X_i\|, \|\hat{X}_j - \mathbf{W}_n X_j\|\big\} \le C\|\hat{\mathbf{X}} - \mathbf{X}\mathbf{W}_n\|_{2\to\infty}, \end{aligned}$$

for some constant C independent of i and j. Similarly

$$\begin{aligned} \big|\kappa(\hat{Y}_k, \hat{Y}_l) - \kappa(\mathbf{V}_m Y_k, \mathbf{V}_m Y_l)\big| &\le C\|\hat{\mathbf{Y}} - \mathbf{Y}\mathbf{V}_m\|_{2\to\infty}, \\ \big|\kappa(\hat{X}_i, \hat{Y}_k) - \kappa(\mathbf{W}_n X_i, \mathbf{V}_m Y_k)\big| &\le C\big(\|\hat{\mathbf{X}} - \mathbf{X}\mathbf{W}_n\|_{2\to\infty} + \|\hat{\mathbf{Y}} - \mathbf{Y}\mathbf{V}_m\|_{2\to\infty}\big). \end{aligned}$$

Thus,

$$\big|U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) - U_{n,m}(\mathbf{X}\mathbf{W}_n, \mathbf{Y}\mathbf{V}_m)\big| \le 2C\big(\|\hat{\mathbf{X}} - \mathbf{X}\mathbf{W}_n\|_{2\to\infty} + \|\hat{\mathbf{Y}} - \mathbf{Y}\mathbf{V}_m\|_{2\to\infty}\big)$$

which converges, by Lemma 2, to zero almost surely as $n, m \to \infty$. Furthermore,

$$U_{n,m}(\mathbf{X}\mathbf{W}_n, \mathbf{Y}\mathbf{V}_m) = U_{n,m}\big(\mathbf{X}, \mathbf{Y}\mathbf{V}_m\mathbf{W}_n^\top\big)$$

as κ is a radial kernel. We have that

$$n^{-1}\mathbf{X}^\top\mathbf{X} = n^{-1}\mathbf{W}_1^\top\mathbf{S}_1\mathbf{W}_1 \quad \text{and}$$
$$m^{-1}\mathbf{Y}^\top\mathbf{Y} = m^{-1}\mathbf{W}_2^\top\mathbf{S}_2\mathbf{W}_2$$

are $\sqrt{n}$-consistent and $\sqrt{m}$-consistent estimators of $\mathbb{E}[X_1X_1^\top]$ and $\mathbb{E}[Y_1Y_1^\top]$, respectively. Since F and G satisfy the distinct eigenvalues condition, we can apply the Davis–Kahan theorem to each individual eigenvectors of $\mathbb{E}[X_1X_1^\top]$ and $\mathbb{E}[Y_1Y_1^\top]$, thereby showing that $\mathbf{W}_n$ and $\mathbf{V}_m$ are $\sqrt{n}$-consistent and $\sqrt{m}$-consistent estimator of the corresponding orthogonal matrices in the eigendecomposition of $\mathbb{E}[X_1X_1^\top]$ and $\mathbb{E}[Y_1Y_1^\top]$, respectively. If $F \doteq G$, that is, $F = G \circ \mathbf{W}_0$ for $\mathbf{W}_0$ orthogonal, then $\mathbf{V}_m\mathbf{W}_n^\top = \mathbf{W}_0 + O(\max\{n^{-1/2}, m^{-1/2}\})$ and hence

$$\begin{aligned}\big|U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) - U_{n,m}(\mathbf{X}, \mathbf{Y}\mathbf{W}_0)\big| &= \big|U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) - U_{n,m}\big(\mathbf{X}, \mathbf{Y}\mathbf{V}_m\mathbf{W}_n^\top\big)\big| \\ &\quad + O\big(\max\big\{n^{-1/2}, m^{-1/2}\big\}\big)\end{aligned}$$

which also converges to zero almost surely. That is to say, the test statistic based on the estimated latent position converges to the statistic based on the true but unknown latent positions. Thus one can construct, using the test statistics $U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$, a test procedure for $\mathbb{H}_0$: $F \neq G$ that is *consistent* against all fixed alternatives $F \neq G$. This is in essence a first order result; in this regard, it is similar in spirit to first order consistency results for spectral clustering [28] and vertex classification [29]. However, as we recall from Theorem 1, in order to obtain a non-degenerate limiting distribution, we want to consider the scaled statistics $(m+n)U_{n,m}(\mathbf{X}, \mathbf{Y})$. Showing the convergence to zero of $(m+n)(U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) - U_{n,m}(\mathbf{X}, \mathbf{Y}\mathbf{V}_m\mathbf{W}_n^\top))$ is much more involved and is the main impetus behind the following lemma.

Lemma 3. *Let κ be a twice continuously differentiable kernel. Let $\mathcal{F}_\Phi = \{\Phi(Z): Z \in \Omega\}$ where $\Phi(Z) = \kappa(\cdot, Z)$ is the feature map of κ, that is, $f \in \mathcal{F}_\Phi$ if $f(X) = \kappa(X, Z)$ for some Z. Suppose $(\mathbf{X}_n, \mathbf{A}_n) \sim \mathrm{RDPG}(F)$ for $n = 1, 2, \ldots$ is a sequence of d-dimensional random dot product graphs and the latent positions distribution F satisfies the distinct eigenvalues condition in Assumption 1. Denote by $\mathbf{W}_n$ the orthogonal matrix in the eigendecomposition $\mathbf{W}_n\mathbf{S}_n\mathbf{W}_n^\top = \mathbf{X}_n^\top\mathbf{X}_n$. Then as $n \to \infty$, the sequence $\mathbf{W}_n$ satisfies*

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

almost surely, where $\hat{\mathbf{X}}_n = \{\hat{X}_i\}_{i=1}^n$ is the adjacency spectral embedding of $\mathbf{A}_n$.

Lemma 3 is the main technical result of this paper. Using the bound on $\|\hat{\mathbf{X}} - \mathbf{X}\mathbf{W}\|_{2\to\infty}$ from Lemma 2 implies that for some class of continuous functions $\mathcal{F}$, e.g., continuous functions of the form $\phi(\|\cdot - c\|)$ for all c in a compact subset of $\mathbb{R}^d$, there exists a sequence of orthogonal matrices $\mathbf{W}_n$ such that

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

almost surely as $n \to \infty$ [20], Theorem 15. Lemma 3 improves upon this; for some special class $\mathcal{F}$, the above also holds with the factor $1/n$ replaced by a factor of $1/\sqrt{n}$.

The proof of Lemma 3 is given in the Appendix. A rough sketch of the proof is as follows. For fixed $f \in \mathcal{F}_\Phi$, a Taylor expansion allows one to write $n^{-1/2}\sum_{i=1}^n (f(\mathbf{W}_n \hat{X}_i) - f(X_i))$ in terms of $\sum_i \lambda_i^{-1/2} \mathbf{v}_i^\top (\mathbf{A} - \mathbf{P})\mathbf{u}_i$ for unit vectors $\mathbf{v}_i$ depending on f and $\mathbf{u}_i$ depending on $\{X_i\}$; here λ_i are the eigenvalues of $\mathbf{P}$. Hoeffding's inequality applied to the sum $\sum_i \lambda_i^{-1/2} \mathbf{u}_i^\top (\mathbf{A} - \mathbf{P})\mathbf{v}_i$ provides an exponential tail bound for each $f \in \mathcal{F}_\Phi$. A chaining argument similar to that in [33], Section 3.2, and bounds for the so-called *covering number* of $\mathcal{F}_\Phi$ (again, see [33], Section 2.3, for a precise definition) lead to an exponential tail bound that is uniform over all $f \in \mathcal{F}_\Phi$.

The application of Lemma 3 to our nonparametric two-sample hypothesis testing problem is presented in Section 3.3. Another interesting consequence of Lemma 3 is a functional central limit theorem for $\hat{\mathbf{X}}$, which is the topic of the following subsection.

3.2. A functional central limit theorem for $\hat{\mathbf{X}}$

By replacing the class of functions $\mathcal{F}_\Phi$ in Lemma 3 with a more general class of functions $\mathcal{F}$ whose covering numbers are still "small," a similar chaining argument can be adapted to yield the following functional central limit theorem. (For a comprehensive discussion of functional central limit theorems, see, for example, [9,34] and the references therein.) We first recall certain definitions, which we reproduce from [34]. Let X_i, $1 \le i \le n$ be identically distributed random variables on a measure space $(\mathcal{X}, \mathcal{B})$, and let $\mathbb{P}_n$ be their associated *empirical measure*; that is, $\mathbb{P}_n$ is the discrete random measure defined, for any $E \in \mathcal{B}$, by

$$\mathbb{P}_n(E) = \frac{1}{n}\sum_{i=1}^n 1_E(X_i).$$

Let P denote the common distribution of the random variables X_i, and suppose that $\mathcal{F}$ is a class of measurable, real-valued functions on $\mathcal{X}$. The *$\mathcal{F}$-indexed empirical process* $\mathbb{G}_n$ is the stochastic process

$$f \mapsto \mathbb{G}_n(f) = \sqrt{n}(\mathbb{P}_n - P)f = \frac{1}{\sqrt{n}}\sum_{i=1}^n \big(f(X_i) - \mathbb{E}\big[f(X_i)\big]\big).$$

Under certain conditions, the empirical process $\{\mathbb{G}_n(f) : f \in \mathcal{F}\}$ can be viewed as a map into $\ell^\infty(\mathcal{F})$, the collection of all uniformly bounded real-valued functionals on $\mathcal{F}$. In particular, let $\mathcal{F}$ be a class of functions for which the empirical process $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$ converges to a limiting process $\mathbb{G}$ where $\mathbb{G}$ is a tight Borel-measurable element of $\ell^\infty(\mathcal{F})$ (more specifically a Brownian bridge). Then $\mathcal{F}$ is said to be a *P-Donsker class*, or for brevity, P-Donsker [34], Section 2.1. A sufficient condition, albeit a rather strong one, for $\mathcal{F}$ to be P-Donsker is via the entropy for the supremum norm. That is, let $N_\infty(\delta, \mathcal{F})$ be the smallest value of N such that there exists $\{f_j\}_{j=1}^N$ with $\sup_{f\in\mathcal{F}} \min_j \|f - f_j\|_\infty \le \delta$. Then $\mathcal{F}$ is P-Donsker for any P if [34], Section 2.5.2,

$$\int_0^\infty \sqrt{\log N_\infty(\delta, \mathcal{F})}\,\mathrm{d}\delta < \infty. \tag{3.3}$$

As an example, let $\mathcal{F}$ be the unit ball associated with a kernel κ on a compact $\Omega \subset \mathbb{R}^d$. Then $\mathcal{F}$ is P-Donsker provided κ is m-times continuously differentiable on Ω for some $m \geq 2d + 1$ [34], Theorems 2.7.1 and 2.5.6. The unit ball associated with the Gaussian kernel on $\mathbb{R}^d$ is thus P-Donsker for all d.

Theorem 4. *Let $(\mathbf{X}_n, \mathbf{A}_n)$ for $n = 1, 2, \ldots,$ be a sequence of d-dimensional* RDPG(P) *where the latent position distribution P satisfies the distinct eigenvalues condition in Assumption 1. Let $\mathcal{F}$ be a collection of (at least) twice continuously differentiable functions on Ω with*

$$\sup_{f \in \mathcal{F}, X \in \Omega} \big\| (\partial f)(X) \big\| < \infty; \qquad \sup_{f \in \mathcal{F}, X \in \Omega} \big\| \big(\partial^2 f\big)(X) \big\| < \infty.$$

Furthermore, suppose $\mathcal{F}$ satisfies equation (3.3) so that $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$ converges to $\mathbb{G}$, a P-Brownian bridge on $\ell^\infty(\mathcal{F})$. Denote by $\mathbf{W}_n$ the orthogonal matrices in the eigendecomposition $\mathbf{W}_n \mathbf{S}_n \mathbf{W}_n^\top = \mathbf{X}_n^\top \mathbf{X}_n$. Then as $n \to \infty$, the $\mathcal{F}$-indexed empirical process

$$f \in \mathcal{F} \mapsto \hat{\mathbb{G}}_n f = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big(f(\mathbf{W}_n \hat{X}_i) - \mathbb{E}\big[f(X_i) \big] \big) \tag{3.4}$$

also converges to $\mathbb{G}$ on $\ell^\infty(\mathcal{F})$.

Theorem 4 is in essence a functional central limit theorem for the estimated latent positions $\{\hat{X}_i\}$ in the random dot product graph setting. We emphasize that for any n, the $\{\hat{X}_i\}_{i=1}^n$ are not jointly independent random variables, that is, Theorem 4 is a functional central limit theorem for *dependent* data. Due to the non-identifiability of random dot product graphs, there is an explicit dependency on the sequence of orthogonal matrices $\mathbf{W}_n$; note, however, that $\mathbf{W}_n$ depends solely on $\mathbf{X}_n$ and not on the $\{\hat{X}_i\}$.

3.3. Consistent testing

We now consider testing the hypothesis $\mathbb{H}_0: F \doteq G$ using the kernel-based framework of Section 2.2. For our purpose, we shall assume henceforth that κ is a twice continuously-differentiable radial kernel and that κ is also universal. Examples of such kernels are the Gaussian kernels and the inverse multiquadric kernels $\kappa(x, y) = (c^2 + \|x - y\|^2)^{-\beta}$ for $c, \beta > 0$.

To justify this assumption on our kernel, we remark that in Theorem 5 below, we show that the test statistic $U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ based on the estimated latent positions converges to the corresponding statistic $U_{n,m}(\mathbf{X}, \mathbf{Y})$ for the true but unknown latent positions. Due to the non-identifiability of the random dot product graph under unitary transformation, *any* estimate of the latent positions is close, only up to an appropriate orthogonal transformations, to $\mathbf{X}$ and $\mathbf{Y}$. We have seen in Section 3.1 that for a radial kernel, this implies the approximations $\kappa(\hat{X}_i, \hat{X}_j) \approx \kappa(X_i, X_j)$, $\kappa(\hat{Y}_k, \hat{Y}_l) \approx \kappa(Y_k, Y_l)$ and the convergence of $U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ to $U_{n,m}(\mathbf{X}, \mathbf{Y})$. If κ is not a radial kernel, the above approximations might not hold and $U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ need not converge to $U_{n,m}(\mathbf{X}, \mathbf{Y})$. The assumption that κ is twice continuously-differentiable is for the technical conditions of

Lemma 3. Finally, the assumption that κ is universal allows the test procedure to be consistent against a large class of alternatives.

Theorem 5. *Let* $(\mathbf{X}, \mathbf{A}) \sim \mathrm{RDPG}(F)$ *and* $(\mathbf{Y}, \mathbf{B}) \sim \mathrm{RDPG}(G)$ *be independent random dot product graphs with latent position distributions* F *and* G. *Furthermore, suppose that both* F *and* G *satisfies the distinct eigenvalues condition in Assumption* 1. *Consider the hypothesis test*

$$H_0: F \doteq G \quad \textit{against} \quad H_A: F \not\doteq G.$$

Denote by $\hat{\mathbf{X}} = \{\hat{X}_1, \ldots, \hat{X}_n\}$ *and* $\hat{\mathbf{Y}} = \{\hat{Y}_1, \ldots, \hat{Y}_m\}$ *the adjacency spectral embedding of* $\mathbf{A}$ *and* $\mathbf{B}$, *respectively. Let* $\mathbf{W}_1$ *and* $\mathbf{W}_2$ *be* $d \times d$ *orthogonal matrices in the eigendecomposition* $\mathbf{W}_1\mathbf{S}_1\mathbf{W}_1^\top = \mathbf{X}^\top\mathbf{X}$, $\mathbf{W}_2\mathbf{S}_2\mathbf{W}_2 = \mathbf{Y}^\top\mathbf{Y}$, *respectively. Suppose that* $m, n \to \infty$ *and* $m/(m+n) \to \rho \in (0,1)$. *Then under the null hypothesis of* $F \doteq G$, *the sequence of matrices* $\mathbf{W}_{n,m} = \mathbf{W}_2\mathbf{W}_1^\top$ *satisfies*

$$(m+n)\big(U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) - U_{n,m}(\mathbf{X}, \mathbf{Y}\mathbf{W}_{n,m})\big) \xrightarrow{\text{a.s.}} 0. \tag{3.5}$$

Under the alternative hypothesis of $F \not\doteq G$, *the sequence of matrices* $\mathbf{W}_{n,m}$ *satisfies*

$$\frac{m+n}{\log^2(m+n)}\big(U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) - U_{n,m}(\mathbf{X}, \mathbf{Y}\mathbf{W}_{n,m})\big) \xrightarrow{\text{a.s.}} 0. \tag{3.6}$$

Proof. We first define the statistic $V_{n,m}(\mathbf{X}, \mathbf{Y})$

$$\begin{aligned} V_{n,m}(\mathbf{X}, \mathbf{Y}) &= \left\| \frac{1}{n}\sum_{i=1}^{n} \Phi(X_i) - \frac{1}{m}\sum_{k=1}^{m} \Phi(Y_k) \right\|_{\mathcal{H}}^2 \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \kappa(X_i, X_j) - \frac{2}{mn}\sum_{i=1}^{n}\sum_{k=1}^{m} \kappa(X_i, Y_k) + \frac{1}{m^2}\sum_{k=1}^{m}\sum_{l=1}^{m} \kappa(Y_k, Y_l). \end{aligned} \tag{3.7}$$

We shall prove that the difference

$$(m+n)\big(V_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) - V_{n,m}(\mathbf{X}, \mathbf{Y}\mathbf{W}_{n,m})\big) \xrightarrow{\text{a.s}} 0 \tag{3.8}$$

under the hypothesis $F \doteq G$. The claim $(m+n)(U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) - U_{n,m}(\mathbf{X}, \mathbf{Y}\mathbf{W}_{n,m})) \xrightarrow{\text{a.s.}} 0$ in Theorem 5 follows from equation (3.8) and the following expression

$$\begin{aligned} &(m+n)\big(V_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) - V_{n,m}(\mathbf{X}, \mathbf{Y}\mathbf{W}_{n,m})\big) \\ &\quad = (m+n)\big(U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) - U_{n,m}(\mathbf{X}, \mathbf{Y}\mathbf{W}_{n,m})\big) + r_1 + r_2, \end{aligned}$$

where r_1 and r_2 are defined as (recall that κ is a radial kernel)

$$r_1 = \frac{m+n}{n(n-1)}\sum_{i=1}^{n}\big(\kappa(X_i, X_i) - \kappa(\hat{X}_i, \hat{X}_i)\big) + \frac{m+n}{m(m-1)}\sum_{k=1}^{m}\big(\kappa(Y_k, Y_k) - \kappa(\hat{Y}_k, \hat{Y}_k)\big),$$

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

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 a constant L such that both $|r_1|$ and $|r_2|$ is bounded from above by

$$L(m+n)\left\{ \frac{\|\hat{\mathbf{X}} - \mathbf{X}\mathbf{W}_1\|_{2\to\infty}}{n-1} + \frac{\|\hat{\mathbf{Y}} - \mathbf{Y}\mathbf{W}_2\|_{2\to\infty}}{m-1} \right\}.$$

We thus proceed to establishing equation (3.8). Define $\xi_W, \hat{\xi} \in \mathcal{H}$ by

$$\xi_W = \frac{\sqrt{m+n}}{n} \sum_{i=1}^{n} \kappa(\mathbf{W}_1 X_i, \cdot) - \frac{\sqrt{m+n}}{m} \sum_{k=1}^{m} \kappa(\mathbf{W}_2 Y_k, \cdot);$$

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

Note that

$$\begin{aligned} \big|(m+n)\big(V_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) - V_{n,m}(\mathbf{X}, \mathbf{Y}\mathbf{W}_{n,m})\big)\big| &= \big|\|\xi_W\|_{\mathcal{H}}^2 - \|\hat{\xi}\|_{\mathcal{H}}^2\big| \\ &\le \|\xi_W - \hat{\xi}\|_{\mathcal{H}} \big(2\|\xi_W\|_{\mathcal{H}} + |\xi_W - \hat{\xi}\|_{\mathcal{H}}\big). \end{aligned}$$

We now bound the terms $\|\xi_W - \hat{\xi}\|_{\mathcal{H}}$ and $\|\xi_W\|_{\mathcal{H}}$. We first bound $\|\xi_W\|_{\mathcal{H}}$. Let $\mathbf{T}_1$ and $\mathbf{T}_2$ be the orthogonal matrices in the eigendecomposition of $\mathbb{E}[X_1 X_1^\top]$ and $\mathbb{E}[Y_1 Y_1^\top]$. The distinct eigenvalues condition in Assumption 1 implies, by the Davis–Kahan theorem, that $\mathbf{W}_1 = \mathbf{T}_1 + O(n^{-1/2})$ and $\mathbf{W}_2 = \mathbf{T}_2 + O(m^{-1/2})$. When $F \doteq G$, $F \circ \mathbf{T}_1 = G \circ \mathbf{T}_2$ and hence by adding and subtracting terms, we have

$$\xi_W = \sqrt{\frac{m+n}{n}} \sum_{i=1}^{n} \frac{\kappa(\mathbf{T}_1 X_i, \cdot) - \mu[F \circ \mathbf{T}_1]}{\sqrt{n}} - \sqrt{\frac{m+n}{m}} \sum_{k=1}^{m} \frac{\kappa(\mathbf{T}_2 Y_k, \cdot) - \mu[G \circ \mathbf{T}_2]}{\sqrt{m}} + O(1).$$

That is, $\xi_W - O(1)$ is a sum of independent mean zero random elements of $\mathcal{H}$. In addition $\|\kappa(Z, \cdot) - \mu[F]\|_{\mathcal{H}} \le 2$ for any $Z \in \mathbb{R}^d$. Using a Hilbert space concentration inequality [23], Theorem 3.5, we obtain that

$$\mathbb{P}\big[\|\xi_W\|_{\mathcal{H}} \ge \sqrt{m+n}(s/\sqrt{n} + t/\sqrt{m})\big] \le 2\big(\exp\big(-(1+m/n)s^2/8\big) + \exp\big(-(1+n/m)t^2/8\big)\big),$$

which implies that $\|\xi_W\|_{\mathcal{H}}$ is bounded in probability. We now bound $\|\xi_W - \hat{\xi}\|_{\mathcal{H}}$. We have

$$\xi_W - \hat{\xi} = \sqrt{\frac{m+n}{n}} \sum_{i=1}^{n} \frac{\kappa(\mathbf{W}_1 X_i, \cdot) - \kappa(\hat{X}_i, \cdot)}{\sqrt{n}} - \sqrt{\frac{m+n}{n}} \sum_{k=1}^{m} \frac{\kappa(\mathbf{W}_2 Y_k, \cdot) - \kappa(\hat{Y}_k, \cdot)}{\sqrt{m}}$$

and Lemma 3 implies (as κ is radial)

$$\sqrt{\frac{m+n}{n}}\sum_{i=1}^{n}\frac{\kappa(\mathbf{W}_1X_i,\cdot)-\kappa(\hat{X}_i,\cdot)}{\sqrt{n}}\xrightarrow{\text{a.s.}}0;\qquad\sqrt{\frac{m+n}{n}}\sum_{k=1}^{m}\frac{\kappa(\mathbf{W}_2Y_k,\cdot)-\kappa(\hat{Y}_k,\cdot)}{\sqrt{m}}\xrightarrow{\text{a.s.}}0$$

as $m,n\to\infty$, $m/n\to\rho\in(0,1)$. Thus $\|\xi_W-\hat{\xi}\|_{\mathcal{H}}\to 0$ and equations (3.8) and (3.5) are established.

We now derive equation (3.6). We note that in the case when $F\neq G$, one still has

$$\big|(m+n)\big(V_{n,m}(\hat{\mathbf{X}},\hat{\mathbf{Y}})-V_{n,m}(\mathbf{X},\mathbf{YW}_{n,m})\big)\big|\leq\|\xi_W-\hat{\xi}\|_{\mathcal{H}}\big(2\|\xi_W\|_{\mathcal{H}}+\|\xi_W-\hat{\xi}\|_{\mathcal{H}}\big),$$

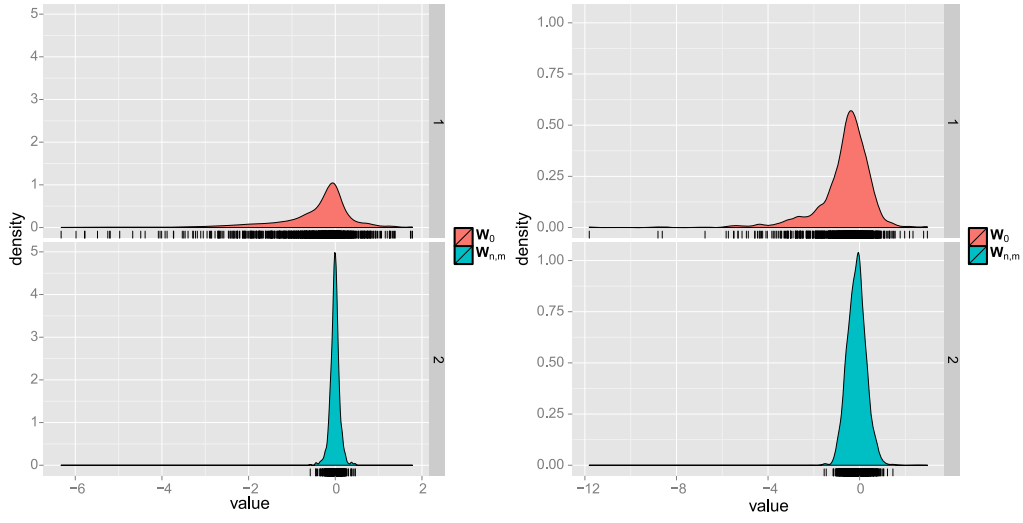
where $\hat{\xi}$ and ξ_W are defined identically to the case when $F\doteq G$. However, when $F\neq G$, the bound $\|\xi_W\|_{\mathcal{H}}=O(1)$ with high probability no longer holds. Indeed, when $F\neq G$,

$$\xi_W-O(1)=\frac{\sqrt{m+n}}{n}\sum_{i=1}^{n}\kappa(\mathbf{T}_1X_i,\cdot)-\frac{\sqrt{m+n}}{m}\sum_{k=1}^{m}\kappa(\mathbf{T}_2Y_k,\cdot)$$

is not a sum of mean 0 random variables. We thus bound $\|\xi_W\|_{\mathcal{H}}=O(\sqrt{n\log n})$ with high probability. The proof of Lemma 3 yields $\|\hat{\xi}-\xi_W\|_{\mathcal{H}}=O(n^{-1/2}\log n)$ with high probability (see equation (A.7) in the Appendix). Hence, $|(m+n)(V_{n,m}(\hat{\mathbf{X}},\hat{\mathbf{Y}})-V_{n,m}(\mathbf{X},\mathbf{YW}_{n,m}))|$ is of order $\log^{3/2}n$ with high probability and equation (3.6) follows. □

Equations (3.5) and (3.6) 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 $U_{n,m}(\mathbf{X},\mathbf{YW}_{n,m})$ defined in equation (2.2) using the true latent positions, under both the null and alternative hypothesis. Because κ is a universal kernel, $U_{n,m}(\mathbf{X},\mathbf{YW}_{n,m})$ 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.

We note that a subtle point in the statement and argument of the theorem is that $\mathbf{W}_{n,m}$ is a random quantity depending on $\mathbf{X}_n$ and $\mathbf{Y}_m$. There does exist a deterministic matrix $\mathbf{W}_0$ depending only on F and G such that $\mathbf{W}_{n,m}\to\mathbf{W}_0$ almost surely as $m,n\to\infty$. Indeed, from the proof of the theorem, we have that $\mathbf{W}_1$ is a $\sqrt{n}$-consistent estimator of $\mathbf{T}_1$ where $\mathbf{T}_1$ is the orthogonal matrix in the eigendecomposition of $\mathbb{E}[X_1X_1^\top]$ and that $\mathbf{W}_2$ is a $\sqrt{m}$-consistent estimator of $\mathbf{T}_2$ where $\mathbf{T}_2$ is the orthogonal matrix in the eigendecomposition of $\mathbb{E}[Y_1Y_1^\top]$. Under the null hypothesis, $F\circ\mathbf{T}_1=G\circ\mathbf{T}_2$; hence if we define $\mathbf{W}_0$ as $\mathbf{T}_2\mathbf{T}_1^\top$, then $\mathbf{W}_2\mathbf{W}_1^\top$ is a $\sqrt{n}$-consistent estimator of $\mathbf{W}_0$. This convergence of order $O(n^{-1/2})$ is, however, not sufficiently fast to guarantee that $(m+n)(U_{n,m}(\hat{\mathbf{X}},\hat{\mathbf{Y}})-U_{n,m}(\mathbf{X},\mathbf{YW}_0))$ converges to zero almost surely when $F\doteq G$. For example, let F be a mixture of two multivariate logit-normal distributions with mean parameters $(0,0)$, $(4,4)$, identity covariance matrices and mixture components $(0.4,0.6)$; let G be a multivariate logit-normal distribution with mean parameter $(2,2)$ and identity covari-



(a) $(\mathbf{X}, \mathbf{A}) \sim \mathrm{RDPG}(F)$, $(\mathbf{Y}, \mathbf{B}) \sim \mathrm{RDPG}(F)$ (b) $(\mathbf{X}, \mathbf{A}) \sim \mathrm{RDPG}(F)$, $(\mathbf{Y}, \mathbf{B}) \sim \mathrm{RDPG}(G)$

Figure 1. Comparison between the random $\mathbf{W}_{n,m}$ and fixed but unknown $\mathbf{W}_0$. The empirical distributions of $(m+n)(U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) - U_{n,m}(\mathbf{X}, \mathbf{Y}\mathbf{W}_0))$ (in red) and $(m+n)(U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) - U_{n,m}(\mathbf{X}, \mathbf{Y}\mathbf{W}_{n,m})$ (in blue) under (a) the null setting of $(\mathbf{X}, \mathbf{A}) \sim F, (\mathbf{Y}, \mathbf{B}) \sim F$ and (b) the alternative setting of $(\mathbf{X}, \mathbf{A}) \sim F, (\mathbf{Y}, \mathbf{B}) \sim G$.

ance matrix. Figure 1 illustrates that the difference $(m+n)(U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) - U_{n,m}(\mathbf{X}, \mathbf{Y}\mathbf{W}_{n,m}))$ is in general smaller compared to the difference $(m+n)(U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) - U_{n,m}(\mathbf{X}, \mathbf{Y}\mathbf{W}_0))$, thereby complicating the derivation of the exact nondegenerate limiting distribution for $(m+n)U_{m,n}(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$. Nevertheless, since the nondegenerate limiting distribution for $(m+n)U_{m,n}(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ will not be distribution-free, the fact that it is currently unknown is, for all practical purposes, irrelevant. Indeed, the proposed test statistic still yields a consistent test procedure whose critical values can be obtained through a simple bootstrapping procedure.

Remark. The computational cost for implementing the test procedure in Theorem 5 consist mainly of two parts, namely computing the adjacency spectral embedding of the graphs $\mathbf{A}$ and $\mathbf{B}$, and computing the test statistic $U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$. Assuming $n \geq m$, the adjacency spectral embedding of $\mathbf{A}$ and $\mathbf{B}$ into $\mathbb{R}^d$ is a (partial) singular value decomposition of $\mathbf{A}$ and $\mathbf{B}$ and thus can be computed in $O(n^2 d)$ time. The test statistic $U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ can be evaluated in $O(n^2)$ time.

Remark. The proof of Theorem 5 can be adapted to show that data-adaptive bandwidth selections behave similarly for $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$ as for $\mathbf{X}$ and $\mathbf{Y}$. That is to say, we can show that under the null hypothesis, $\Delta_\theta = (m+n)(U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) - U_{n,m}(\mathbf{X}, \mathbf{Y}\mathbf{W}_{n,m}))$ 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 $\Theta \subset \mathbb{R}_+$.

4. Experimental results

In this section, we illustrate our test statistic and procedure with two examples. The first example investigates the comparison of distinct two-block stochastic blockmodels. The second example considers graphs from a protein network dataset and uses our proposed test statistic to build a classifier.

4.1. Stochastic blockmodel example

We illustrate the hypothesis tests through several simulated and real data examples. For our first example, let F_ε for a given $\varepsilon > 0$ be mixture of point masses corresponding to a two-block stochastic block model with block membership probabilities $(0.4, 0.6)$ and block probabilities $\mathbf{B}_\varepsilon = \left[\begin{smallmatrix} 0.5+\varepsilon & 0.2 \\ 0.2 & 0.5+\varepsilon \end{smallmatrix}\right]$. We then test, for a given $\varepsilon > 0$, the hypothesis $H_0: F_0 \doteq F_\varepsilon$ against the alternative $H_A: F_0 \neq F_\varepsilon$ using the kernel-based testing procedure of Section 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 RDPG(F_0) and one graph on n vertices from RPDG(F_ε). We then perform an adjacency spectral embedding on the graphs, in which we embed the graphs into $\mathbb{R}^2$, and we proceed to compute the kernel-based test statistic. We 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 power of the test statistic for various choices of $n \in \{100, 200, 500, 1000\}$ and $\varepsilon \in \{0.02, 0.05, 0.1\}$ 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.

4.2. Classification of protein networks

For our last example, we show how the statistics $U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ can also be adapted for use in graphs classification. More concretely, we consider the problem of classifying proteins network into en-

Table 1. Power estimates for testing the null hypothesis $F \doteq G$ 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

	$\varepsilon = 0.02$		$\varepsilon = 0.05$		$\varepsilon = 0.1$	
n	$\{\mathbf{X}, \mathbf{Y}\}$	$\{\hat{\mathbf{X}}, \hat{\mathbf{Y}}\}$	$\{\mathbf{X}, \mathbf{Y}\}$	$\{\hat{\mathbf{X}}, \hat{\mathbf{Y}}\}$	$\{\mathbf{X}, \mathbf{Y}\}$	$\{\hat{\mathbf{X}}, \hat{\mathbf{Y}}\}$
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 2. Classification accuracy on the enzyme dataset

Classifier	Accuracy (%)
SVM with optimized feature vector kernel [8]	80.17
SVM with random walk kernel with secondary structure [5]	77.30
k-NN with dissimilarities based on $U_{n,m}$	78.20

zyme versus non-enzymes. We use the dataset of [8], which consists of 1178 protein networks labeled as enzymes (691 networks) and non-enzymes (487 networks). For our classification procedure, we first embed each of the protein networks into $\mathbb{R}^5$ using adjacency spectral embedding. The choice of $d = 5$ is chosen from among the choices of embedding dimensions ranging from $d = 2$ through $d = 15$ to minimize the classification error rate. 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 2. For the purpose of comparison, we also include the accuracy of several other classifiers that were previously applied on this data set, see [5,8]. The results of [8] 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 [5] 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. The accuracy of our straightforward classifier, which does not use any information about associated secondary structure, is comparable to that obtained from using SVM with a well-designed features kernel or well-designed graph kernels.

5. Extensions

In this section, we will consider extensions to alternative hypothesis tests that consider looser notions of equality between the two distributions. These notions may be quite useful in practice due to variations in graph properties that one may want to ignore in a comparison of the graphs. We do not formally state results for these extensions but we note that they can be derived in a similar manner to Theorem 5; see Sections A.1 and A.2 in the Appendix.

5.1. Scaling case

We now consider the case of testing the hypothesis that the distributions F and G are equal up to scaling. In particular, the test

$$H_0: F \doteq G \circ c \qquad \text{for some } c > 0 \quad \text{against} \quad H_A: F \not\doteq G \circ c \qquad \text{for any } c > 0,$$

Table 3. Power estimates for testing the null hypothesis $F \doteq G \circ c$ 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 $\varepsilon = 0$ coincides with bootstrap estimate for the size of the test

	$\varepsilon = 0$		$\varepsilon = 0.05$		$\varepsilon = 0.1$		$\varepsilon = 0.2$	
n	$\{\mathbf{X}, \mathbf{Y}\}$	$\{\hat{\mathbf{X}}, \hat{\mathbf{Y}}\}$	$\{\mathbf{X}, \mathbf{Y}\}$	$\{\hat{\mathbf{X}}, \hat{\mathbf{Y}}\}$	$\{\mathbf{X}, \mathbf{Y}\}$	$\{\hat{\mathbf{X}}, \hat{\mathbf{Y}}\}$	$\{\mathbf{X}, \mathbf{Y}\}$	$\{\hat{\mathbf{X}}, \hat{\mathbf{Y}}\}$
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

where $Y \sim F \circ c$ if $cY \sim F$. The test statistic is now a simple modification of the one in Theorem 5, that is, we first scale the adjacency spectral embeddings by the norm of the empirical means before computing the kernel test statistic. In particular, if we let

$$\hat{s}_X = n^{-1/2}\|\hat{\mathbf{X}}\|_F, \qquad \hat{s}_Y = m^{-1/2}\|\hat{\mathbf{Y}}\|_F, \qquad s_X = n^{-1/2}\|\mathbf{X}\|_F, \qquad s_Y = m^{-1/2}\|\mathbf{Y}\|_F,$$

then the conclusions of Theorem 5 hold where we use $U_{n,m}(\hat{\mathbf{X}}/\hat{s}_X, \hat{\mathbf{Y}}/\hat{s}_Y)$ as the test statistic in comparison to $U_{n,m}(\mathbf{X}/s_X, \mathbf{Y}\mathbf{W}_{n,m}/s_Y)$. Note that we must restrict c so that $G \circ c$ is still a valid distribution for an RDPG.

As an example let F_ε be the uniform distribution on $[\varepsilon, 1/\sqrt{2}]^2$ where $\varepsilon \geq 0$ and let G be the uniform distribution on $[0, 1/\sqrt{3}]^2$. For a given ε, we test the hypothesis $H_0: F_\varepsilon \doteq G \circ c$ for some constant $c > 0$ against the alternative $H_A: F_\varepsilon \not\doteq G \circ c$ 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 $\sigma = 0.5$. Table 3 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.

5.2. Projection case

We next consider the case of testing

$$H_0: F \circ \pi^{-1} \doteq G \circ \pi^{-1} \quad \text{against} \quad H_A: F \circ \pi^{-1} \not\doteq G \circ \pi^{-1},$$

where π is the projection $x \mapsto x/\|x\|$ that maps x onto the unit sphere in $\mathbb{R}^d$. In an abuse of notation, we will also write $\pi(\mathbf{X})$ to denote the row-wise projection of the rows of $\mathbf{X}$ onto the unit sphere.

We shall assume that 0 is not an atom of either F or G, that is, $F(0) = G(0) = 0$, 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, that is, there exists some $\varepsilon > 0$ such that $F(\{x: \|x\| \leq \varepsilon\}) = G(\{x: \|x\| \leq \varepsilon\}) = 0$. A truncation argument with $\varepsilon \to 0$ allows us to handle the general case of distributions on Ω where 0 is not an atom.

To contextualize the test of equality up to projection, consider the very specific case of the degree-corrected stochastic blockmodel [16]. A degree-corrected stochastic blockmodel can be view as a random dot product graph whose latent position X_v for an arbitrary vertex v is of the form $X_v = \theta_v \nu_v$ where ν_v is sampled from a mixture of point masses and θ_v (the degree-correction factor) is sampled from a distribution on $(0, 1]$. Thus, given two degree-corrected stochastic blockmodel graphs, equality up to projection tests whether the underlying mixture of point masses (that is, the distribution of the ν_v) are the same modulo the distribution of the degree-correction factors θ_v.

For this test, under the assumption that both F and G have supports bounded away from the origin, the conclusions of Theorem 5 hold where we use $U_{n,m}(\pi(\hat{\mathbf{X}}), \pi(\hat{\mathbf{Y}}))$ as the test statistic and compare it to $U_{n,m}(\pi(\mathbf{X}), \pi(\mathbf{Y})\mathbf{W}_{n,m})$.

5.3. Local alternatives and sparsity

We now consider the test procedure of Theorem 5 in the context of (1) local alternatives and (2) sparsity. It is not hard to show that the test statistic $U_{n,m}(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ is also consistent against local alternatives, in particular the setting $(\mathbf{X}_n, \mathbf{A}_n) \sim \text{RDPG}(F_n)$, $(\mathbf{Y}_n, \mathbf{B}_n) \sim \text{RDPG}(G_n)$ with $\|\mu[F_n] - \mu[G_n]\|_{\mathcal{H}} \to 0$. In this setting, the accuracy of $\hat{\mathbf{X}}_n$ and $\hat{\mathbf{Y}}_n$ as estimates for $\mathbf{X}_n$ and $\mathbf{Y}_n$ is unchanged; the only difference is that the distance between F_n and G_n is shrinking. Thus equations (3.5) and (3.6) continue to hold and the test procedure is consistent against all local alternatives for which $\|\mu[F] - \mu[G]\|_{\mathcal{H}} = \omega(n^{-1/2}\log^K(n))$ for some integer $K \geq 2$ (cf. [11], Theorem 13).

Another related setting is that of sparsity, in which $(\mathbf{X}_n, \mathbf{A}_n) \sim \text{RDPG}(\alpha_n^{1/2} F)$, $(\mathbf{Y}_n, \mathbf{B}_n) \sim \text{RDPG}(\alpha_n^{1/2} G)$, with F and G being fixed distributions but the sparsity factor $\alpha_n \to 0$. That is to say, $(\mathbf{X}_n, \mathbf{A}_n) \sim \text{RDPG}(\alpha_n^{1/2} F)$ for $\alpha_n \leq 1$ if the rows of $\mathbf{X}_n$ are sampled i.i.d. from F and, conditioned on $\mathbf{X}_n$, $\mathbf{A}_n$ is a random $n \times n$ adjacency matrix with probability

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

Now the accuracy of $\hat{\mathbf{X}}_n$ and $\hat{\mathbf{Y}}_n$ as estimates for $\mathbf{X}_n$ and $\mathbf{Y}_n$ decreases with α_n due to increasing sparsity. More specifically, if $\hat{\mathbf{X}}_n$ denotes the adjacency spectral embedding of $\mathbf{A}_n$ where $(\mathbf{X}_n, \mathbf{A}_n) \sim \text{RDPG}(\alpha_n F)$, then Lemma 2 can be extended to yield that, with probability at least $1 - 4\eta$, there exists an orthogonal matrix $\mathbf{W}_n$ such that

$$\big\|\alpha_n^{-1/2}\hat{\mathbf{X}}_n - \mathbf{X}_n\mathbf{W}_n\big\|_F \leq \alpha_n^{-1/2} C_1 \tag{5.1}$$

for some constant C_1. We note that there are n rows in $\hat{\mathbf{X}}_n$ and hence, on average, we have that for each index i, $\|\alpha_n^{-1/2}\hat{X}_i - \mathbf{W}_n X_i\| \leq (n\alpha_n)^{-1/2} C_1$ with high probability. Thus, if $n\alpha_n \to \infty$, we have that, on average, each $\alpha_n^{-1/2}\hat{X}_i$ is a consistent estimate of the corresponding X_i. Thus we should expect that there exists some sequence of orthogonal matrices $\mathbf{V}_n$ such that $|U_{n,m}(\alpha_n^{-1/2}\hat{\mathbf{X}}_n, \alpha_n^{-1/2}\hat{\mathbf{Y}}_n) - U_{n,m}(\mathbf{X}_n, \mathbf{Y}_n\mathbf{V}_n)| \to 0$ as $n \to \infty$. More formally, we have the following.

Proposition 6. *Let* $(\mathbf{X}_n, \mathbf{A}_n) \sim \mathrm{RDPG}(\alpha_n^{1/2} F)$ *and* $(\mathbf{Y}_m, \mathbf{B}_m) \sim \mathrm{RDPG}(\beta_m^{1/2} G)$ *be independent random dot product graphs with latent position distributions* F *and* G *and sparsity factor* α_n *and* β_m*, respectively. Furthermore, suppose that both* F *and* G *satisfies the distinct eigenvalues condition in Assumption 1 and that* α_n *and* β_m *are known. Consider the hypothesis test*

$$H_0: F \doteq G \quad \textit{against} \quad H_A: F \not\doteq G.$$

Denote by $\hat{\mathbf{X}}_n = \{\hat{X}_1, \ldots, \hat{X}_n\}$ *and* $\hat{\mathbf{Y}}_m = \{\hat{Y}_1, \ldots, \hat{Y}_m\}$ *the adjacency spectral embedding of* $\mathbf{A}_n$ *and* $\mathbf{B}_m$*, respectively. Let* $\mathbf{W}_1$ *and* $\mathbf{W}_2$ *be* $d \times d$ *orthogonal matrices in the eigendecomposition* $\mathbf{W}_1 \mathbf{S}_1 \mathbf{W}_1^\top = \mathbf{X}_n^\top \mathbf{X}_n$, $\mathbf{W}_2 \mathbf{S}_2 \mathbf{W}_2 = \mathbf{Y}_m^\top \mathbf{Y}_m$*, respectively. Suppose that* $m, n \to \infty$, $\frac{m}{m+n} \to \rho \in (0,1)$ *and furthermore that* $n\alpha_n = \omega(\log^4 n)$ *and* $m\beta_m = \omega(\log^4 m)$*. Then the sequence of matrices* $\mathbf{W}_{n,m} = \mathbf{W}_2 \mathbf{W}_1^\top$ *satisfies*

$$U_{n,m}\big(\alpha_n^{-1/2}\hat{\mathbf{X}}_n, \beta_m^{-1/2}\hat{\mathbf{Y}}_m\big) - U_{n,m}(\mathbf{X}_n, \mathbf{Y}_m \mathbf{W}_{n,m}) \xrightarrow{\text{a.s.}} 0. \tag{5.2}$$

Proof Sketch. Let $\psi = U_{n,m}(\alpha_n^{-1/2}\hat{\mathbf{X}}_n, \beta_m^{-1/2}\hat{\mathbf{Y}}_m) - U_{n,m}(\mathbf{X}_n, \mathbf{Y}_m \mathbf{W}_{n,m})$. We have

$$\begin{aligned}
\psi = {} & \frac{1}{n(n-1)} \sum_{j \neq i} \big(\kappa\big(\alpha_n^{-1/2}\hat{X}_i, \alpha_n^{-1/2}\hat{X}_j\big) - \kappa(\mathbf{W}_1 X_i, \mathbf{W}_1 X_j)\big) \\
& - \frac{2}{mn} \sum_{i=1}^{n} \sum_{k=1}^{m} \big(\kappa\big(\alpha_n^{-1/2}\hat{X}_i, \beta_m^{-1/2}\hat{Y}_k\big) - \kappa(\mathbf{W}_1 X_i, \mathbf{W}_2 Y_k)\big) \\
& + \frac{1}{m(m-1)} \sum_{l \neq k} \big(\kappa\big(\beta_m^{-1/2}\hat{Y}_k, \beta_m^{-1/2}\hat{Y}_l\big) - \kappa(\mathbf{W}_2 Y_k, \mathbf{W}_2 Y_l)\big).
\end{aligned}$$

Let $\mathcal{S}_X \subset \{1, 2, \ldots, n\}$ and $\mathcal{S}_Y \subset \{1, 2, \ldots, m\}$ be defined by

$$\begin{aligned}
\mathcal{S}_X &= \big\{i: \big\|\alpha_n^{-1/2}\hat{X}_i - \mathbf{W}_1 X_i\big\| \leq C_1 (n\alpha_n)^{-1/2} \log n\big\}, \\
\mathcal{S}_Y &= \big\{k: \big\|\beta_m^{-1/2}\hat{Y}_k - \mathbf{W}_2 Y_k\big\| \leq C_2 (m\beta_m)^{-1/2} \log m\big\}.
\end{aligned}$$

From equation (5.1), the number of indices $i \notin S_X$ is of order $o(n)$, with high probability. Similarly, the number of indices $k \notin S_Y$ is of order $o(m)$ with high probability. Therefore,

$$\begin{aligned}
\psi = {} & \frac{1}{n(n-1)} \sum_{i \in S_X} \sum_{j \in S_X} \big(\kappa\big(\alpha_n^{-1/2}\hat{X}_i, \alpha_n^{-1/2}\hat{X}_j\big) - \kappa(\mathbf{W}_1 X_i, \mathbf{W}_1 X_j)\big) \\
& - \frac{2}{mn} \sum_{i \in S_X} \sum_{k \in S_Y} \big(\kappa\big(\alpha_n^{-1/2}\hat{X}_i, \beta_m^{-1/2}\hat{Y}_k\big) - \kappa(\mathbf{W}_1 X_i, \mathbf{W}_2 Y_k)\big) \\
& + \frac{1}{m(m-1)} \sum_{k \in S_Y} \sum_{l \in S_Y} \big(\kappa\big(\beta_m^{-1/2}\hat{Y}_k, \beta_m^{-1/2}\hat{Y}_l\big) - \kappa(\mathbf{W}_2 Y_k, \mathbf{W}_2 Y_l)\big) + o(1).
\end{aligned}$$

We consider the term $1/(n(n-1))\sum_{i\in S_X}\sum_{j\in S_X}(\kappa(\alpha_n^{-1/2}\hat{X}_i,\alpha_n^{-1/2}\hat{X}_j)-\kappa(\mathbf{W}_1X_i,\mathbf{W}_1X_j))$. By the differentiability of κ and compactness of Ω, we have

$$\begin{aligned}&\left|\kappa\left(\alpha_n^{-1/2}\hat{X}_i,\alpha_n^{-1/2}\hat{X}_j\right)-\kappa(\mathbf{W}_1X_i,\mathbf{W}_1X_j)\right|\\&\quad\le C\max\left\{\left\|\alpha_n^{-1/2}\hat{X}_i-\mathbf{W}_1X_i\right\|,\left\|\alpha_n^{-1/2}\hat{X}_j-\mathbf{W}_1X_j\right\|\right\}\end{aligned}$$

for some constant C independent of i and j. Thus,

$$\begin{aligned}&\left|\frac{1}{n(n-1)}\sum_{i\in S_X}\sum_{j\in S_X}\left(\kappa\left(\alpha_n^{-1/2}\hat{X}_i,\alpha_n^{-1/2}\hat{X}_j\right)-\kappa(\mathbf{W}_1X_i,\mathbf{W}_nX_j)\right)\right|\\&\quad\le\max_{i\in S_X}C\left\|\alpha_n^{-1/2}\hat{X}_i-\mathbf{W}_1X_i\right\|.\end{aligned}$$

Similar reasoning yields

$$\begin{aligned}|\psi|&\le 2C\left(\max_{i\in S_X}\left\|\alpha_n^{-1/2}\hat{X}_i-\mathbf{W}_1X_i\right\|+\max_{k\in S_Y}\left\|\beta_m^{-1/2}\hat{Y}_k-\mathbf{W}_2Y_k\right\|\right)+o(1)\\&\le 2C\left(C_1(n\alpha_n)^{-1/2}\log n+C_2(m\beta_m)^{-1/2}\log m\right)+o(1)\end{aligned}$$

with high probability. As $n\alpha_n=\omega(\log^4 n)$ and $m\beta_m=\omega(\log^4 m)$, we have $\psi\xrightarrow{\text{a.s.}}0$ as $m,n\to\infty$. □

We assume in the statement of Proposition 6 that the sparsity factors α_n and β_m are known. If α_n and β_m are unknown, then they can be estimated from the adjacency spectral embedding of $\mathbf{A}_n$ and $\mathbf{B}_m$, but only up to some constant factor. Hence the hypothesis test of

$$H_0\colon F\doteq G\quad\text{against}\quad H_A\colon F\not\doteq G$$

is no longer meaningful (as the sparsity factors α_n and β_m cannot be determined uniquely) and one should consider instead the hypothesis test of equality up to scaling of Section 5.1.

We note that the conclusion of Proposition 6, namely equation (5.2), is weaker than that of Theorem 5 due to the lack of the $m+n$ factor in equation (5.2) as compared to equations (3.5) and (3.6). The more difficult question, and one which we will not address in this paper, is to refine the rate of convergence to zero of equation (5.2) in the sparse setting. We suspect, however, that equation (3.5) will not hold in the case where $\alpha_n=o(n^{-1/2})$ and $\beta_m=o(m^{-1/2})$. Nevertheless, Proposition 6 still yields a test procedure that is consistent against any alternative, provided that both F and G satisfy Assumption 1, and that the sparsity factors α_n and β_m do not converge to 0 too quickly.

6. 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 exist 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, for example, [35,36]. There is thus potential theoretical value in the formulation of 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 [36] for a brief discussion, but much remains to be investigated. For example, the limiting distribution of the test statistic in [36] 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 of [30], 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 $(\mathbf{X}, \mathbf{A}) \sim \mathrm{RDPG}(F_X)$ and $(\mathbf{Y}, \mathbf{B}) \sim \mathrm{RDPG}(F_Y)$ 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.

Appendix: Additional proofs

Proof of Lemma 3. As κ is twice continuously differentiable, $\mathcal{F}$ is also twice continuously differentiable [27], Corollary 4.36. Denote by $\mathbf{W}$ the orthogonal matrix such that $\mathbf{X} = \mathbf{U_P S_P}^{1/2}\mathbf{W}$.

Let $f \in \mathcal{F}_\Phi$, a Taylor expansion of f then yields

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

where, for any i, $X_i^* \in \mathbb{R}^d$ is such that $\|X_i^* - X_i\| \le \|\mathbf{W}\hat{X}_i - X_i\|$. We first bound the quadratic terms, i.e. those depending on $\partial^2 f$. We note that since κ is twice continuously differentiable, $\sup_{f\in\mathcal{F}_\Phi, X\in\Omega} \|(\partial^2 f)(X)\|$ is bounded (the norm under consideration is the spectral norm on matrices). Therefore,

$$\sup_{f\in\mathcal{F}_\Phi}\left|\sum_{i=1}^n \frac{(\mathbf{W}\hat{X}_i - X_i)^\top(\partial^2 f)(X_i^*)(\mathbf{W}\hat{X}_i - X_i)}{\sqrt{n}}\right| \le \sup_{f\in\mathcal{F},X\in\Omega} \frac{\|(\partial^2 f)(X)\|\|\hat{\mathbf{X}}\mathbf{W} - \mathbf{X}\|_F^2}{\sqrt{n}}.$$

Hence, by applying Lemma 2 to bound $\|\hat{\mathbf{X}}\mathbf{W} - \mathbf{X}\|_F^2$ in the above expression, we have

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

as $n\to\infty$.

We now bound the linear terms, i.e., those depending on ∂f. For any $f\in\mathcal{F}_\Phi$, and any $X_1,\dots,X_n$, let $\mathbf{M}(\partial f) = \mathbf{M}(\partial f; X_1,\dots,X_n) \in \mathbb{R}^{n\times d}$ be the matrix whose rows are the vectors $(\partial f)(X_i)$. We then have

$$\begin{aligned}\zeta(f) &:= \frac{1}{\sqrt{n}}\sum_{i=1}^n (\partial f)(X_i)^\top(\mathbf{W}\hat{X}_i - X_i)\\ &= \frac{1}{\sqrt{n}}\operatorname{tr}\big((\hat{\mathbf{X}}\mathbf{W} - \mathbf{X})\big[\mathbf{M}(\partial f)\big]^\top\big) = \frac{1}{\sqrt{n}}\operatorname{tr}\big(\big(\mathbf{U}_\mathbf{A}\mathbf{S}_\mathbf{A}^{1/2} - \mathbf{U}_\mathbf{P}\mathbf{S}_\mathbf{P}^{1/2}\big)\mathbf{W}\big[\mathbf{M}(\partial f)\big]^\top\big).\end{aligned}$$

Now $\mathbf{A} = \mathbf{U}_\mathbf{A}\mathbf{S}_\mathbf{A}\mathbf{U}_\mathbf{A}^\top + \mathbf{E}$ where, as we recall in Definition 2, $\mathbf{S}_\mathbf{A}$ is the diagonal matrix containing the d largest eigenvalues of $|\mathbf{A}|$ (which coincides, with high probability, with the eigenvalues of A) and $\mathbf{U}_\mathbf{A}$ is the matrix whose columns are the corresponding eigenvectors. The eigendecomposition of $\mathbf{E}$ can be written in terms of the eigenvalues and eigenvectors that are not included in $\mathbf{S}_\mathbf{A}$ and $\mathbf{U}_\mathbf{A}$. Thus $\mathbf{E}\mathbf{U}_\mathbf{A} = \mathbf{0}$ and $\mathbf{U}_\mathbf{A}\mathbf{S}_\mathbf{A}^{1/2} = \mathbf{U}_\mathbf{A}\mathbf{S}_\mathbf{A}\mathbf{U}_\mathbf{A}^\top\mathbf{U}_\mathbf{A}\mathbf{S}_\mathbf{A}^{-1/2} = (\mathbf{U}_\mathbf{A}\mathbf{S}_\mathbf{A}\mathbf{U}_\mathbf{A}^\top + \mathbf{E})\mathbf{U}_\mathbf{A}\mathbf{S}_\mathbf{A}^{-1/2} = \mathbf{A}\mathbf{U}_\mathbf{A}\mathbf{S}_\mathbf{A}^{-1/2}$. Similarly, $\mathbf{P} = \mathbf{U}_\mathbf{P}\mathbf{S}_\mathbf{P}\mathbf{U}_\mathbf{P}^\top$ (because $\mathbf{P}$ is rank d) and $\mathbf{U}_\mathbf{P}\mathbf{S}_\mathbf{P}^{1/2} = \mathbf{P}\mathbf{U}_\mathbf{P}\mathbf{S}_\mathbf{P}^{-1/2}$. Thus,

$$\begin{aligned}\zeta(f) &= \frac{1}{\sqrt{n}}\operatorname{tr}\big(\big(\mathbf{A}\mathbf{U}_\mathbf{A}\mathbf{S}_\mathbf{A}^{-1/2} - \mathbf{P}\mathbf{U}_\mathbf{P}\mathbf{S}_\mathbf{P}^{-1/2}\big)\mathbf{W}\big[\mathbf{M}(\partial f)\big]^\top\big)\\ &= \frac{1}{\sqrt{n}}\operatorname{tr}\big(\big(\mathbf{A}(\mathbf{U}_\mathbf{A} - \mathbf{U}_\mathbf{P})\mathbf{S}_\mathbf{A}^{-1/2} + \mathbf{A}\mathbf{U}_\mathbf{P}\big(\mathbf{S}_\mathbf{A}^{-1/2} - \mathbf{S}_\mathbf{P}^{-1/2}\big) + (\mathbf{A} - \mathbf{P})\mathbf{U}_\mathbf{P}\mathbf{S}_\mathbf{P}^{-1/2}\big)\mathbf{W}\big[\mathbf{M}(\partial f)\big]^\top\big).\end{aligned}$$

We therefore have

$$\begin{aligned}\sup_{f\in\mathcal{F}_\Phi}\big|\zeta(f)\big| \le{}& \frac{\sup_{f\in\mathcal{F}_\Phi}\|\mathbf{M}(\partial f)\|_F}{\sqrt{n}}\big(\big\|\mathbf{A}(\mathbf{U}_\mathbf{A} - \mathbf{U}_\mathbf{P})\mathbf{S}_\mathbf{A}^{-1/2}\mathbf{W}\big\|_F + \big\|\mathbf{A}\mathbf{U}_\mathbf{P}\big(\mathbf{S}_\mathbf{A}^{-1/2} - \mathbf{S}_\mathbf{P}^{-1/2}\big)\mathbf{W}\big\|_F\big)\\ &+ \frac{1}{\sqrt{n}}\sup_{f\in\mathcal{F}_\Phi}\big|\operatorname{tr}\big(\big[\mathbf{M}(\partial f)\big]^T(\mathbf{A} - \mathbf{P})\mathbf{U}_\mathbf{P}\mathbf{S}_\mathbf{P}^{-1/2}\mathbf{W}\big)\big|.\end{aligned} \tag{A.1}$$

We bound the first two terms on the right-hand side of equation (A.1) using the following result from [20]. □

Lemma 7. *Let* $(\mathbf{X}, \mathbf{A}) \sim \mathrm{RDPG}(F)$ *and 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*

$$\big\|\mathbf{A}(\mathbf{U_A} - \mathbf{U_P})\mathbf{S_A}^{-1/2}\big\|_F \le \frac{24\sqrt{2}d\log(n/\eta)}{\sqrt{\gamma^5(F)n}}, \tag{A.2}$$

$$\big\|\mathbf{A}\mathbf{U_P}\big(\mathbf{S_A}^{-1/2} - \mathbf{S_P}^{-1/2}\big)\big\|_F \le \frac{48d\log(n/\eta)}{\sqrt{\gamma^7(F)n}}, \tag{A.3}$$

where $\gamma(F)$ *is the minimum gap between the distinct eigenvalues of the matrix* $\mathbb{E}[X_1X_1^\top]$ *with* $X_1 \sim F$.

Equation (A.2) in the above lemma is a restatement of Lemma 3.4 in [20] where we have used the fact that the maximum row sum of $\mathbf{A}$ is n. Equation (A.3) follows from Lemma 3.2 in [20] and the fact that $\|\mathbf{M}_1\mathbf{M}_2\|_{2\to\infty} \le \|\mathbf{M}_1\|_{2\to\infty}\|\mathbf{M}_2\|$ for any matrices $\mathbf{M}_1$ and $\mathbf{M}_2$. As the individual bound in equations (A.2) and (A.3) holds with probabilty $1 - \eta$, they hold simultaneously with probability $1 - 2\eta$.

Lemma 7 then yields

$$\frac{\sup_{f\in\mathcal{F}_\Phi}\|\mathbf{M}(\partial f)\|_F}{\sqrt{n}}\big(\big\|\mathbf{A}(\mathbf{U_A} - \mathbf{U_P})\mathbf{S_A}^{-1/2}\mathbf{W}\big\|_F + \big\|\mathbf{A}\mathbf{U_P}\big(\mathbf{S_A}^{-1/2} - \mathbf{S_P}^{-1/2}\big)\mathbf{W}\big\|_F\big) \le \frac{C(F)\log n}{\sqrt{n}}$$

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 equation (A.1) 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. Denote by $\partial\mathcal{F}_\Phi$ the space of functions $\partial\mathcal{F}_\Phi = \{\partial f : f \in \mathcal{F}_\Phi\}$ from $\mathbb{R}^d$ to $\mathbb{R}^d$. For a given $\partial f \in \partial\mathcal{F}_\Phi$ let $\|\partial f\|_\infty$ denote the quantity $\sup_{X\in\Omega}\|(\partial f)(X)\|_2$, where $\|\cdot\|_2$ is the Euclidean norm in $\mathbb{R}^d$. Similarly, for given $\partial f, \partial g \in \partial\mathcal{F}_\Phi$, let $\|\partial f - \partial g\|_\infty$ denote $\sup_{X\in\Omega}\|(\partial f - \partial g)(X)\|_2$. As κ is twice continuously differentiable and Ω is compact, $\partial\mathcal{F}_\Phi$ is totally bounded with respect to $\|\cdot\|_\infty$. Put $\delta = \sup_{\partial f\in\partial\mathcal{F}_\Phi}\|\partial f\|_\infty$. Then for any $j \in \mathbb{N}$, we can find a finite subset $S_j = \{\partial f_1, \partial f_2, \ldots, \partial f_{n_j}\}$ of $\partial\mathcal{F}_\Phi$ such that for any $\partial f \in \partial\mathcal{F}_\Phi$, there exists a $\partial f_l \in S_j$ with $\|\partial f - \partial f_l\|_\infty \le \delta_j := 2^{-j}\delta$. We shall assume that S_j is *minimal* among all sets with the above property.

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

$$\begin{aligned}\tilde{\zeta}(f) &:= \frac{1}{\sqrt{n}}\sup_{f\in\mathcal{F}_\Phi}\big|\mathrm{tr}\big[\mathbf{M}(\partial f)\big]^\top(\mathbf{A} - \mathbf{P})\mathbf{U_P}\mathbf{S_P}^{-1/2}\mathbf{W}\big| \\ &= \sup_{f\in\mathcal{F}_\Phi}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n(\partial f)(X_i)^\top(\tilde{X}_i - X_i)\right|\end{aligned}$$

$$
\begin{aligned}
&= \sup_{f\in\mathcal{F}_\Phi}\left|\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=0}^{\infty}(\Pi_{j+1}\partial f-\Pi_j\partial f)(X_i)^\top(\tilde{X}_i-X_i)\right)+\frac{c_0}{\sqrt{n}}\right|\\
&= \sup_{f\in\mathcal{F}_\Phi}\left|\left(\frac{1}{\sqrt{n}}\sum_{j=0}^{\infty}\sum_{i=1}^{n}(\Pi_{j+1}\partial f-\Pi_j\partial f)(X_i)^\top(\tilde{X}_i-X_i)\right)+\frac{c_0}{\sqrt{n}}\right|\\
&\le \sum_{j=0}^{\infty}\sup_{f\in\mathcal{F}_\Phi}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\Pi_{j+1}\partial f-\Pi_j\partial f)(X_i)^\top(\tilde{X}_i-X_i)\right|+\left|\frac{c_0}{\sqrt{n}}\right|,
\end{aligned}
$$

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

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

$$
\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\Pi_{j+1}\partial f-\Pi_j\partial f)(X_i)^\top(\tilde{X}_i-X_i)\\
&\quad=\frac{1}{\sqrt{n}}\sum_{s=1}^{d}\big(\boldsymbol{\pi}_s^{(j,j+1)}(\partial f)\big)^\top(\mathbf{A}-\mathbf{P})\mathbf{u}_s\lambda_s^{-1/2},
\end{aligned}
\tag{A.4}
$$

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

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

$$
\mathbf{b}^T(\mathbf{A}-\mathbf{P})\mathbf{c}=2\sum_{i<j}b_i(\mathbf{A}-\mathbf{P})_{ij}c_j+\sum_i\mathbf{P}_{ii}b_ic_i.
$$

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

$$
\mathbb{P}\left[\left|2\sum_{i<j}b_i(\mathbf{A}-\mathbf{P})c_j\right|\ge t\right]\le 2\exp\left(-\frac{t^2}{8\sum_{i<j}b_i^2c_j^2}\right)\le 2\exp\left(-\frac{t^2}{8\|\mathbf{b}\|^2\|\mathbf{c}\|^2}\right).
$$

In addition, $\sum_i\mathbf{P}_{ii}b_ic_i\le\|\mathbf{b}\|\|\mathbf{c}\|$. We apply the above argument to equation (A.4). First, $\|\boldsymbol{\pi}_s^{(j,j+1)}(\partial f)\|_2\le 3/2\delta_j\sqrt{n}$ for all $\partial f\in\partial\mathcal{F}$. In addition, $\|\mathbf{u}_s\|=1$ for all s. Hence, for all $t\ge 2\delta_j\lambda_d^{-1/2}$,

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

for some universal constant $K > 0$. Let N_j be the cardinality of $\{\Pi_{j+1}\partial f - \Pi_j \partial f : f \in \mathcal{F}_\Phi\}$. Then by the union bound,

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

Now $N_j \leq |S_{j+1}|^2$ and hence, for any $t_j > 0$,

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}_\Phi}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\Pi_{j+1}\partial f - \Pi_j\partial f)(X_i)^T(\tilde{X}_i - X_i)\right| \geq \eta_j\right] \leq 2d\exp\bigl(-t_j^2\bigr), \tag{A.5}$$

where $\eta_j = d\sqrt{K\delta_j^2\lambda_d^{-1}(t_j^2 + \log|S_{j+1}|^2)}$. Summing equation (A.5) over all $j \geq 0$, and bounding $n^{-1/2}c_0$ using another application of Hoeffding's inequality, we arrive at

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}_\Phi}\bigl|\tilde{\zeta}(f)\bigr| \geq \sum_{j=0}^{\infty}K'\eta_j\right] \leq 2d\sum_{j=0}^{\infty}\exp\bigl(-t_j^2\bigr)$$

for some constant $K' > 0$. We now bound $\sum_{j=0}^{\infty}\eta_j = \sum_{j=0}^{\infty} d\sqrt{K\delta_j^2\lambda_d^{-1}(t_j^2 + \log|S_{j+1}|^2)}$. To bound $|S_j|$, we use the covering number for Ω, i.e., $|S_j| \leq (L/\delta_j)^d$ [33], Lemma 2.5, for some constant L independent of δ_j. Then by taking $t_j^2 = 2(\log j + \log n)$,

$$\mathbb{P}\Bigl[\sup_{f\in\mathcal{F}_\Phi}\bigl|\tilde{\zeta}(f)\bigr| \geq d\lambda_d^{-1/2}(C_1\log n + C_2)\Bigr] \leq \frac{2dC_3}{n^2} \tag{A.6}$$

for some constants C_1, C_2 and C_3. Equations (A.6) and (A.1) implies

$$\sup_{f\in\mathcal{F}_\Phi}\bigl|\zeta(f)\bigr| \leq \frac{C(F)\log n}{\sqrt{n}} + d\lambda_d^{-1/2}(C_1\log n + C_2) \tag{A.7}$$

with probability at least $1 - (1 + 2dC_3)n^{-2}$. Since there exists some constant $c > 0$ for which $\lambda_d/(cn) \to 1$ almost surely, an application of the Borel–Cantelli lemma to equation (A.7) yields $\sup_{f\in\mathcal{F}_\Phi}|\zeta(f)| \to 0$ almost surely. Lemma 3 is thus established.

A.1. Proof for the scaling case Section 5.1

The proof parallels that of Theorem 5. We sketch here the requisite modifications for the case when the null hypothesis $F \doteq G \circ c$ holds. Namely, we show that when $F \doteq G \circ c$ for some constant $c > 0$,

$$(m+n)(V_{n,m}(\hat{\mathbf{X}}/\hat{s}_X, \hat{\mathbf{Y}}/\hat{s}_Y) - V_{n,m}(\mathbf{X}/s_X, \mathbf{Y}\mathbf{W}_{n,m}/s_Y) \xrightarrow{\text{a.s}} 0. \tag{A.8}$$

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

$$\xi_W = \frac{\sqrt{m+n}}{n}\sum_{i=1}^{n}\kappa(\mathbf{W}_1 X_i/s_X, \cdot) - \frac{\sqrt{m+n}}{m}\sum_{k=1}^{m}\kappa(\mathbf{W}_2 Y_k/s_Y, \cdot),$$

$$\hat{\xi} = \frac{\sqrt{m+n}}{n}\sum_{i=1}^{n}\kappa(\hat{X}_i/\hat{s}_X, \cdot) - \frac{\sqrt{m+n}}{m}\sum_{k=1}^{m}\kappa(\hat{Y}_k/\hat{s}_Y, \cdot).$$

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

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

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

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

Lemma 2 implies $|r_1 + r_2| \to 0$ almost surely. Now denote $\sigma_X = (\mathbb{E}[\|X\|^2])^{1/2}$ and $\sigma_Y = (\mathbb{E}[\|Y\|^2])^{1/2}$. Then s_X and s_Y are $\sqrt{n}$-consistent and $\sqrt{m}$-consistent estimators of σ_X and σ_Y, respectively. When $F \stackrel{+}{=} G \circ c$, $\mu[F \circ \mathbf{T}_1 \circ \sigma_X^{-1}] = \mu[G \circ \mathbf{T}_2 \circ \sigma_Y^{-1}]$. Denote by $\xi_W^{(X)}$ and $\xi_W^{(Y)}$ the quantities

$$\xi_W^{(X)} = \sqrt{m+n}\left(\sum_{i=1}^{n}\frac{\kappa(\mathbf{T}_1 X_i/\sigma_X, \cdot) - \mu[F \circ \mathbf{T}_1 \circ \sigma_X^{-1}]}{n}\right),$$

$$\xi_W^{(Y)} = \sqrt{m+n}\left(\sum_{k=1}^{m}\frac{\kappa(\mathbf{T}_2 Y_k/\sigma_Y, \cdot) - \mu[G \circ \mathbf{T}_2 \circ \sigma_Y^{-1}]}{m}\right).$$

Then $\xi_W = \xi_W^{(X)} + \xi_W^{(Y)} + O(1)$ and hence $\xi_W - O(1)$ is once again a sum of independent mean zero random elements of $\mathcal{H}$. A Hilbert space concentration inequality similar to that of [23], Theorem 3.5, yields that $\|\xi\|_{\mathcal{H}}$ is bounded in probability.

We next bound $\|\xi_W - \hat{\xi}\|_{\mathcal{H}}$. We mimic the proof of Lemma 3, paying attention to the terms $\hat{s}_X$ and s_X. A Taylor expansion of κ yields

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

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

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

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

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

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 Lemma 3 and an application of Slutsky's theorem (for $\hat{s}_X \to (\mathbb{E}[\|X\|^2])^{1/2}$ almost surely). For the second sum on the right-hand side, we have

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

We note that $\|\hat{\mathbf{X}}\|_F^2 = \|\mathbf{S}_{\mathbf{A}}^{1/2}\|_F^2$ and $\|\mathbf{X}\|_F^2 = \|\mathbf{S}_{\mathbf{P}}^{1/2}\|_F^2$. Thus $|\|\hat{\mathbf{X}}\|_F^2 - \|\mathbf{X}\|_F^2| \le \sqrt{d}\|\mathbf{S}_{\mathbf{A}} - \mathbf{S}_{\mathbf{P}}\|_F$ by the Cauchy–Schwarz inequality. Lemma 3.2 in [20] can then be applied to $\|\mathbf{S}_{\mathbf{A}} - \mathbf{S}_{\mathbf{P}}\|_F$ to

show that $|\|\hat{\mathbf{X}}\|_F^2 - \|\mathbf{X}\|_F^2|$ is of order $O(\log n)$ with probability at least $1 - n^{-2}$; note that this bound for $\|\mathbf{S_A} - \mathbf{S_P}\|_F$ is much stronger than that obtained from Weyl's inequality and a concentration bound for $\|\mathbf{A} - \mathbf{P}\|$ from [17,18,22]. Hence by the compactness of Ω, smoothness of Φ and Slutsky's theorem, the second sum also converges to 0 almost surely, thereby establishing equation (A.8).

A.2. Proof for the projection case Section 5.2

The proof of this result is almost identical to that of Theorem 5. We note here the requisite modifications for the case when the null hypothesis of $F \circ \pi^{-1} \doteq G \circ \pi^{-1}$ holds. Define $\xi_W, \hat{\xi} \in \mathcal{H}$ by

$$\xi_W = \frac{\sqrt{m+n}}{n} \sum_{i=1}^{n} \kappa\big(\mathbf{W}_1 \pi(X_i), \cdot\big) - \frac{\sqrt{m+n}}{m} \sum_{k=1}^{m} \kappa\big(\mathbf{W}_2 \pi(Y_k), \cdot\big),$$

$$\hat{\xi} = \frac{\sqrt{m+n}}{n} \sum_{i=1}^{n} \kappa\big(\pi(\hat{X}_i), \cdot\big) - \frac{\sqrt{m+n}}{m} \sum_{k=1}^{m} \kappa\big(\pi(\hat{Y}_k), \cdot\big).$$

In addition, define $r_1 = r_{11} + r_{12}$ and $r_2 = r_{21} + r_{22}$ by

$$r_{11} = \frac{m+n}{n(n-1)} \sum_{i=1}^{n} \big(\kappa\big(\pi(X_i), \pi(X_i)\big) - \kappa\big(\pi(\hat{X}_i), \pi(\hat{X}_i)\big)\big),$$

$$r_{12} = \frac{m+n}{m(m-1)} \sum_{k=1}^{m} \big(\kappa\big(\pi(Y_k), \pi(Y_k)\big) - \kappa\big(\pi(\hat{Y}_k), \pi(\hat{Y}_k)\big)\big),$$

$$r_{21} = \frac{m+n}{n^2(n-1)} \sum_{i,j} \big(\kappa\big(\pi(X_i), \pi(X_j)\big) - \kappa\big(\pi(\hat{X}_i), \pi(\hat{X}_j)\big)\big),$$

$$r_{22} = \frac{m+n}{m^2(m-1)} \sum_{k,l} \big(\kappa\big(\pi(Y_k), \pi(Y_l)\big) - \kappa\big(\pi(\hat{Y}_k), \pi(\hat{Y}_l)\big)\big).$$

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

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

for some constant L depending only on κ. To complete the proof, we adapt the argument in the proof of Lemma 3 to the family of functions

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

to show that $\|\xi_W - \hat{\xi}\|_{\mathcal{H}} \to 0$ almost surely as $n \to \infty$.

References

[1] Airoldi, E.M., Blei, D.M., Fienberg, S.E. and Xing, E.P. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9** 1981–2014.

[2] Alba Fernández, V., Jiménez Gamero, M.D. and Muñoz García, J. (2008). A test for the two-sample problem based on empirical characteristic functions. *Comput. Statist. Data Anal.* **52** 3730–3748. MR2427377

[3] Anderson, N.H., Hall, P. and Titterington, D.M. (1994). Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *J. Multivariate Anal.* **50** 41–54. MR1292607

[4] Baringhaus, L. and Henze, N. (1988). A consistent test for multivariate normality based on the empirical characteristic function. *Metrika* **35** 339–348. MR0980849

[5] 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.

[6] Davis, C. and Kahan, W.M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.* **7** 1–46. MR0264450

[7] Diaconis, P. and Janson, S. (2008). Graph limits and exchangeable random graphs. *Rend. Mat. Appl.* (7) **28** 33–61. MR2463439

[8] Dobson, P.D. and Doig, A.J. (2003). Distinguishing enzyme structures from non-enzymes without alignments. *J. Mol. Biol.* **330** 771–781.

[9] Dudley, R.M. (1999). *Uniform Central Limit Theorems*. *Cambridge Studies in Advanced Mathematics* **63**. Cambridge: Cambridge Univ. Press. MR1720712

[10] Fishkind, D.E., Shen, C. and Priebe, C.E. (2015). On the incommensurability phenomenon. *J. Classification*. To appear.

[11] Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B. and Smola, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.* **13** 723–773. MR2913716

[12] 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. MR3176521

[13] Harchaoui, Z., Bach, F., Cappé, O. and Moulines, E. (2013). Kernel-based methods for hypothesis testing: A unified view. *IEEE Signal Process. Mag.* **30** 87–97.

[14] Hoff, P.D., Raftery, A.E. and Handcock, M.S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* **97** 1090–1098. MR1951262

[15] Holland, P.W., Laskey, K.B. and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. MR0718088

[16] Karrer, B. and Newman, M.E.J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* (3) **83** 016107, 10. MR2788206

[17] Lei, J. and Rinaldo, A. (2015). Consistency of spectral clustering in stochastic block models. *Ann. Statist.* **43** 215–237. MR3285605

[18] Lu, L. and Peng, X. (2013). Spectra of edge-independent random graphs. *Electron. J. Combin.* **20** Paper 27, 18. MR3158266

[19] Lyons, R. (2013). Distance covariance in metric spaces. *Ann. Probab.* **41** 3284–3305. MR3127883

[20] Lyzinski, V., Sussman, D.L., Tang, M., Athreya, A. and Priebe, C.E. (2014). Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. *Electron. J. Stat.* **8** 2905–2922. MR3299126

[21] Maa, J.-F., Pearl, D.K. and Bartoszyński, R. (1996). Reducing multidimensional two-sample data to one-dimensional interpoint comparisons. *Ann. Statist.* **24** 1069–1074. MR1401837

[22] Oliveira, R.I. (2009). Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. Preprint. Available at arXiv:0911.0600.

[23] Pinelis, I. (1994). Optimum bounds for the distributions of martingales in Banach spaces. *Ann. Probab.* **22** 1679–1706. MR1331198

[24] Sejdinovic, D., Sriperumbudur, B., Gretton, A. and Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Statist.* **41** 2263–2291. MR3127866

[25] Sriperumbudur, B.K., Fukumizu, K. and Lanckriet, G.R.G. (2011). Universality, characteristic kernels and RKHS embedding of measures. *J. Mach. Learn. Res.* **12** 2389–2410. MR2825431

[26] Steinwart, I. (2002). On the influence of the kernel on the consistency of support vector machines. *J. Mach. Learn. Res.* **2** 67–93. MR1883281

[27] Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. *Information Science and Statistics*. New York: Springer. MR2450103

[28] Sussman, D.L., Tang, M., Fishkind, D.E. and Priebe, C.E. (2012). A consistent adjacency spectral embedding for stochastic blockmodel graphs. *J. Amer. Statist. Assoc.* **107** 1119–1128. MR3010899

[29] Sussman, D.L., Tang, M. and Priebe, C.E. (2014). Consistent latent position estimation and vertex classification for random dot product graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* **36** 48–57.

[30] Székely, G.J. and Rizzo, M.L. (2009). Brownian distance covariance. *Ann. Appl. Stat.* **3** 1236–1265. MR2752127

[31] Székely, G.J. and Rizzo, M.L. (2013). Energy statistics: A class of statistics based on distances. *J. Statist. Plann. Inference* **143** 1249–1272. MR3055745

[32] Tang, M., Athreya, A., Sussman, D.L., Lyzinski, V. and Priebe, C.E. (2014). A semiparametric two-sample hypothesis testing problem for random dot product graphs. Preprint. Available at arXiv:1403.7249.

[33] van de Geer, S.A. (2000). *Empirical Processes in M-Estimation*. Cambridge: Cambridge Univ. Press.

[34] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes*: *With Applications to Statistics*. *Springer Series in Statistics*. New York: Springer. MR1385671

[35] Wolfe, P.J. and Olhede, S.C. (2013). Nonparametric graphon estimation. Preprint. Available at arXiv:1309.5936.

[36] 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. Reykjavik, Iceland.

[37] Young, S.J. and Scheinerman, E.R. (2007). Random dot product graph models for social networks. In *Algorithms and Models for the Web-Graph*. *Lecture Notes in Computer Science* **4863** 138–149. Berlin: Springer. MR2504912

Received July 2015 and revised November 2015