Two-sample smooth tests for the equality of distributions

WEN-XIN ZHOU[1,2,*], CHAO ZHENG[2,**] and ZHEN ZHANG[3]

[1]*Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA. E-mail:* [*]*wenxinz@princeton.edu*
[2]*School of Mathematics and Statistics, University of Melbourne, Parkville, VIC 3010, Australia. E-mail:* [**]*zhengc1@student.unimelb.edu.au*
[3]*Department of Statistics, University of Chicago, Chicago, IL 60637, USA. E-mail: zhangz19@galton.uchicago.edu*

This paper considers the problem of testing the equality of two unspecified distributions. The classical omnibus tests such as the Kolmogorov–Smirnov and Cramér–von Mises are known to suffer from low power against essentially all but location-scale alternatives. We propose a new two-sample test that modifies the Neyman's smooth test and extend it to the multivariate case based on the idea of projection pursue. The asymptotic null property of the test and its power against local alternatives are studied. The multiplier bootstrap method is employed to compute the critical value of the multivariate test. We establish validity of the bootstrap approximation in the case where the dimension is allowed to grow with the sample size. Numerical studies show that the new testing procedures perform well even for small sample sizes and are powerful in detecting local features or high-frequency components.

Keywords: goodness-of-fit; high-frequency alternations; multiplier bootstrap; Neyman's smooth test; two-sample problem

1. Introduction

Let X and Y be two $\mathbb{R}^p$-valued random variables with continuous distribution functions F and G, respectively, where $p \geq 1$ is a positive integer. Given data from each of the two *unspecified* distributions F and G, we are interested in testing the null hypothesis of the equality of distributions

$$H_0 : F = G \quad \text{versus} \quad H_1 : F \neq G. \tag{1.1}$$

This is the two-sample version of the conventional goodness-of-fit problem, which is one of the most fundamental hypothesis testing problems in statistics [34].

1.1. Univariate case: $p = 1$

Suppose we have two independent univariate random samples $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ and $\mathcal{Y}_m = \{Y_1, \ldots, Y_m\}$ from F and G, respectively. The empirical distribution functions (EDF) are given

by

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x) \quad \text{and} \quad G_m(y) = \frac{1}{m} \sum_{j=1}^{m} I(Y_j \leq y). \tag{1.2}$$

For testing the equality of two univariate distributions, conventional approaches in the literature use a measure of discrepancy between F_n and G_m as a test statistic. Prototypical examples include the Kolmogorov–Smirnov (KS) test $\hat{\Psi}_{\text{KS}} = \sqrt{nm/(n+m)} \sup_{t \in \mathbb{R}} |F_n(t) - G_m(t)|$ and the Cramér–von Mises (CVM) family of statistics

$$\hat{\Psi}_{\text{CVM}} = \frac{nm}{n+m} \int_{-\infty}^{\infty} \{F_n(t) - G_m(t)\}^2 w(H_{n+m}(t)) \, dH_{n,m}(t),$$

where $H_{n,m}(t) := \{nF_n(t) + mG_m(t)\}/(n+m)$ denotes the pooled EDF and w is a non-negative weight function. Taking $w \equiv 1$ yields the Cramér–von Mises statistic, and $w(t) = \{t(1-t)\}^{-1}$ yields the Anderson–Darling statistic [14].

The traditional omnibus tests, which have been widely used for testing the two-sample goodness-of-fit hypothesis (1.1) due to their simplicity with which they can be performed, suffer from low power in detecting densities containing high-frequency components or local features such as bumps, and thus may have poor finite sample power properties [21]. It is known from empirical studies that the CVM test has poor power against essentially all but location-scale alternatives [20]. The same issue arises in the KS test as well. To enhance power under local alternatives, Neyman's smooth method [38] was introduced earlier than the traditional omnibus tests, to test only the first d-dimensional sub-problem if there is prior that most of the discrepancies fall within the first d orthogonal directions. Essentially, Neyman's smooth tests represent a compromise between omnibus and directional tests. As evidenced by numerous empirical studies over the years, smooth tests have been shown to be more powerful than traditional omnibus tests over a broad range of realistic alternatives. See, for example, [4,20,21,31] and [19].

A two-sample analogue of the Neyman's smooth test was recently proposed by [5] for testing the equality of F and G based on two independent samples. The test statistic is asymptotically chi-square distributed and as a special case of Rao's score test, it enjoys certain optimality properties. Specifically, [5] motivated the two-sample Neyman's smooth test by considering the random variable $V = F(Y)$ with distribution and density functions given by

$$H(z) := G(F^{-1}(z)) \quad \text{and} \quad \rho(z) := g(F^{-1}(z))/f(F^{-1}(z)) \tag{1.3}$$

for $0 < z < 1$, respectively, where F^{-1} is the quantile function of X, i.e. $F^{-1}(z) = \inf\{x \in \mathbb{R} : F(x) \geq z\}$, and f and g denote the density functions of X and Y. Assume that F and G are strictly increasing, then H is also increasing, $\rho(z) \geq 0$ for $0 < z < 1$ and $\int_0^1 \rho(z) \, dz = 1$. Under the null hypothesis H_0, $\rho \equiv 1$ so that $V =_d U(0, 1)$. In other words, the null hypothesis H_0 in (1.1) is equivalent to

$$\tilde{H}_0 : \rho(z) = 1 \qquad \text{for all } 0 < z < 1, \tag{1.4}$$

where ρ is as in (1.3). Throughout, the function ρ is referred as the ratio density function. Based on Neyman's smooth test principle, we restrict attention to the following smooth alternatives to

the null of uniformity

$$\rho_\theta(z) = C_d(\theta) \exp\left\{ \sum_{k=1}^{d} \theta_k \psi_k(z) \right\} \qquad \text{for } \theta := (\theta_1, \ldots, \theta_d)^\mathsf{T} \in \mathbb{R}^d \text{ and } 0 < z < 1, \qquad (1.5)$$

which include a broad family of parametric alternatives, where $d = \dim(\theta)$ is some positive integer and $\{C_d(\theta)\}^{-1} = \int_0^1 \exp\{\sum_{k=1}^d \theta_k \psi_k(z)\}\, dz$. Setting $\psi_0 \equiv 1$, the functions $\psi_1, \ldots, \psi_d$ are chosen in such a way that $\{\psi_0, \psi_1, \ldots, \psi_d\}$ forms a set of orthonormal functions, that is,

$$\int_0^1 \psi_k(z)\psi_\ell(z)\, dz = \delta_{k\ell} = \begin{cases} 1, & \text{if } k = \ell, \\ 0, & \text{if } k \neq \ell. \end{cases} \qquad (1.6)$$

The null hypothesis asserts $H_0^d : \theta = 0$. Assuming that $m \leq n$ and the truncation parameter d is fixed, the two-sample smooth test proposed by [5] is defined as $\hat{\Psi}_{\text{BGX}} = m\hat{\boldsymbol{\psi}}^\mathsf{T}\hat{\boldsymbol{\psi}}$, where $\hat{\boldsymbol{\psi}} = m^{-1}\sum_{j=1}^m \boldsymbol{\psi}(\hat{V}_j)$, $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_d)^\mathsf{T}$ and $\hat{V}_j = F_n(Y_j)$. Under certain moment conditions and if the sample sizes (n, m) satisfy $m \log \log n = o(n)$ as $n, m \to \infty$, the test statistic $\hat{\Psi}_{\text{BGX}}$ converges in distribution to the χ^2 distribution with d degrees of freedom. Accounting for the error of estimating F, Bera, Ghosh and Xiao [5] further considered a generalized version of the smooth test that is asymptotically $\chi^2(d)$ distributed and can be applied when n and m are of the same magnitude.

However, Bera, Ghosh and Xiao [5] only focused on the fix d scenario (i.e. $d = 4$) so that their two-sample smooth test is consistent in power against alternative where $V = F(Y)$ does not have the same first k moments as that of the uniform distribution [34]. If there is a priori evidence that most of the energy is concentrated at low frequencies, that is, large θ_k are located at small k, it is reasonable to use Neyman's smooth test. Otherwise, Neyman's test is less powerful when testing contiguous alternatives with local characters [21]. As [31] pointed out, achieving reasonable power over more than a few orthogonal directions is hopeless. Indeed, the larger the value of d, the greater the number of orthogonal directions used to construct the test statistic. Therefore, it is possible to obtain consistency against all distributions if the truncation parameter d is allowed to increase with the sample size. Motivated by [10], we regard $H_0^d : \theta = 0$ as a global mean testing problem with dimension increasing with the sample size. When d is large, Neyman's smooth test which is based on the ℓ_2-norm of θ may also suffer from low powers under sparse alternatives. In part, this is because that the quadratic statistic accumulates high-dimensional estimation errors under H_0^d, resulting in large critical values that can dominate the signals under sparse alternatives.

To overcome the foregoing drawbacks, we first note that the traditional omnibus tests aim to capture the differences of two entire distributions as opposed to only assessing a particular aspect of the distributions, by contrast, Neyman's smooth principle reduces the original nonparametric problem to a d-dimensional parametric one. Lying in the middle, we are interested in enhancing the power in detecting two adjacent densities where one has local features or contains high-frequency components, while maintaining the same capability in detecting smooth alternative densities as the traditional tests. We expect to arrive at a compromise between desired significance level and statistical power by allowing the truncation parameter d to increase with sample sizes. In Section 2, we introduce a new test statistic by taking maximum over d univariate statistics. The limiting null distribution is derived under mild conditions, while d is allowed to grow

with n and m. To conduct inference, a novel intermediate approximation to the null distribution is proposed to compute the critical value. In fact, when n and m are comparable, d can be of order $n^{1/4}$ (resp. $n^{1/9}$) (up to logarithmic in n factors) if the trigonometric series (resp. Legendre polynomial series) is used to construct the test statistic.

1.2. Multivariate case: $p \geq 2$

As a canonical problem in multivariate analysis, testing the equality of two multivariate distributions based on the two samples has been extensively studied in the literature that can be dated back to [47], under the conventional fix p setting. Friedman and Rafsky [22] constructed a two-sample test based on the minimal spanning tree formed from the interpoint distances and their test statistic was shown to be asymptotically distribution free under the null; Schilling [42] and Henze [27] proposed nearest neighbor tests which are based on the number of times that the nearest neighbors come from the same group; Rosenbaum [40] proposed an exact distribution-free test based on a matching of the observations into disjoint pairs to minimize the total distance within pairs. Work in the context of nonparametric tests include that of [1,2,25] and [6], among others.

Most aforementioned existing methods are tailored for the case where the dimension p is fixed. Driven by a broad range of contemporary statistical applications, analysis of high-dimensional data is of significant current interest. In the high-dimensional setting, the classical testing procedures may have poor power performance, as evidenced by the numerical investigations in [6]. Several tests for the equality of two distributions in high dimensions have been proposed. See, for example, [25] and [6]. However, limiting null distributions of the test statistics introduced in [25] and [6] were derived when the dimension p is fixed.

In the present paper, we propose a new test statistic that extends Neyman's smooth test principle to higher dimensions based on the idea of projection pursue. To conduct inference for the test, we employ the multiplier (wild) bootstrap method which is similar in spirit to that used in [26] and [3]. We refer to Section 3 for details on methodologies. It can be shown that (Propositions 7.2 and 7.3), under mild conditions, the error in size of our multivariate smooth test decays polynomially in sample sizes (n, m). It is noteworthy that we allow the dimension p to grow as a function of (n, m), a type of framework the existing methods do not rigorously address. More importantly, we do not limit the dependency structure among the coordinates in X and Y and no shape constraints of the distribution curves are known as a priori which inhibits a pure parametric approach to the problem.

1.3. Organization of the paper

The rest of the paper is organized as follows. In Section 2, we describe the two-sample smooth testing procedure in the univariate case. An extension to the multivariate setting based on projection pursue is introduced in Section 3. Section 4 establishes theoretical properties of the proposed smooth tests in both univariate and multivariate settings. Finite sample performance of the proposed tests is investigated in Section 5 through Monte Carlo experiments. The proofs of the main results are given in Section 7 and some additional technical arguments are contained in Section 8.

Notation

For a positive integer p, we write $[p] = \{1, 2, \ldots, p\}$ and denote by $|\cdot|_2$ and $|\cdot|_\infty$ the ℓ_2- and ℓ_∞-norm in $\mathbb{R}^p$, respectively, that is, $|x|_2 = (x_1^2 + \cdots + x_p^2)^{1/2}$ and $|x|_\infty = \max_{k \in [p]} |x_k|$ for $x = (x_1, \ldots, x_p)^\mathsf{T} \in \mathbb{R}^p$. The unit sphere in $\mathbb{R}^p$ is denoted by $\mathcal{S}^{p-1} = \{x \in \mathbb{R}^p : |x|_2 = 1\}$. For two $\mathbb{R}^p$-valued random variables X and Y, we write $X =_d Y$ if they have the same probability distribution and denote by P_X the probability measure on $\mathbb{R}^p$ induced by X. For two real numbers a and b, we use the notation $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. For two sequences of positive numbers a_n and b_n, we write $a_n \asymp b_n$ if there exist constants $c_1, c_2 > 0$ such that for all sufficiently large n, $c_1 \le a_n/b_n \le c_2$, we write $a_n = O(b_n)$ if there is a constant $C > 0$ such that for all n large enough, $a_n \le C b_n$, and we write $a_n \sim b_n$ or $a_n \simeq b_n$ and $a_n = o(b_n)$, respectively, if $\lim_{n \to \infty} a_n/b_n = 1$ and $\lim_{n \to \infty} a_n/b_n = 0$. For any two functions $f, g : \mathbb{R} \mapsto \mathbb{R}$, we denote with $f \circ g$ the composite function $f \circ g(x) = f\{g(x)\}$ for $x \in \mathbb{R}$.

For any probability measure Q on a measurable space $(\mathbb{S}, \mathcal{S})$, let $\|\cdot\|_{Q,2}$ be $L_2(Q)$-seminorm defined by $\|f\|_{Q,2} = (Q|f|^2)^{1/2} = (\int |f|^2 \, dQ)^{1/2}$ for $f \in L_2(Q)$. For a class of measurable functions $\mathcal{F}$ equipped with an envelope $F(s) = \sup_{f \in \mathcal{F}} |f(s)|, s \in \mathbb{S}$, let $N(\mathcal{F}, L_2(Q), \varepsilon\|F\|_{Q,2})$ denote the ε-covering number of the class of functions $\mathcal{F}$ with respect to the $L_2(Q)$-distance for $0 < \varepsilon \le 1$. We say that the class $\mathcal{F}$ is *Euclidean* or *VC-type* [39,46] if there are constants $A, v > 0$ such that $\sup_Q N(\mathcal{F}, L_2(Q), \varepsilon\|F\|_{Q,2}) \le (A/\varepsilon)^v$ for all $0 < \varepsilon \le 1$, where the supremum ranges over all finitely discrete probability measures on $(\mathbb{S}, \mathcal{S})$. When $\mathbb{S} = \mathbb{R}^p$ for $p \ge 1$, we use $\mathcal{S}$ to denote the Borel σ-algebra unless otherwise stated.

2. Testing equality of two univariate distributions

2.1. Oracle procedure

Without loss of generality, we assume $n \ge m$ and recall that the null hypothesis $H_0 : F = G$ is equivalent to $\tilde{H}_0 : V =_d U(0, 1)$ for $V = F(Y)$ as in (1.4). Following [5], we consider the smooth alternatives lying in the family of densities (1.5) which is a d-parameter exponential family, where $d = d_{n,m}$ is allowed to increase with n and m in order to obtain power against a large array of alternatives. In particular, this family is quadratic mean differentiable at $\theta = 0$ and therefore the score vector at $\theta = 0$ is given by $m^{-1/2}(\frac{\partial}{\partial \theta_1} \log L_m(\theta), \ldots, \frac{\partial}{\partial \theta_d} \log L_m(\theta))|_{\theta=0}$ [34], where $L_m(\theta) = \{C_d(\theta)\}^m \exp\{\sum_{j=1}^m \sum_{k=1}^d \theta_k \psi_k(V_j)\}$ is the likelihood function and $V_j = F(Y_j)$, such that $\frac{\partial}{\partial \theta_k} \log L_m(\theta) = \sum_{j=1}^m [\psi_k(V_j) - E_\theta\{\psi_k(V_j)\}]$. As $\{\psi_0 \equiv 1, \psi_1, \ldots, \psi_d\}$ forms a set of orthonormal functions, it is easy to see that if $\theta = 0$, $E_\theta\{\psi_k(V)\} = 0$ and $E_\theta\{\psi_k(V)^2\} = 1$ for every $k \in [d]$.

To provide a more omnibus test against a broader range of alternatives, we allow a large truncation parameter d and for the reduced null hypothesis $H_0^d : \theta = 0$, it is instructive to consider the following oracle statistic

$$\Psi(d) = \max_{1 \le k \le d} \left| \frac{1}{\sqrt{m}} \sum_{j=1}^m \psi_k(V_j) \right|, \tag{2.1}$$

which can be regarded as a *smoothed* version of the KS statistic. Throughout, the number of orthogonal directions $d = \dim(\theta)$ is chosen such that $d \leq n \wedge m$. Intuitively, this extreme value statistic is appealing when most of the energy (non-zero θ_k) is concentrated on a few dimensions but with unknown locations, meaning that both low- and high-frequency alternations are possible. Now it is a common belief [8] that maximum-type statistics are powerful against sparse alternatives, which in the current context is the case where the two densities only differ in a small number of orthogonal directions (not necessarily in the first few). To see this, consider a contiguous alternative where there exists some ℓ^* such that $\theta_{\ell^*} \neq 0$ with $|\theta_{\ell^*}|$ sufficiently small and $\theta_\ell = 0$ for all other ℓ, then informally we have $C_d^{-1}(\theta) = \int_0^1 \exp\{\theta_{\ell^*}\psi_{\ell^*}(z)\}\,dz \simeq \int_0^1 \{1 + \theta_{\ell^*}\psi_{\ell^*}(z)\}\,dz = 1$ and

$$E_\theta\{\psi_k(V)\} = C_d(\theta) \int_0^1 \psi_k(z) \exp\{\theta_{\ell^*}\psi_{\ell^*}(z)\}\,dz \simeq \int_0^1 \psi_k(z)\{1 + \theta_{\ell^*}\psi_{\ell^*}(z)\}\,dz = \theta_{\ell^*}\delta_{k\ell^*}.$$

Under a sparse alternative where only a few components of $\theta = (\theta_1, \ldots, \theta_d)^\mathsf{T}$ are non-zero, the power typically depends on the magnitudes of the signals (non-zero coordinates of θ) and the number of the signals.

2.2. Data-driven procedure

For the oracle statistic $\Psi(d)$ in (2.1), the random variables $V_j = F(Y_j)$ are not directly observed as the distribution function F is *unspecified*. Indeed, this is the major difference of the two-sample problem from the classical (one-sample) goodness-of-fit problem. We therefore consider the following data-driven procedure. In the first stage, an estimate $\hat{V}_j$ of V_j is obtained by using the empirical distribution function F_n:

$$\hat{V}_j = F_n(Y_j) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq Y_j). \tag{2.2}$$

Then the data-driven version of $\Psi(d)$ in (2.1) is defined by

$$\hat{\Psi} = \hat{\Psi}(d) = \sqrt{\frac{nm}{n+m}} \max_{1 \leq k \leq d} |\hat{\psi}_k|, \tag{2.3}$$

where $\hat{\psi}_k = m^{-1}\sum_{j=1}^{m} \psi_k(\hat{V}_j)$. In the case $m > n$, we may use G_m instead of F_n, leading to an alternative test statistic $\tilde{\Psi}(d) = \sqrt{nm/(n+m)} \max_{1 \leq k \leq d} |n^{-1}\sum_{i=1}^{n} \psi_k(G_m(X_i))|$.

Typically, large values of $\hat{\Psi}$ lead to a rejection of the null $H_0^d : \theta = 0$ and hence of $H_0 : F = G$, or equivalently, $\tilde{H}_0$ in (1.4). For conducting inference, we need to compute the critical value so that the corresponding test has approximately size α. A natural approach is to derive the limiting distribution of the test statistic $\hat{\Psi}(d)$ under the null. Under certain smoothness conditions on ψ_k, it can be shown that for every $k \in [d]$,

$$\hat{\psi}_k = \frac{1}{m}\sum_{j=1}^{m} \psi_k(\hat{V}_j) \simeq \frac{1}{m}\sum_{j=1}^{m} \psi_k(V_j) - \frac{1}{n}\sum_{i=1}^{n} \psi_k(U_i),$$

where $U_i = G(X_i)$. See, for example, (7.7) and (7.10) in the proof of Proposition 7.1. Under H_0 and when $d \geq 1$ is fixed, a direct application of the multivariate central limit theorem is that as $n, m \to \infty$,

$$\sqrt{\frac{nm}{n+m}} (\hat{\psi}_1, \ldots, \hat{\psi}_d)^\intercal \xrightarrow{d} \mathbf{G} =_d N(0, \mathbf{I}_d), \tag{2.4}$$

where $\mathbf{I}_d$ is the d-dimensional identity matrix. This implies by the continuous mapping theorem that $\hat{\Psi}(d) \xrightarrow{d} |\mathbf{G}|_\infty$ when d is fixed.

For every $0 < \alpha < 1$, denote by z_α the $(1-\alpha)$-quantile of the standard normal distribution, i.e. $z_\alpha = \Phi^{-1}(1-\alpha)$. Then, the $(1-\alpha)$-quantile of $|\mathbf{G}|_\infty$ can be expressed as $c_\alpha(d) = z_{1/2-(1-\alpha)^{1/d}/2} = \Phi^{-1}(1/2 + (1-\alpha)^{1/d}/2)$. The corresponding asymptotic α-level *Smooth* test is thus defined as

$$\Phi_\alpha^S(d) = I\{\hat{\Psi}(d) \geq c_\alpha(d)\}. \tag{2.5}$$

The null hypothesis H_0 is rejected if and only if $\Phi_\alpha^S(d) = 1$.

To construct test that has better power for alternative densities with large energy at high frequencies, we allow the truncation parameter $d = \dim(\theta)$ to increase with sample sizes n and m. This setup was previously considered by [21] in the context of the Gaussian white noise model, where it was argued that if there is a priori evidence that large θ_k's are located at small k, then it is reasonable to select a relatively small d; otherwise the resulting test may suffer from low power in detecting densities containing high-frequency components. However, by letting d to increase with sample sizes we allow for different asymptotics than Neyman's fix d large sample scenario. This type of asymptotics aims to illustrate how the truncation parameter d may affect the quality of the test statistic, and to depict a more accurate picture of the behavior for fixed samples. In the present two-sample context, it will be shown (Proposition 7.1) that the distribution of $\hat{\Psi}(d)$ can still be consistently estimated by that of $|\mathbf{G}|_\infty$ with the truncation parameter d increasing polynomially in n and m, where $\mathbf{G}$ is a d-dimensional centered Gaussian random vector with covariance matrix $\mathbf{I}_d$. Consequently, the asymptotic size of the smooth test $\Phi_\alpha^S(d)$ in (2.5) coincides with the nominal size α (Theorem 4.1).

2.3. Choice of the function basis

In this paper, we shall focus the following two sets of orthonormal functions with respect to the Lebesgue measure on $[0, 1]$, which are the most commonly used basis for constructing smooth-type goodness-of-fit tests.

(i) (Legendre Polynomial (LP) series). Neyman's original proposal [38] was to use orthonormal polynomials, now known as the normalized Legendre polynomials. Specifically, ψ_k is chosen to be a polynomial of degree k which is orthogonal to all the ones before it and is normalized to size 1 as in (1.6). Setting $\psi_0 \equiv 1$, the next four ψ_k's are explicitly given by: $\psi_1(z) = \sqrt{3}(2z - 1)$, $\psi_2(z) = \sqrt{5}(6z^2 - 6z + 1)$, $\psi_3(z) = \sqrt{7}(20z^3 - 30z^2 + 12z - 1)$ and $\psi_4(z) = 3(70z^4 - 140z^3 + 90z^2 - 20z + 1)$. In general, the normalized Legendre polynomial of

order k can be written as

$$\psi_k(z) = \frac{\sqrt{2k+1}}{k!} \frac{d^k}{dz^k} (z^2 - z)^k, \qquad 0 \le z \le 1, k = 1, 2, \ldots. \tag{2.6}$$

See, for example, Lemma 1 in [5].

(ii) (Trigonometric series). Another widely used basis of orthonormal functions is a trigonometric series given by

$$\psi_k(z) = \sqrt{2} \cos(\pi k z), \qquad 0 \le z \le 1, k = 1, 2, \ldots. \tag{2.7}$$

This particular choice arises in the construction of the weighted quadratic type test statistics, including the Cramér–von Mises and the Anderson–Darling test statistics as prototypical examples [20,34]. Alternatively, one could use the Fourier series which is also a popular trigonometric series given by $\{\cos(2\pi k z), \sin(2\pi k z) : k = 1, \ldots, d/2\}$ for d even.

Other commonly used compactly supported orthonormal series include spline series [15], Cohen-Deubechies-Vial wavelet series [35] and local polynomial partition series [9], among others. As the two-sample test statistics constructed in this paper use orthonormal functions that are at least twice continuously differentiable on $[0, 1]$, we restrict attention to the Legendre polynomial series (2.6) and the trigonometric series (2.7) only. Indeed, the idea developed here can be directly applied to construct (one-sample) goodness-of-fit tests in one and higher dimensions without imposing smoothness conditions on the series.

3. Testing equality of two multivariate distributions

Evidenced by both theoretical (Section 4.1) and numerical (Section 5) studies, we see that Neyman's smooth test principle leads to convenient and powerful tests for univariate data. However, the presence of multivariate joint distributions makes it difficult, or even unrealistic, to consider a direct multivariate extension of the smooth alternatives given in (1.5). In the case of complete independence where all the components of X and Y are independent, the problem for testing equality of two multivariate distributions is equivalent to that for testing equality of many marginal distributions. Neyman's smooth principle can therefore be employed to each of the p marginals.

In this section, we do not impose assumption that limits the dependence among the coordinates in X and Y and note here that the null hypothesis $H_0 : F = G$ is equivalent to $H_0 : u^\mathsf{T} X =_d u^\mathsf{T} Y, \forall u \in \mathcal{S}^{p-1}$. This observation and the idea of projection pursue now allow to apply Neyman's smooth test principle, yielding a family of univariate smooth tests indexed by $u \in \mathcal{S}^{p-1}$ based on which we shall construct our test that incorporates the correlations among all the one-dimensional projections.

3.1. Test statistics

Assume that two independent random samples, $X_1, \ldots, X_n$ from the distribution F and $Y_1, \ldots, Y_m$ from the distribution G are observed, where the two samples sizes are comparable and $m \le n$. Along every direction $u \in \mathcal{S}^{p-1}$, let F^u and G^u be the distribution functions of

one-dimensional projections $u^\mathsf{T} X$ and $u^\mathsf{T} Y$, respectively, and define the corresponding empirical distribution functions by

$$F_n^u(x) = \frac{1}{n} \sum_{i=1}^n I\left(u^\mathsf{T} X_i \le x\right) \quad \text{and} \quad G_m^u(y) = \frac{1}{m} \sum_{j=1}^m I\left(u^\mathsf{T} Y_j \le y\right). \tag{3.1}$$

As a natural multivariate extension of the KS test, we consider the following test statistic

$$\hat{\Psi}_{\mathrm{MKS}} = \sqrt{\frac{nm}{n+m}} \sup_{(u,t)\in\mathcal{S}^{p-1}\times\mathbb{R}} \left| F_n^u(t) - G_m^u(t) \right|,$$

which coincides with the KS test when $p = 1$. Baringhaus and Franz [1] proposed a multivariate extension of the Cramér–von Mises test which is of the form

$$\hat{\Psi}_{\mathrm{BF}} = \frac{nm}{n+m} \int_{\mathcal{S}^{p-1}} \int_{-\infty}^{+\infty} \left\{ F_n^u(t) - G_m^u(t) \right\}^2 dt\, \vartheta(du),$$

where ϑ denotes the Lebesgue measure on $\mathcal{S}^{p-1}$. Despite their popularities in practice, the classical omnibus distribution-based testing procedures suffer from low power in detecting fine features such as sharp and short aberrants as well as global features such as high-frequency alternations [21]. Now it is well known that the foregoing drawbacks can be well repaired via smoothing-based test statistics. This motivates the following multivariate smooth test statistic.

As in Section 2, let $\{\psi_0 \equiv 1, \psi_1, \ldots, \psi_d\}$ $(d \ge 1)$ be a set of orthonormal functions and put $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_d)^\mathsf{T} : \mathbb{R} \mapsto \mathbb{R}^d$. Using the union-intersection principle, the two-sample problem of testing $H_0 : F = G$ versus $H_1 : F \ne G$ can be expressed a collection of univariate testing problems, by noting that H_0 and H_1 are equivalent to $\bigcap_{u\in\mathcal{S}^{p-1}} H_{u,0}$ and $\bigcup_{u\in\mathcal{S}^{p-1}} H_{u,1}$, respectively, where

$$H_{u,0} : u^\mathsf{T} X =_d u^\mathsf{T} Y, \qquad H_{u,1} : u^\mathsf{T} X \ne_d u^\mathsf{T} Y.$$

For every marginal null hypothesis $H_{u,0}$, we consider a smooth-type test statistic in the same spirit as in Section 2.1 that

$$\Psi_u(d) = \left| \frac{1}{m} \sum_{j=1}^m \boldsymbol{\psi}\left(V_j^u\right) \right|_\infty \qquad \text{with } V_j^u = F^u\left(u^\mathsf{T} Y_j\right). \tag{3.2}$$

For diagnostic purposes, it is interesting to find the best separating direction, that is, $u_{\max} := \arg\max_{u\in\mathcal{S}^{p-1}} \Psi_u(d)$, along which the two distributions differ most. For the purpose of conducting inference which is the main objective in this paper, we just plug $u_{\max}$ into (3.2) to get the oracle test statistic $\Psi_{\max}(d) := \Psi_{u_{\max}}(d)$, though it is practically infeasible as the distribution function F is unspecified. The most natural and convenient approach is to replace V_j^u in (3.2) with $\hat{V}_j^u = F_n^u(u^\mathsf{T} Y_j)$, leading to the following extreme value statistic for testing $H_0 : F = G$,

$$\hat{\Psi}_{\max} = \hat{\Psi}_{\max}(d) = \sqrt{\frac{nm}{n+m}} \sup_{u\in\mathcal{S}^{p-1}} \hat{\Psi}_u(d), \tag{3.3}$$

where $\hat{\Psi}_u(d) = |m^{-1}\sum_{j=1}^m \boldsymbol{\psi}(\hat{V}_j^u)|_\infty = \max_{1\le k\le d}|\hat{\psi}_{u,k}|$ with $\hat{\psi}_{u,k} = m^{-1}\sum_{j=1}^m \psi_k(\hat{V}_j^u)$. Rejection of the null is thus for large values of $\hat{\Psi}_{\max}$, say $\hat{\Psi}_{\max} > c_\alpha(d)$, where $c_\alpha(d)$ is a critical value to be determined so that the resulting test has the prespecified significance level $\alpha \in (0,1)$ asymptotically.

3.2. Critical values

Due to the highly complex dependence structure among $\{\hat{\psi}_{u,k}\}_{(u,k)\in\mathcal{S}^{p-1}\times[d]}$, the limiting (null) distribution of $\hat{\Psi}_{\max}(d)$ may not exist. In fact, $\hat{\Psi}_{\max}(d)$ can be regarded as the supremum of an empirical process indexed by the class

$$\hat{\mathcal{F}}_{n,m} := \left\{ x \mapsto \psi_k\left(\frac{1}{n}\sum_{i=1}^n I\{u^\mathsf{T}(x - X_i) \ge 0\}\right) : (u,k) \in \mathcal{S}^{p-1} \times [d] \right\}$$

of functions $\mathbb{R}^p \mapsto \mathbb{R}$; that is, $\hat{\Psi}_{\max}(d) = \sqrt{nm/(n+m)}\sup_{f\in\hat{\mathcal{F}}_{n,m}}|m^{-1}\sum_{j=1}^m f(Y_j)|$. As we allow the dimension p to grow with sample sizes, the "complexity" of $\hat{\mathcal{F}}_{n,m}$ increases with n,m and is thus non-Donsker. Therefore, the extreme value statistic $\hat{\Psi}_{\max}(d)$, even after proper normalization, may not be weakly convergent as $n, m \to \infty$. Tailored for such non-Donsker classes of functions that change with the sample size, Chernozhukov, Chetverikov and Kato [11] developed Gaussian approximations for certain maximum-type statistics under weak regularity conditions. This motivates us to take a different route by using the multiplier (wild) bootstrap method to compute the critical value $c_\alpha(d)$ for the statistic $\hat{\Psi}_{\max}(d)$ so that the resulting test has approximately size $\alpha \in (0,1)$.

Let $\{Z_1, \ldots, Z_N\} = \{Y_1, \ldots, Y_m, X_1, \ldots, X_n\}$ denote the pooled sample with a total sample size $N = n + m$. For every $(u,k) \in \mathcal{S}^{p-1} \times [d]$, we shall prove in Section 7.4.2 that

$$\sqrt{\frac{nm}{n+m}}\,\hat{\psi}_{u,k} \simeq \frac{1}{\sqrt{N}}\sum_{j=1}^N w_j\psi_k \circ F^u\big(u^\mathsf{T} Z_i\big),$$

where $w_j = \sqrt{n/m}$ for $j \in [m]$ and $w_j = -\sqrt{m/n}$ for $j \in m + [n]$. This implies that, under certain regularity conditions, $\hat{\Psi}_{\max}(d) \simeq \sup_{f\in\mathcal{F}_d^p}|N^{-1/2}\sum_{j=1}^N w_j f(Z_j)|$, where $\mathcal{F}_d^p = \{x \mapsto \psi_k \circ F^u(u^\mathsf{T} x) : (u,k) \in \mathcal{S}^{p-1} \times [d]\}$. Furthermore, we shall prove in Proposition 7.2 that, under the null hypothesis $H_0 : F = G$,

$$\hat{\Psi}_{\max}(d) \simeq_d \|\mathbb{G}\|_{\mathcal{F}_d^p} := \sup_{f\in\mathcal{F}_d^p}|\mathbb{G}f|, \tag{3.4}$$

where $\mathbb{G}$ is a centered Gaussian process indexed by $\mathcal{F}_d^p$ with covariance function

$$E\big\{(\mathbb{G}f_{u,k})(\mathbb{G}f_{v,\ell})\big\}$$
$$= P_X(f_{u,k}f_{v,\ell}) = E\big\{f_{u,k}(X)f_{v,\ell}(X)\big\} = \int_{\mathbb{R}^p}\psi_k\big(F^u(u^\mathsf{T}x)\big)\psi_\ell\big(F^v(v^\mathsf{T}x)\big)\,dF(x), \tag{3.5}$$

where $U^u = F^u(u^\mathsf{T} X) =_d \mathrm{Unif}(0, 1)$ for $u \in \mathcal{S}^{p-1}$. In particular, $E\{(\mathbb{G} f_{u,k})(\mathbb{G} f_{u,\ell})\} = \delta_{k\ell}$. The distribution of $\|\mathbb{G}\|_{\mathcal{F}_d^p}$, however, is unspecified because its covariance function is unknown. Therefore, in practice we need to replace it with a suitable estimator, and then simulate the Gaussian process $\mathbb{G}$ to compute the critical value $c_\alpha(d)$ numerically, as described below.

Multiplier bootstrap

(i) Independent of the observed data $\{X_i\}_{i=1}^n$ and $\{Y_j\}_{j=1}^m$, generate i.i.d. standard normal random variables $e_1, \ldots, e_n$. Then construct the *Multiplier Bootstrap* statistic

$$\hat{\Psi}_{\max}^{\mathrm{MB}} = \hat{\Psi}_{\max}^{\mathrm{MB}}(d) = \sup_{(u,k) \in \mathcal{S}^{p-1} \times [d]} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n e_i \psi_k(\hat{U}_i^u) \right|, \tag{3.6}$$

where $\hat{U}_i^u = F_n^u(u^\mathsf{T} X_i)$ for $F_n^u(\cdot)$ as in (3.1).

(ii) Calculate the data-driven critical value $\hat{c}_\alpha^{\mathrm{MB}}(d)$ which is defined as the conditional $(1-\alpha)$-quantile of $\hat{\Psi}_{\max}^{\mathrm{MB}}$ given $\{X_i\}_{i=1}^n$; that is,

$$\hat{c}_\alpha^{\mathrm{MB}}(d) = \inf\{t \in \mathbb{R} : P_e(\hat{\Psi}_{\max}^{\mathrm{MB}} > t) \leq \alpha\}, \tag{3.7}$$

where P_e denotes the probability measure induced by the normal random variables $\{e_i\}_{i=1}^n$ conditional on $\{X_i\}_{i=1}^n$.

For every $t \geq 0$, $P_e(\hat{\Psi}_{\max}^{\mathrm{MB}} \leq t)$ is a random variable depending on $\{X_i\}_{i=1}^n$ and so is $\hat{c}_\alpha^{\mathrm{MB}}(d)$, which can be computed with arbitrary accuracy via Monte Carlo simulations. Consequently, we propose the following *Multivariate Smooth* test

$$\Phi_\alpha^{\mathrm{MS}}(d) = I\{\hat{\Psi}_{\max}(d) \geq \hat{c}_\alpha^{\mathrm{MB}}(d)\}. \tag{3.8}$$

The null hypothesis $H_0 : F = G$ is rejected if and only if $\Phi_\alpha^{\mathrm{MS}}(d) = 1$.

4. Theoretical properties

Assume that we are given independent samples from the two (univariate and multivariate) distributions. As different sample sizes are allowed, for technical reasons we need to impose assumptions about the way in which samples sizes grow. The following gives the basic assumptions on the sampling process.

Assumption 4.1. (i) $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_m\}$ *are two independent random samples from* X *and* Y, *with absolute continuous distribution functions* F *and* G, *respectively;*

(ii) *The sample sizes,* n *and* m, *are comparable in the sense that* $c_0 n \leq m \leq n$ *for some constant* $0 < c_0 \leq 1$.

Let $\{\psi_0 \equiv 1, \psi_1, \ldots, \psi_d\}$ be a sequence of twice differentiable orthonormal functions $[0, 1] \mapsto \mathbb{R}$, where $d \geq 1$ is the truncation parameter. Moreover, for $\ell = 0, 1, 2$, define

$$B_{\ell d} = \max_{1 \leq k \leq d} \|\psi_k^{(\ell)}\|_\infty = \max_{1 \leq k \leq d} \max_{z \in [0,1]} |\psi_k^{(\ell)}(z)|. \tag{4.1}$$

These quantities will play a key role in our analysis. For the particular choice of the function basis as in (2.6) and (2.7), we specify below the order of $B_{\ell d}$, as a function of d, for $\ell = 0, 1, 2$.

(i) (Legendre polynomial series). For the normalized Legendre polynomials ψ_k, it is known that $B_{0d} = \max_{1 \le k \le d} \max_{z \in [0,1]} |\psi_k(z)| = \sqrt{2d+1}$. See, for example, [41]. Moreover, by the Markov inequality [43], $\|\psi_k'\|_\infty \le k^2 \|\psi_k\|_\infty$ and $\|\psi_k''\|_\infty \le \frac{k^2(k^2-1)}{3} \|\psi_k\|_\infty$. Together with (4.1), this implies

$$B_{0d} = \sqrt{2d+1}, \qquad B_{1d} \le \sqrt{3} d^{5/2} \quad \text{and} \quad B_{2d} \le 3^{-1/2} d^{9/2}. \tag{4.2}$$

(ii) (Trigonometric series). For the trigonometric series $\psi_k(z) = \sqrt{2} \cos(\pi k z)$, it is straight-forward to see that $\psi_k'(z) = -\sqrt{2}\pi k \sin(\pi k z)$ and $\psi_k''(z) = -\sqrt{2}\pi^2 k^2 \cos(\pi k z)$. Consequently, we have

$$B_{0d} = \sqrt{2}, \qquad B_{1d} \le \sqrt{2}\pi d \quad \text{and} \quad B_{2d} \le \sqrt{2}\pi^2 d^2. \tag{4.3}$$

4.1. Asymptotic properties of $\Phi_\alpha^S(d)$

Assumption 4.2. *The truncation parameter d is such that $d \le m$ and as $n \to \infty$,*

$$(\log n)^{7/6} B_{0d} = o(n^{1/6}), \qquad (\log n)^{3/2} B_{1d} = o(n^{1/2}), \qquad (\log n)^{1/2} B_{2d} = o(n^{1/2}).$$

The next theorem establishes the validity of the univariate smooth test $\Phi_\alpha^S(d)$ in (2.5).

Theorem 4.1. *Suppose that Assumptions 4.1 and 4.2 hold. Then as $n, m \to \infty$,*

$$\sup_{0 < \alpha < 1} \left| P_{H_0}\{\Phi_\alpha^S(d) = 1\} - \alpha \right| \to 0. \tag{4.4}$$

Remark 4.1. In view of (4.2) and (4.3), it follows from Theorem 4.1 that the error in size of the smooth test $\Phi_\alpha^S(d)$ using the trigonometric series (2.7) (resp. Legendre polynomials series (2.6)) tends to zero provided that $d = o\{(n/\log n)^{1/4}\}$ (resp. $d = o\{(n/\log n)^{1/9}\}$) as $n \to \infty$.

Next, we consider the asymptotic power of $\Phi_\alpha^S(d)$ against local alternatives when $d = d_{n,m} \to \infty$ as $n, m \to \infty$. For the following results, let $\bar{n} = 2(n^{-1} + m^{-1})^{-1}$ denote the *harmonic mean* of the two sample sizes. Our oracle statistic Ψ_S given in (2.1) mimics $\sqrt{\bar{n}/2} \max_{1 \le k \le d} |E_\theta\{\psi_k(V)\}|$, where $\theta = (\theta_1, \ldots, \theta_d)^\mathsf{T} \in \mathbb{R}^d$. Consider testing $\tilde{H}_0 : \rho \equiv 0$ in (1.4) against the following local alternatives

$$H_1^d : \rho = \rho_\theta, \qquad \text{for } \theta \in \Theta := \left\{ b = (b_1, \ldots, b_d)^\mathsf{T} \in \mathbb{R}^d : \max_{1 \le k \le d} |b_k| = \lambda \sqrt{\frac{\log d}{\bar{n}}} \right\}, \tag{4.5}$$

where ρ_θ is as in (1.5) and $\lambda > 0$ is a separation parameter. It is clear that the difficulty of testing between $\tilde{H}_0$ and H_1^d depends on the value of λ; that is, the smaller λ is, the harder it is to distinguish between the two hypotheses. The power of the test $\Phi_\alpha^S(d)$ in (2.5) against H_1^d is provided by the following theorem.

Theorem 4.2. *Suppose that Assumption* 4.1 *holds. The truncation parameter* $d = d_{n,m}$ *is such that* $d = o(n^{1/4})$ *if the trigonometric series* (2.7) *is used to construct the test statistic* $\hat{\Psi}(d)$ *in* (2.3) *and* $d = o(n^{1/9})$ *if the Legendre polynomials series* (2.6) *is used. Then, under* H_1^d *with* $\lambda \geq 2 + \varepsilon$ *for some* $\varepsilon > 0$,

$$\lim_{n,d \to \infty} P_{H_1^d}\{\Phi_\alpha^S(d) = 1\} = 1. \tag{4.6}$$

4.2. Asymptotic properties of $\Phi_\alpha^{MS}(d)$

In this section, we consider the multivariate case where the dimension $p = p_{n,m}$ is allowed to grow with sample sizes, and hence our results hold naturally for the fix dimension scenario. Specifically, we impose the following assumption on the quadruplet (n, m, p, d).

Assumption 4.3. *There exist constants* $C_0, C_1 > 0$ *and* $c_1 \in (0, 1)$ *such that*

$$d \leq \min\{n, m, \exp(C_0 p)\}, \qquad \max(p^7 B_{1d}^2, p B_{2d}^2) \leq C_1 n^{1-c_1}. \tag{4.7}$$

The next theorem establishes the validity of the multivariate smooth test $\Phi_\alpha^{MS}(d)$.

Theorem 4.3. *Suppose that Assumptions* 4.1 *and* 4.3 *hold. Then as* $n, m \to \infty$,

$$\sup_{0 < \alpha < 1} \left| P_{H_0}\{\Phi_\alpha^{MS}(d) = 1\} - \alpha \right| \to 0. \tag{4.8}$$

5. Numerical studies

In this section, we illustrate the finite sample performance of the proposed smooth tests described in Sections 2 and 3 via Monte Carlo simulations. The univariate and the multivariate cases will be studied separately.

5.1. Univariate case

Proposition 7.1 in Section 7 shows that the distribution of the test statistic $\hat{\Psi}(d)$ in (2.3) can be consistently estimated by that of the absolute Gaussian maximum $|\mathbf{G}|_\infty$, where $\mathbf{G} =_d N(0, \mathbf{I}_d)$. To see how close this approximation is, we compare in Figure 1 the cumulative distribution function of $|\mathbf{G}|_\infty$ and the empirical distributions of $\hat{\Psi}(d)$ using the trigonometric series (2.7) and the Legendre polynomial (LP) series (2.6), when the data are generated from Student's $t(7)$-distribution, with $n = 180, m = 150$ and $d = 12$. We only present the upper half of the curve since the $(1 - \alpha)$ quantile of $|\mathbf{G}|_\infty$ with $\alpha \in (0, 1/2)$ is of particular interest. It can be seen from Figure 1 that the cumulative distribution curves of $|\mathbf{G}|_\infty$ and the trigonometric series based statistic, denoted by T-$\hat{\Psi}(d)$, almost coincide, while there is a slightly noticeable difference between $|\mathbf{G}|_\infty$ and the LP polynomial series based statistic LP-$\hat{\Psi}(d)$. Indeed, this phenomenon can be expected from the theoretical discoveries. See, for example, the rate of convergence in (7.1) and (7.2), where dependence of $\{B_{\ell d}\}_{\ell=0,1,2}$ on d can be found in (4.2) and (4.3).
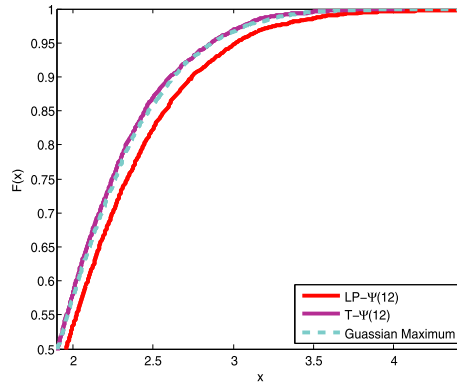
Figure 1. Comparison of the empirical cumulative distributions of LP-$\hat{\Psi}(12)$, T-$\hat{\Psi}(12)$ and the limiting cumulative distribution with $n = 180$ and $m = 150$. The plot is based on 5000 simulations.

Next, we carry out 5000 simulations with nominal significance level $\alpha = 0.05$ to calculate the empirical sizes of the proposed smooth test $\Phi_{\alpha}^{S}(d)$. We denote with T-$\Phi_{\alpha}^{S}(d)$ and LP-$\Phi_{\alpha}^{S}(d)$, respectively, the tests based on the trigonometric series (2.7) and the LP polynomial series (2.6). The sample sizes (n, m) are taken to be $(80, 60)$, $(120, 90)$, $(180, 150)$, and d takes values 4, 8, 12. We compare the proposed smooth test with the testing procedure proposed by [5], the two-sample Kolmogorov–Smirnov test and the two-sample Cramér–von Mises test in five examples when the data are generated from Gamma, Logistic, Gaussian, Pareto and Stable distributions. The results are summarized in Table 1, from which we see that among all the five examples considered, the empirical sizes of T-$\Phi_{\alpha}^{S}(d)$ with $d \in \{4, 8, 12\}$ are close to 0.05. This highlights the robustness of the testing procedure T-$\Phi_{\alpha}^{S}(d)$ with respect to the choice of the truncation parameter d. Further, we note that the empirical sizes of LP-$\Phi_{\alpha}^{S}(4)$ are comparable to those of Φ_{BGX}, while as d increases, the test LP-$\Phi_{\alpha}^{S}(d)$ suffers from size distortion gradually. In fact, as pointed out by [38] and [4], when the Legendre polynomials series is used to construct the test statistic, the effectiveness of the corresponding test in each direction could be diluted if d is too large. Nevertheless, the test based on the trigonometric series remains to be efficient as d increases and can be very powerful as we shall see later.

The power performance is evaluated through the following five examples. In each example, the result reported is based on 1000 simulations where samples sizes (n, m) are taken to be $(120, 90)$ and $(180, 150)$. Because of the distortion of empirical sizes of LP-$\Phi_{\alpha}^{S}(d)$, we only compare the power of the trigonometric series based smooth test T-$\Phi_{\alpha}^{S}(d)$ with that of the KS, CVM and BGX tests. The plots of power functions against different families of alternative distributions from Examples 1–5 are given in Figure 2.

Example 1.

$$X : F = \mathrm{uniform}(-1, 1) \quad \text{versus}$$

$$Y : G = G_{\mu} \qquad \text{with density } g_{\mu}(x) = \frac{1}{2} + 2x \frac{\mu - |x|}{\mu^2} I(|x| < \mu) \qquad (0 \leq \mu \leq 1).$$

Table 1. Comparison of empirical sizes with nominal significance level $\alpha = 0.05$

Model	(n, m)	T-$\Phi_\alpha^S(d)$			LP-$\Phi_\alpha^S(d)$			BGX	KS	CVM
		$d=4$	$d=8$	$d=12$	$d=4$	$d=8$	$d=12$			
Gamma(2, 2)	(80, 60)	0.0504	0.0500	0.0490	0.0584	0.0724	0.1078	0.0634	0.0494	0.0524
	(120, 90)	0.0504	0.0510	0.0484	0.0542	0.0654	0.0830	0.0590	0.0438	0.0434
	(180, 150)	0.0496	0.0486	0.0484	0.0500	0.0554	0.0706	0.0530	0.0440	0.0444
Logistic(0, 1)	(80, 60)	0.0498	0.0496	0.0482	0.0576	0.0748	0.1038	0.0618	0.0456	0.0494
	(120, 90)	0.0504	0.0500	0.0496	0.0528	0.0666	0.0860	0.0552	0.0504	0.0466
	(180, 150)	0.0502	0.0498	0.0500	0.0508	0.0570	0.0696	0.0574	0.0438	0.0424
N(0, 1)	(80, 60)	0.0504	0.0488	0.0470	0.0570	0.0764	0.1060	0.0648	0.0494	0.0504
	(120, 90)	0.0502	0.0494	0.0482	0.0548	0.0694	0.0850	0.0566	0.053	0.0504
	(180, 150)	0.0502	0.0514	0.0516	0.0504	0.0544	0.0616	0.0542	0.0446	0.0500
Pareto(0.5, 1, 1)	(80, 60)	0.0502	0.0484	0.0468	0.0582	0.0766	0.1064	0.0640	0.0460	0.0494
	(120, 90)	0.0500	0.0494	0.0494	0.0540	0.0640	0.0824	0.0592	0.0468	0.0480
	(180, 150)	0.0498	0.0496	0.0500	0.0530	0.0586	0.0724	0.0542	0.0436	0.0456
Stable(1.5, 0, 1, 1)	(80, 60)	0.0480	0.0470	0.0456	0.0544	0.0758	0.1088	0.0606	0.0474	0.0488
	(120, 90)	0.0496	0.0498	0.0494	0.0578	0.0692	0.0790	0.0614	0.0492	0.0514
	(180, 150)	0.0510	0.0514	0.0506	0.0508	0.0570	0.0690	0.0608	0.0490	0.0484

Example 2.

$$X : F = \text{uniform}(-1, 1) \quad \text{versus}$$

$$Y : G = G_\sigma \qquad \text{with density } g_\sigma(x) = \tfrac{1}{2}\left\{1 + \sin(2\pi\sigma x)\right\} \qquad (0.5 \leq \sigma \leq 5).$$

Example 3.

$$X : F = \text{lognormal}(0, 1) \qquad \text{with density } f(x) = (2\pi)^{-1/2} x^{-1} \exp\left\{-(\log x)^2/2\right\} \quad \text{versus}$$

$$Y : G = G_a \qquad \text{with density } g_a(x) = f(x)\left\{1 + a\sin(2\pi \log x)\right\} \qquad (-1 \leq a \leq 1).$$

Example 4.

$$X : F = \text{uniform}(0, 1) \quad \text{versus}$$

$$Y : G = G_c \qquad \text{with density } g_c(x) = \exp\left\{c\sin(5\pi x)\right\} \qquad (0 \leq c \leq 2).$$

Example 5.

$$X : F = \text{uniform}(0, 1) \quad \text{versus}$$

$$Y : G = G_c \qquad \text{with density } g_c(x) = 1 + c\cos(5\pi x) \qquad (0 \leq c \leq 2).$$
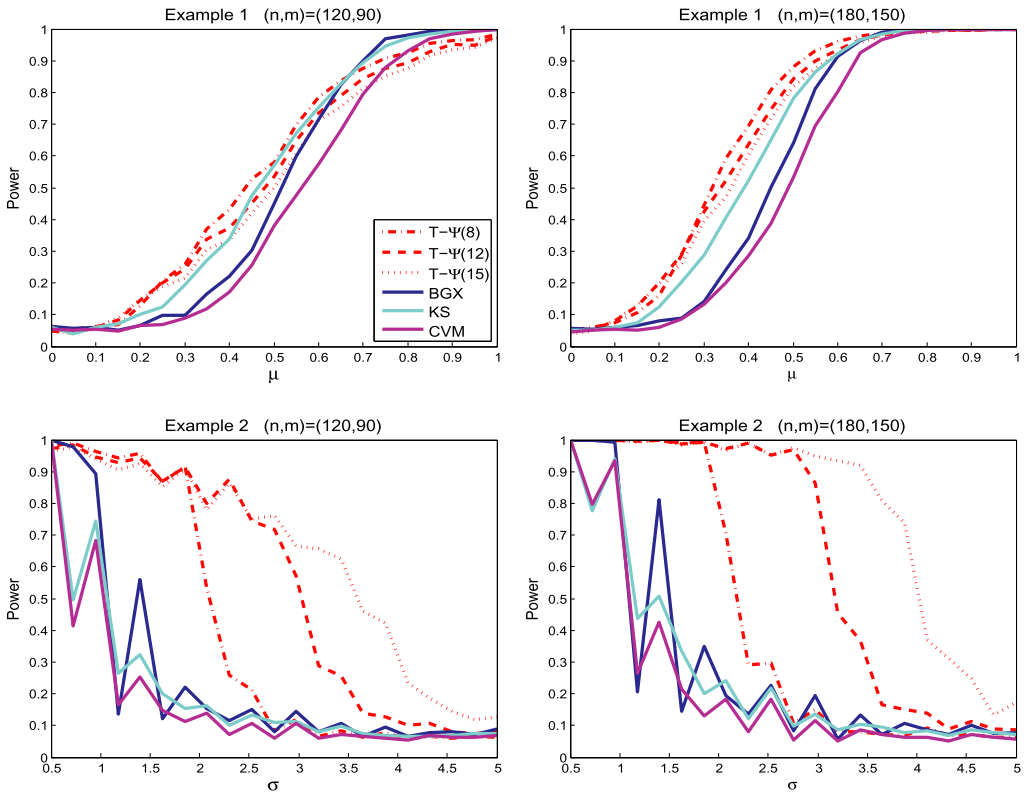
Figure 2. Empirical powers for Examples 1–5 based on 1000 replications with $\alpha = 0.05$.

The first two examples are, respectively, Examples 5 and 6 in [21] which were designed to demonstrate the performance of the adaptive Neyman's test proposed there. In Example 1, when $\mu = 0$, G_μ coincides with F. For this family of alternatives index by μ, the strength of the local feature depends on μ in the sense that the larger the μ, the stronger the local feature. As expected, the powers of all the tests considered grow with μ and when sample sizes are large enough, the smooth tests T-$\Phi_\alpha^S(d)$ uniformly outperform the others. Example 2, on the other hand, is designed to test the global features with various frequencies. It can be seen from the second row in Figure 2 that the test T-$\Phi_\alpha^S(16)$ has the highest power that approaches to 1 rapidly as σ decreases to 0. The third example is from [28], where g_a is a density and has the same moments as f_0 of any order. In this example, all the KS, CVM and BGX tests suffer from very poor power, while surprisingly, the smooth tests based on the trigonometric series remain powerful. The last two examples aim to cover the high-frequency alternations. Again, the proposed tests have the highest powers. In fact, the BGX test was originally constructed to identify deviations in the directions of mean, variance, skewness and kurtosis, and hence it can be relatively less powerful in detecting local features or high-frequency components.
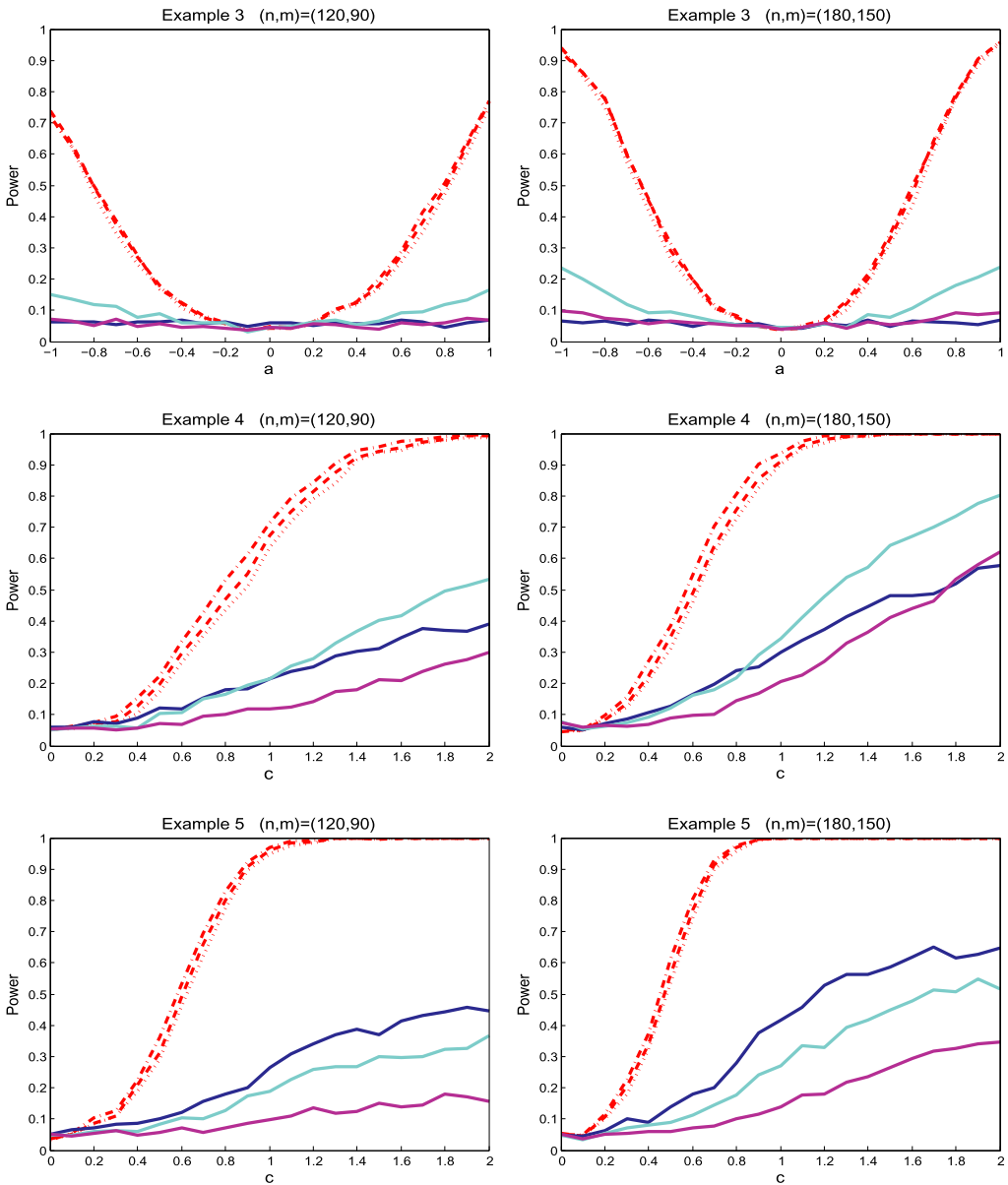
Figure 2. (Continued).

Table 2. Empirical size with significance level $\alpha = 0.05$

		Multivariate normal		Multivariate t			
		$N(0, I_p)$	$N(0, \Sigma)$	$t_4(0, I_p)$	$t_8(0, I_p)$	$t_4(0, \Sigma)$	$t_8(0, \Sigma)$
$p = 3$	T-$\Phi_\alpha^{\mathrm{MS}}(d)$	0.0446	0.0456	0.0514	0.0442	0.0494	0.0458
	BF	0.0480	0.0494	0.0504	0.0488	0.0448	0.0484
$p = 5$	T-$\Phi_\alpha^{\mathrm{MS}}(d)$	0.0496	0.0472	0.0494	0.0560	0.0450	0.0514
	BF	0.0466	0.0472	0.0502	0.0458	0.0488	0.0484
$p = 10$	T-$\Phi_\alpha^{\mathrm{MS}}(d)$	0.0582	0.0594	0.0512	0.0516	0.0570	0.0602
	BF	0.0422	0.0454	0.0364	0.0422	0.0482	0.0438

5.2. Multivariate case

The computation of the proposed multivariate smooth test and the critical value requires to find optimal directions $\hat{u}_{\max}$ and $\hat{u}_{\max}^{\mathrm{MB}}$ on the unit sphere $\mathcal{S}^{p-1}$ that maximize non-smooth objective functions (3.3) and (3.6), respectively. To solve these optimization problems, we convert the data into spherical coordinates and employ the Nelder-Mead algorithm. As a trade-off between the power and the computational feasibility of the test, we keep the value of d fixed at 4.

Similar to the univariate case, we first carry out 5000 simulations with nominal significance level $\alpha = 0.05$ to calculate the empirical sizes of the proposed test T-$\Phi_\alpha^{\mathrm{MS}}(d)$ with trigonometric series. For each $p \in \{3, 5, 10\}$, the data are generated from multivariate normal and t-distributions with different degrees of freedom (4 and 8) and covariance structures (I_p and Σ). Sample sizes (n, m) are taken to be (180, 160). We summarize the results in Table 2, comparing with the method proposed by [1], which will be referred as the BF test. From Table 2 we see that when $p = 3, 5$, both methods have an empirical size fairly close to 0.05; when $p = 10$, the empirical size of the proposed smooth test increases since the optimization over the unit sphere becomes more challenging, while the empirical size of the BF test is typically smaller than the nominal level.

The power performance of the multivariate smooth test is evaluated through Examples 6–9. The first two are multivariate versions of Examples 1 and 4, which demonstrate, respectively, the alternations with local feature and high frequency. The last two examples are designed to examine a rotation effect in the alternations. In each one, the power reported is based on 1000 simulations where samples sizes (n, m) are taken to be (180, 160). Again, we compare the power of the trigonometric series based smooth test T-$\Phi_\alpha^{\mathrm{S}}(d)$ with that of the BF test. The power curve are depicted in Figure 3.

Example 6.

$$X = (X_1, X_2, X_3)^{\mathsf{T}}, \qquad X_1, X_2 \overset{\text{i.i.d.}}{\sim} \text{uniform}(-1, 1), \qquad X_3 = 0.3X_1 + 0.7X_2 \quad \text{versus}$$

$$Y = (Y_1, Y_2, Y_3)^{\mathsf{T}}, \qquad Y_1, Y_2 \overset{\text{i.i.d.}}{\sim} g_\mu(x) = \frac{1}{2} + 2x \frac{\mu - |x|}{\mu^2} I(|x| < \mu) \qquad (0 \le \mu \le 1)$$

$$Y_3 = 0.3Y_1 + 0.7Y_2.$$

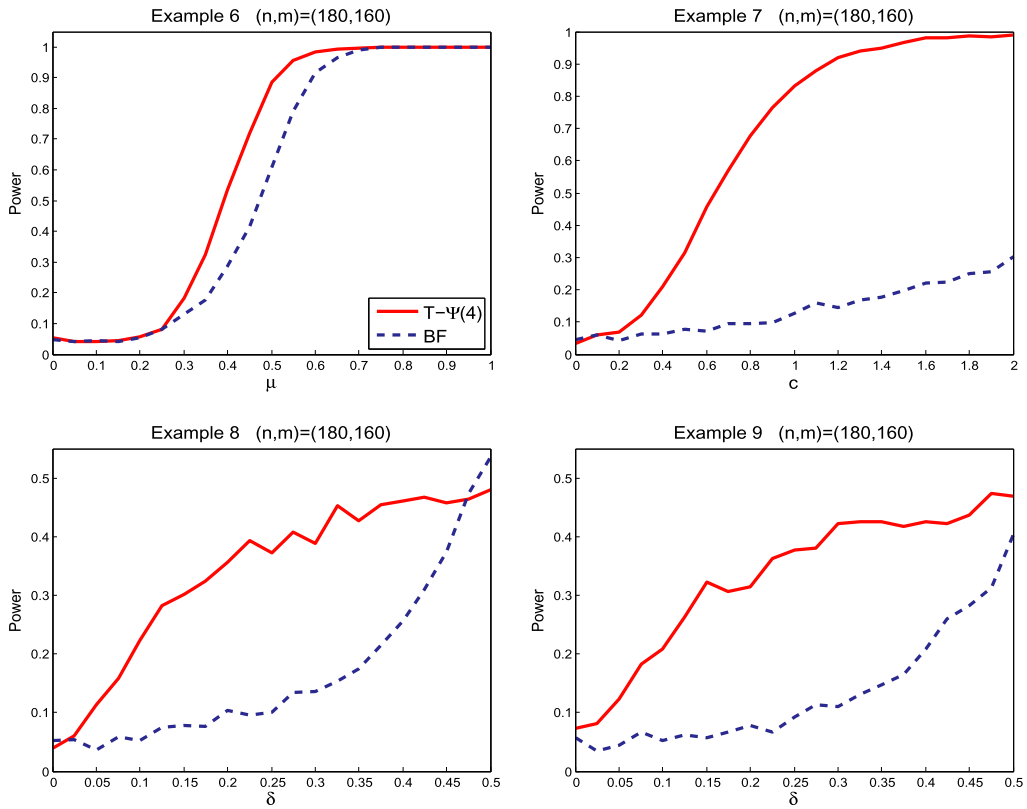**Figure 3.** Empirical powers for Examples 6–9 based on 1000 replications with $\alpha = 0.05$.

Example 7.

$$X = (X_1, X_2, X_3)^\mathsf{T}, \qquad X_1, X_2 \overset{\text{i.i.d.}}{\sim} \text{uniform}(0, 1), \qquad X_3 = 0.3X_1 + 0.7X_2 \quad \text{versus}$$

$$Y = (Y_1, Y_2, Y_3)^\mathsf{T}, \qquad Y_1, Y_2 \overset{\text{i.i.d.}}{\sim} g_c(x) = \exp\{c\sin(5\pi x)\} \qquad (0 \le c \le 2),$$

$$Y_3 = 0.3Y_1 + 0.7Y_2.$$

Example 8.

$$X \sim \mathcal{N}(0, I_5) \quad \text{versus} \quad Y = AZ, \qquad Z \sim \mathcal{N}(0, I_5),$$

where

$$A = \begin{pmatrix} A_0 & 0 \\ 0 & I_3 \end{pmatrix}, \qquad A_0 = \begin{pmatrix} \sqrt{1-\delta} & \sqrt{\delta} \\ \sqrt{\delta} & \sqrt{1-\delta} \end{pmatrix} \qquad (0 \le \delta \le 0.5).$$

Example 9.

$$X \sim t_4(0, I_5) \quad \text{versus} \quad Y = AZ, \qquad Z \sim t_4(0, I_5),$$

where

$$A = \begin{pmatrix} A_0 & 0 \\ 0 & I_3 \end{pmatrix}, \qquad A_0 = \begin{pmatrix} \sqrt{1-\delta} & \sqrt{\delta} \\ \sqrt{\delta} & \sqrt{1-\delta} \end{pmatrix} \qquad (0 \le \delta \le 0.5).$$

Figure 3 shows that the proposed smooth test uniformly outperforms the BF test in all the examples in terms of power. Since we are using trigonometric series, the test is powerful especially if the data contains high frequency components (Example 7), which is difficult to be detected by the BF test.

6. Discussion

We introduced in this paper a smooth test for the equality of two unknown distributions, which is shown to maintain the pre-specified significance level asymptotically. Moreover, it was shown theoretically and numerically that the test is especially powerful in detecting local features or high-frequency components.

The proposed procedure depends on a user-specific parameter d, which is the number of orthogonal directions used to construct the test statistic. Theoretically, the size of d is allowed to grow with n and can be as large as $o(n^c)$ for some $0 < c < 1$. Since the optimal value of d depends on how far the two unknown distributions deviate from each other, it is not possible to practically define an optimal choice of d. As suggested by our numerical studies, $d = 10$ is a reasonable choice when the sample sizes are in the order of 10^2, which leads a good compromise between the computational cost and the performance of the test. Alternatively, a data-driven approach based on a modification of Schwarz's rule was proposed by [30], that is, $\hat{d} = \arg\max_{1 \le d \le D(n,m)} \{T(d) - d\log(n+m)\}$ for some $D(n, m) \to \infty$ as $\min(n, m) \to \infty$, where $T(d)$ is the test statistic using the first d orthonormal functions. This principal can be applied to the proposed testing procedure by setting $D(n, m)$ to be some large value, say 20. Nevertheless, the optimal choice of $D(n, m)$ remains unclear.

The computation of the multivariate test statistic $\hat{\Psi}_{\max}(d)$ requires solving the optimization problem with an ℓ_2-norm constraint. To solve this problem when the dimension p is relatively small, we first convert the data into spherical coordinates and then use the Nelder-Mead algorithm. An interesting extension is to combine our method with the smoothing technique as in [29]. Let $K : \mathbb{R} \mapsto \mathbb{R}$ be a symmetric, bounded density function. For a predetermined small number $h = h_n > 0$, $\hat{\psi}_{u,k}$ is approximated by a continuous function $\hat{\psi}_{u,k,h} = m^{-1} \sum_{j=1}^m \psi_k(\hat{V}_{j,h}^u)$, where

$$\hat{V}_{j,h}^u = \frac{1}{n} \sum_{i=1}^n \mathcal{K}\left\{ \frac{u^\mathsf{T}(Y_j - X_i)}{h} \right\} \qquad \text{with } \mathcal{K}(t) = \int_{-\infty}^t K(z)\, dz.$$

As $h \to 0$, $\hat{V}_{j,h}^u$ converges to $\hat{V}_j^u$ almost surely, and hence for each $k \in [d]$, $\sup_{u \in \mathcal{S}^{p-1}} |\hat{\psi}_{k,u,h}|$ is a smoothed version of $\sup_{u \in \mathcal{S}^{p-1}} |\hat{\psi}_{u,k}|$. The smoothing technique can be similarly applied

to the multiplier bootstrap statistic. Consequently, we can employ the gradient descent algorithm to solve the optimization for smooth functions. We leave a thorough comparison of various algorithms for different values of p as an interesting problem for future research.

7. Proof of the main results

In this section, we prove Theorems 4.1–4.3. Proofs of the lemmas and some additional technical arguments are given in Section 8. Throughout this section, we write $N = n + m$ and use C and c to denote absolute positive constants, which may take different values at each occurrence. We write $a \lesssim b$ if a is smaller than or equal to b up to an absolute positive constant, and $a \gtrsim b$ if $b \lesssim a$.

7.1. Proof of Theorem 4.1

Recall that $\mathbf{G} = (G_1, \ldots, G_d)^{\mathsf{T}}$ is a d-dimensional standard Gaussian random vector, the distribution of $|\mathbf{G}|_\infty$ is absolute continuous so that $P\{|\mathbf{G}|_\infty \geq c_\alpha(d)\} = \alpha$. Therefore, under the assumption that $d \leq n \wedge m$, the conclusion (4.8) follows from the following proposition immediately.

Proposition 7.1. *Assume that the conditions of Theorem* 4.1 *hold and let*

$$\gamma_{0n} = \frac{(\log n)^{7/8}}{n^{1/8}} B_{0d}, \qquad \gamma_{1n} = \frac{(\log d)^{3/2}}{\sqrt{n}} B_{1d}, \qquad \gamma_{2n} = \sqrt{\frac{\log d}{n}} B_{2d}. \qquad (7.1)$$

Then under $H_0 : F = G$,

$$\sup_{t \geq 0} \left| P\{\hat{\Psi}(d) \leq t\} - P(|\mathbf{G}|_\infty \leq t) \right| \lesssim \gamma_{0n} + \sqrt{\gamma_{1n}} + \sqrt{\gamma_{2n}}. \qquad (7.2)$$

The proof of Proposition 7.1 is provided in Section 7.4.1.

7.2. Proof of Theorem 4.2

For the d-dimensional Gaussian random vector $\mathbf{G}$, applying the Borell–TIS (Borell–Tsirelson–Ibragimov–Sudakov) inequality [46] yields that for every $t > 0$, $P\{|\mathbf{G}|_\infty > E(|\mathbf{G}|_\infty) + t\} \leq \exp(-t^2/2)$. By taking $t = \sqrt{2\log(1/\alpha)}$, we get

$$c_\alpha(d) \leq E(|\mathbf{G}|_\infty) + \sqrt{2\log(1/\alpha)}, \qquad (7.3)$$

where $c_\alpha(d)$ denotes the $(1 - \alpha)$-quantile of $|\mathbf{G}|_\infty$. A standard result on Gaussian maxima yields $E(|\mathbf{G}|_\infty) \leq \{1 + (2\log d)^{-1}\}\sqrt{2\log d}$.

Let $k^* = \arg\max_{k \in [d]} |\theta_k|$ under H_1^d and assume without loss of generality that $\theta_{k^*} > 0$. By (7.7) and (7.10) in the proof of Proposition 7.1, we have

$$
\begin{aligned}
P_{H_1^d}\{\hat{\Psi} > c_\alpha(d)\} &\geq P_{H_1^d}\left\{\sqrt{\frac{nm}{N}}\,\hat{\psi}_{k^*} > c_\alpha(d)\right\} \\
&= P_{H_1^d}\left\{\frac{1}{\sqrt{N}}\sum_{j=1}^N \xi_{jk^*} + \sqrt{\frac{nm}{N}}(R_{1k^*} + R_{2k^*}) > c_\alpha(d)\right\},
\end{aligned}
\tag{7.4}
$$

where $\xi_{jk} = \sqrt{n/m}\{\psi_k(V_j) - \vartheta_k\}I\{j \in [m]\} + \sqrt{m/n}\,h_{1k}(X_{j-m})I\{j \in m + [n]\}$ for $(j,k) \in [N] \times [d]$ with $V_j = F(Y_j)$, $\vartheta_k = E_{H_1^d}\{\psi_k(V)\}$ and $h_{1k}(x) = E_{H_1^d}(\psi_k'(V)[I\{V \geq F(x)\} - V])$. Note that $E\{h_{1k}(X)\} = 0$ and thus $E(\xi_{jk}) = 0$. Let $\mathcal{E}(t_1, t_2)$ be as in (7.12) for $t_1, t_2 > 0$ to be specified. Put $\delta = t_1 B_{2d} + t_2 B_{1d} + \sqrt{nm/N}(\theta_{k^*} - \vartheta_{k^*}) + \sqrt{2\log(1/\alpha)}$, then it follows from (7.3) and (7.4) that

$$
\begin{aligned}
&P_{H_1^d}\{\hat{\Psi} > c_\alpha(d)\} \\
&\geq P_{H_1^d}\left\{\frac{1}{\sqrt{N}}\sum_{j=1}^N \xi_{jk^*} > \left(1 + \frac{1}{2\log d}\right)\sqrt{2\log d} + \delta - \sqrt{\frac{nm}{N}}\theta_{k^*}\right\} - P\{\mathcal{E}(t_1, t_2)^c\} \\
&\geq P_{H_1^d}\left\{\frac{1}{\sqrt{N}}\sum_{j=1}^N \xi_{jk^*} > \left(\frac{1}{2\log d} - \frac{\varepsilon}{2}\right)\sqrt{2\log d} + \delta\right\} - P\{\mathcal{E}(t_1, t_2)^c\}.
\end{aligned}
$$

In particular, taking $t_1 = t_{1n}(d) \asymp n^{-1/2}\sqrt{\log d}$ and $t_2 = t_{2n}(d) \asymp n^{-1/2}\log d$ implies by (7.16) that $P\{\mathcal{E}(t_1, t_2)^c\} \to 0$ as $d \to \infty$. Further, by (8.8) and the conditions of the theorem, we have $\delta = o(\sqrt{\log d})$. Consequently, as $d \to \infty$,

$$
P_{H_1^d}\{\hat{\Psi} > c_\alpha(d)\} \geq 1 - P_{H_1^d}\left(\frac{1}{\sqrt{N}}\sum_{j=1}^N \xi_{jk^*} \leq -\frac{\varepsilon}{2}\sqrt{\log d}\right) - P\{\mathcal{E}(t_1, t_2)^c\} \to 1.
$$

This completes the proof of Theorem 4.2.

7.3. Proof of Theorem 4.3

We first introduce two propositions describing the limiting null properties of the multivariate smooth and multiplier bootstrap statistics used to construct the test. The conclusion of Theorem 4.3 follows immediately.

The first proposition characterizes the non-asymptotic behavior of the multivariate smooth statistic $\hat{\Psi}_{\max}(d)$ which involves the supremum of a centered Gaussian process. Let $\mathcal{F} = \mathcal{F}_d^p$ be as in (3.4) and for simplicity, the dependence of $\mathcal{F}$ on (p, d) will be assumed without displaying.

Proposition 7.2. *Suppose that Assumptions* 4.1 *and* 4.3 *hold. Then there exists a centered, tight Gaussian process* $\mathbb{G}$ *indexed by* $\mathcal{F}$ *such that under the null hypothesis* $H_0 : F = G$,

$$\sup_{t \geq 0} \left| P\left\{ \hat{\Psi}_{\max}(d) \leq t \right\} - P\left(\|\mathbb{G}\|_{\mathcal{F}} \leq t \right) \right| \leq C n^{-c}, \tag{7.5}$$

where C *and* c *are positive constants depending only on* c_0, c_1, C_0 *and* C_1.

Proposition 7.2 implies that the "limiting" distribution of $\hat{\Psi}_{\max}$ depends on unknown the covariance structure given in (3.5). To compute a critical value, we suggest to use multiplier bootstrapping as described in Section 3.2. The following result, which can be regarded as a multiplier central limit theorem, provides the theoretical justification of its validity. In fact, the construction of the multiplier bootstrap statistic $\hat{\Psi}_{\max}^{\mathrm{MB}}(d)$ involves the use of artificial random numbers to simulate a process, the supremum of which is (asymptotically) equally distributed as $\|\mathbb{G}\|_{\mathcal{F}}$ according to Proposition 7.3 below.

Proposition 7.3. *Suppose that Assumptions* 4.1 *and* 4.3 *hold. Then with probability at least* $1 - 3n^{-1}$,

$$\sup_{t \geq 0} \left| P_e\left\{ \hat{\Psi}_{\max}^{\mathrm{MB}}(d) \leq t \right\} - P\left(\|\mathbb{G}\|_{\mathcal{F}} \leq t \right) \right| \leq C n^{-c} \tag{7.6}$$

for $\mathbb{G}$ *as defined in Proposition* 7.2, *where* C *and* c *are positive constants depending only on* c_0, c_1, C_0 *and* C_1.

Proofs of the above two propositions are given in Section 7.4.

7.4. Proof of Propositions 7.1–7.3

7.4.1. *Proof of Proposition* 7.1

For every $k \in [d]$, it follows from (2.2) and Taylor expansion that

$$\frac{1}{m} \sum_{j=1}^{m} \psi_k(\hat{V}_j) = \frac{1}{m} \sum_{j=1}^{m} \psi_k(V_j) + \frac{1}{m} \sum_{j=1}^{m} \psi_k'(V_j)(\hat{V}_j - V_j) + \frac{1}{2m} \sum_{j=1}^{m} \psi_k''(\xi_j)(\hat{V}_j - V_j)^2$$

$$= \frac{1}{m} \sum_{j=1}^{m} \psi_k(V_j) + \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \psi_k'(V_j)\left\{ I(X_i \leq Y_j) - F(Y_j) \right\} + R_{1k}, \tag{7.7}$$

where $R_{1k} := (2m)^{-1} \sum_{j=1}^{m} \psi_k''(\zeta_j)(\hat{V}_j - V_j)^2$ and ζ_j is a random variable lying between $\hat{V}_j$ and V_j. It is straightforward to see that $R_{1k} \leq \frac{1}{2} \|\psi_k''\|_{\infty} \max_{1 \leq j \leq m}(\hat{V}_j - V_j)^2$. A direct consequence of the Dvoretzky–Kiefer–Wolfwitz inequality [36], that is, for every $t > 0$, $P\{\sqrt{n} \sup_x |F_n(x) - F(x)| > t\} \leq 2 \exp(-2t^2)$, is that

$$P\left(n \max_{1 \leq k \leq d} |R_{1k}| / \|\psi_k''\|_{\infty} > t \right) \leq 2 \exp(-4t). \tag{7.8}$$

Let $h_k(x, y) = \psi_k'(F(y))\{I(x \le y) - F(y)\}$ for $x, y \in \mathbb{R}$ be a kernel function $\mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$. Then the second addend on the right-hand side of (7.7) can be written as $U_{n,m}(k) = (nm)^{-1} \sum_{i=1}^n \sum_{j=1}^m h_k(X_i, Y_j)$ with $E\{h_k(X, Y)\} = 0$. Observer that $U_{n,m}(k)$ is a two-sample U-statistic with a bounded kernel h_k satisfying $b_k := \|h_k\|_\infty \le \|\psi_k'\|_\infty$ and

$$\sigma_k^2 := E\{h_k(X, Y)^2\} = E\{(V - V^2)\psi_k'(V)^2\} \le \|\psi_k'\|_\infty^2/4. \tag{7.9}$$

Let $h_{1k}(x) = E_{H_0^d}\{h_k(X, Y)|X = x\}$ and $h_{2k}(y) = E_{H_0^d}\{h_k(X, Y)|Y = y\}$ be the first order projections of the kernel h_k under H_0^d. Since X and Y are independent and under H_0, $V = F(Y) =_d$ Unif$(0, 1)$ under H_0, we have $h_{2k} \equiv 0$ and

$$h_{1k}(x) = E\big(\psi_k'(V)[I\{V \ge F(x)\} - V]\big) = \int_{F(x)}^1 \psi_k'(v)\, dv - \int_0^1 v\psi_k'(v)\, dv = -\psi_k\big(F(x)\big).$$

Define random variables $U_i = F(X_i) =_d$ Unif$(0, 1)$ that are independent of $V_j = F(Y_j)$. Then, using the Hoeffding's decomposition gives

$$U_{n,m}(k) = -\frac{1}{n}\sum_{i=1}^n \psi_k(U_i) + \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m h_{0k}(X_i, Y_j) := -\frac{1}{n}\sum_{i=1}^n \psi_k(U_i) + R_{2k}, \tag{7.10}$$

where $h_{0k}(x, y) = h_k(x, y) - h_{1k}(x) - h_{2k}(y)$.

In view of (7.7) and (7.10), we introduce a new sequence of independent random vectors $\{\boldsymbol{\xi}_j = (\xi_{j1}, \dots, \xi_{jK})^\mathsf{T}\}_{j=1}^N$ for $N = n + m$, defined by

$$\xi_{jk} = \begin{cases} \sqrt{n/m}\,\psi_k(V_j), & 1 \le j \le m, \\ -\sqrt{m/n}\,\psi_k(U_{j-m}) & m + 1 \le j \le N. \end{cases} \tag{7.11}$$

Put $\boldsymbol{\psi} = (\psi_1, \dots, \psi_d)^\mathsf{T}$, $\mathbf{R}_1 = (R_{11}, \dots, R_{1d})^\mathsf{T}$ and $\mathbf{R}_2 = (R_{21}, \dots, R_{2d})^\mathsf{T}$, such that

$$\sqrt{\frac{n}{mN}}\sum_{j=1}^m \boldsymbol{\psi}(\hat{V}_j) = \frac{1}{\sqrt{N}}\sum_{j=1}^N \boldsymbol{\xi}_j + \sqrt{\frac{nm}{N}}(\mathbf{R}_1 + \mathbf{R}_2).$$

Recall that $\{\psi_0 \equiv 1, \psi_1, \dots, \psi_d\}$ is a set of orthonormal functions and $V =_d$ Unif$(0, 1)$ under H_0. By (7.11), the covariance matrix of $N^{-1/2}\sum_{j=1}^N \boldsymbol{\xi}_j$ is equal to $\mathbf{I}_d$.

For any $t_1, t_2 > 0$, define the event

$$\mathcal{E}(t_1, t_2) = \bigcap_{k=1}^d \{\sqrt{m}|R_{1k}| \le \|\psi_k''\|_\infty t_1\} \cap \{\sqrt{m}|R_{2k}| \le \|\psi_k'\|_\infty t_2\}. \tag{7.12}$$

Under H_0, we have for every $t > 0$,

$$P_{H_0}\{\hat{\Psi}(d) \le t\}$$
$$= P\left\{\max_{1 \le k \le d}\left|\sqrt{\frac{n}{mN}}\sum_{j=1}^m \psi_k(\hat{V}_j)\right| \le t\right\} \tag{7.13}$$

$$= P\left\{ \max_{1\le k\le d} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \xi_{jk} + \sqrt{\frac{nm}{N}}(R_{1k}+R_{2k}) \right| \le t \right\}$$

$$\le P\left\{ \max_{1\le k\le d} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \xi_{jk} \right| \le t + \sqrt{\frac{n}{N}}(t_1 B_{2d}+t_2 B_{1d}) \right\} + P\{\mathcal{E}(t_1,t_2)^{\mathrm{c}}\},$$

where $B_{\ell d}$ $(\ell = 1, 2)$ are as in (4.1). To get rid of the absolute value in (7.13), a similar argument as in the proof of Theorem 1 in [10] gives

$$P\left(\max_{1\le k\le d} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \xi_{jk} \right| \le t \right) = P\left(\max_{1\le k\le 2d} \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \xi_{jk}^{\mathrm{ext}} \le t \right), \tag{7.14}$$

where $\{\boldsymbol{\xi}_j^{\mathrm{ext}}\}_{j=1}^{N}$ is a sequence of dilated random vectors taking values in $\mathbb{R}^{2d}$ defined by $\boldsymbol{\xi}_j^{\mathrm{ext}} = (\xi_{j1}^{\mathrm{ext}},\ldots,\xi_{j,2d}^{\mathrm{ext}})^{\mathsf{T}} = (\boldsymbol{\xi}_j^{\mathsf{T}}, -\boldsymbol{\xi}_j^{\mathsf{T}})^{\mathsf{T}}$. In view of (7.14), we only need to focus on $\max_{1\le k\le d} N^{-1/2} \sum_{j=1}^{N} \xi_{jk}$ without losing generality.

Note that ξ_{jk} are bounded random variables satisfying $E(\xi_{jk}) = 0$ and $|\xi_{jk}| \le \sqrt{\frac{n}{m}} \|\psi_k\|_{\infty}$. Applying Lemmas 2.3 and 2.1 in [11], respectively, yields

$$\sup_{t\in\mathbb{R}} \left| P\left(\max_{1\le k\le d} \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \xi_{jk} \le t \right) - P\left(\max_{1\le k\le d} G_k \le t \right) \right| \lesssim \frac{\{\log(dn)\}^{7/8}}{n^{1/8}} B_d,$$

where $B_d := [E\{\max_{1\le k\le d} |\psi_k(V)|^3\}]^{1/4} \le B_{0d}^{3/4}$, $\mathbf{G} = (G_1,\ldots,G_d)^{\mathsf{T}} =_d N(0,\mathbf{I}_d)$ and for every $\varepsilon > 0$,

$$\sup_{t\in\mathbb{R}} P\left(\left| \max_{1\le k\le d} G_k - t \right| \le \varepsilon \right) \le 4\varepsilon(1 + \sqrt{2\log d}).$$

The last two displays jointly imply

$$P\left\{ \max_{1\le k\le d} \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \xi_{jk} \le t + \sqrt{\frac{n}{N}}(t_1 B_{2d}+t_2 B_{1d}) \right\}$$

$$\le P\left(\max_{1\le k\le d} G_k \le t \right) + C\left\{ \frac{\{\log(dn)\}^{7/8}}{n^{1/8}} B_{0d}^{3/4} + (\log d)^{1/2}(t_1 B_{2d}+t_2 B_{1d}) \right\}. \tag{7.15}$$

For $P\{\mathcal{E}(t_1,t_2)^{\mathrm{c}}\}$ in (7.13), it follows from (7.8) and (8.3) in Lemma 8.2 that

$$P\{\mathcal{E}(t_1,t_2)^{\mathrm{c}}\} \le 2\exp(-4t_1 n/\sqrt{m}) + \sum_{k=1}^{d} P\left(\sqrt{m}|R_{2k}| > \|\psi_k'\|_{\infty} t_2/2 \right)$$

$$\lesssim \exp(-4t_1\sqrt{n}) + d\exp(-ct_2\sqrt{n}). \tag{7.16}$$

Taking $t_1 \asymp (\gamma_{2n} n)^{-1/2}$, $t_2 \asymp (\gamma_{1n} n)^{-1/2} \log d$ in (7.13) implies by (7.15) and (7.16) that

$$P_{H_0}(\hat{\Psi} \leq t) \leq P\big(|\mathbf{G}|_\infty \leq t\big) + C\left[\frac{\{\log(dn)\}^{7/8}}{n^{1/8}} B_{0d}^{3/4} + \sqrt{\gamma_{1n}} + \sqrt{\gamma_{2n}}\right],$$

where $\gamma_{\ell n}$ ($\ell = 1, 2$) are as in (7.1). Here, the last inequality relies on the fact that $\sup_{t \geq 0}(te^{-t}) \leq e^{-1}$. A similar argument leads to the reverse inequality and thus completes the proof.

7.4.2. *Proof of Proposition 7.2*

In view of (4.7), we assume without loss of generality that $B_{1d} \leq \sqrt{n}$. Let $\mathcal{T} = \mathcal{T}_d^p$ be the product space $\mathcal{S}^{p-1} \times [d]$. For every $u \in \mathcal{S}^{p-1}$, let $\hat{V}_j^u = F_n^u(u^\mathsf{T} Y_j)$, $V_j^u = F^u(u^\mathsf{T} Y_j)$, $j \in [m]$ and $U_i^u = F^u(u^\mathsf{T} X_i)$, $i \in [n]$. By Taylor expansion and arguments similar to those employed in the proof of Proposition 7.1, we obtain that for every $(u, k) \in \mathcal{T}$,

$$
\begin{aligned}
\sqrt{\frac{nm}{N}} \hat{\psi}_{u,k} &= \sqrt{\frac{n}{mN}} \sum_{j=1}^m \psi_k\big(\hat{V}_j^u\big) \\
&= \sqrt{\frac{n}{mN}} \sum_{j=1}^m \big\{\psi_k\big(V_j^u\big) - \psi_k'\big(V_j^u\big)V_j^u\big\} + \sqrt{\frac{1}{nmN}} U_{n,m}(u, k) + \sqrt{\frac{nm}{N}} R_{u,k},
\end{aligned}
\tag{7.17}
$$

where $U_{n,m}(u, k) := \sum_{i=1}^n \sum_{j=1}^m \psi_k'(V_j^u) I\{u^\mathsf{T}(X_i - Y_j) \leq 0\}$ is a two-sample U-statistic with $E\{U_{n,m}(u, k)\} = \psi_k(1)$ under H_0 and $|R_{u,k}| \leq \frac{1}{2}\|\psi_k''\|_\infty \max_{j \in [m]}(\hat{V}_j^u - V_j^u)^2$. Let

$$\mathcal{H} = \mathcal{H}_d^p = \big\{h_{u,k}(\cdot, \cdot) : \mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbb{R} | (u, k) \in \mathcal{T}\big\} \tag{7.18}$$

be a class of measurable functions, where $h_{u,k}(x, y) = \psi_k'(F^u(u^\mathsf{T} y)) I\{u^\mathsf{T}(x - y) \leq 0\}$ for $x, y \in \mathbb{R}^p$. For ease of exposition, the dependence of $\mathcal{T}$ and $\mathcal{H}$ on (p, d) will be assumed without displaying. In the above notation, each $h = h_{u,k} \in \mathcal{H}$ determines a two-sample U-statistic $U_{n,m}(h) := \sum_{i=1}^n \sum_{j=1}^m h(X_i, X_j) = U_{n,m}(u, k)$, such that $\{U_{n,m}(h)\}_{h \in \mathcal{H}}$ forms a two-sample U-process indexed by the class $\mathcal{H}$ of kernels. Moreover, define the degenerate version of $\mathcal{H}$ as

$$\mathcal{H}_0 = \big\{h_0(x, y) = h(x, y) - (P_Y h)(x) - (P_X h)(y) + (P_X \times P_Y)(h)|h \in \mathcal{H}\big\}, \tag{7.19}$$

where for $h(\cdot, \cdot) : \mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbb{R}$,

$$(P_X h)(\cdot) = \int_{\mathbb{R}^p} h(x, \cdot) \, dF(x), \qquad (P_Y h)(\cdot) = \int_{\mathbb{R}^p} h(\cdot, y) \, dG(y)$$

and $(P_X \times P_Y)(h) = \int \int h(x, y) \, dF(x) \, dG(y)$. Under H_0, it is easy to verify that for every $(u, k) \in \mathcal{T}$, $(P_X \times P_Y)(h_{u,k}) = \psi_k(1)$,

$$(P_Y h_{u,k})(x) = \psi_k(1) - \psi_k\big(F^u\big(u^\mathsf{T} x\big)\big) \quad \text{and} \quad (P_X h_{u,k})(y) = \psi_k'\big(F^u\big(u^\mathsf{T} y\big)\big) F^u\big(u^\mathsf{T} y\big).$$

In addition to $\mathcal{H}$ and $\mathcal{H}_0$, we define the following class of measurable functions on $\mathbb{R}^p$:

$$\mathcal{F} = \mathcal{F}_d^p = \big\{x \mapsto f_{u,k}(x) = \psi_k \circ f_u(x) : k \in [d],\, f_u \in \mathcal{F}^p\big\}, \tag{7.20}$$

where

$$\mathcal{F}^p = \left\{y \mapsto f_u(y) = \int I_u(x, y)\, dF(x) : I_u \in \mathcal{I}^p\right\} \tag{7.21}$$

with $\mathcal{I}^p = \{(x, y) \mapsto I\{u^{\mathsf{T}}(x - y) \leq 0\} : u \in \mathcal{S}^{p-1}\}$.

Together, (7.17) and (7.19)–(7.21) lead to

$$|\hat{\Psi}_{\max} - \Psi_0| \leq \frac{1}{\sqrt{N}}\left\{\frac{1}{\sqrt{nm}}\|U_{n,m}\|_{\mathcal{H}_0} + \sqrt{nm}\, \sup_{(u,k)\in\mathcal{T}} |R_{u,k}|\right\}, \tag{7.22}$$

where $\|U_{n,m}\|_{\mathcal{H}_0} = \sup_{h_0 \in \mathcal{H}_0} |U_{n,m}(h_0)|$,

$$\Psi_0 = \sup_{f\in\mathcal{F}} \left|\frac{1}{\sqrt{N}} \sum_{j=1}^{N} w_j \big\{f(Z_j) - P_X f\big\}\right| \tag{7.23}$$

with $\{Z_1, \ldots, Z_N\} = \{Y_1, \ldots, Y_m, X_1, \ldots, X_n\}$ and $w_j = \sqrt{n/m}\,I\{j \in [m]\} - \sqrt{m/n}\,I\{j \in [n] + m\}$ for $j = 1, \ldots, N$.

With the above preparations, the rest of the proof involves three steps: First, approximation of the test statistic $\hat{\Psi}_{\max}$ by Ψ_0 requires the uniform negligibility of the right side of (7.22). Second, we prove the Gaussian approximation of Ψ_0 by the supremum of a centered, tight Gaussian process $\mathbb{G}$ indexed by $\mathcal{F}$ with covariance function

$$E\big\{(\mathbb{G}f_{u,k})(\mathbb{G}f_{v,\ell})\big\} = \int_{\mathbb{R}^p} \psi_k\big(F^u(u^{\mathsf{T}}x)\big)\psi_\ell\big(F^v(v^{\mathsf{T}}x)\big)\, dF(x) \tag{7.24}$$

for $(u, k), (v, \ell) \in \mathcal{T}$. Finally, we apply an anti-concentration argument due to [13] to construct the Berry–Esseen type bound.

Step 1. The following two results show the uniform negligibility of the right-hand side of (7.22).

Lemma 7.1. *Assume that the conditions of Proposition* 7.2 *hold. Then under* $H_0 : F = G$,

$$E\big(\|U_{n,m}\|_{\mathcal{H}_0}\big) \lesssim B_{2d}\sqrt{(p + \log d)nm}. \tag{7.25}$$

Lemma 7.2. *With probability at least* $1 - 2n^{-1}$, *we have*

$$\sup_{(u,t)\in\mathcal{S}^{p-1}\times\mathbb{R}} \big|F_n^u(t) - F^u(t)\big| \lesssim \sqrt{\frac{p + \log n}{n}}. \tag{7.26}$$

By (7.25) and (7.26), it follows from the Markov inequality that for $t > 0$,

$$P\big\{(nm)^{-1/2}\|U_{n,m}\|_{\mathcal{H}_0} > t\big\} \lesssim t^{-1} B_{2d}\sqrt{p + \log d} \tag{7.27}$$

and with probability at least $1 - 2n^{-1}$, $\sqrt{nm}\sup_{(u,k)\in\mathcal{T}}|R_{u,k}| \lesssim B_{2d}\sqrt{p + \log n}$. Taking $t = \gamma^{-1} B_{2d}\sqrt{p + \log d}$ for some $\gamma \in (0, 1)$ in (7.27) implies by (7.22) that

$$P\left(|\hat{\Psi}_{\max} - \Psi_0| \gtrsim B_{2d}\frac{\sqrt{p + \log d + \log n}}{\gamma\sqrt{n}}\right) \lesssim \gamma + n^{-1}. \tag{7.28}$$

Step 2. The following result establishes the Gaussian approximation for Ψ_0.

Lemma 7.3. *Assume that the conditions of Proposition* 7.2 *hold. Then under* H_0, *there exists a centered, tight Gaussian process* $\mathbb{G}$ *indexed by* $\mathcal{F} = \mathcal{F}_d^p$ *given in* (7.20) *with covariance function* (7.24) *and a random variable* $\Psi^* =_d \|\mathbb{G}\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}}|\mathbb{G}f|$ *such that for every* $\gamma \in (0, 1)$,

$$P\left\{|\Psi_0 - \Psi^*| \gtrsim B_{1d}\frac{K_d^p \log n}{\sqrt{\gamma n}} + B_{1d}^{1/2}\frac{(K_d^p \log n)^{3/4}}{\gamma^{1/2} n^{1/4}} + B_{1d}^{1/3}\frac{(K_d^p \log n)^{2/3}}{\gamma^{1/3} n^{1/6}}\right\}$$
$$\leq \gamma + n^{-1}\log n, \tag{7.29}$$

where $K_d^p = p + \log d$.

By (7.28) and (7.29) with $K_d^p = p + \log d$,

$$P\big\{|\hat{\Psi}_{\max} - \Psi^*| \gtrsim \Delta_{1n}(\gamma)\big\} \lesssim \Delta_{2n}(\gamma), \tag{7.30}$$

where

$$\Delta_{1n}(\gamma) = B_{2d}\frac{(K_d^p + \log n)^{1/2}}{\gamma\sqrt{n}} + B_{1d}\frac{K_d^p \log n}{\gamma^{1/2}\sqrt{n}} + B_{1d}^{1/2}\frac{(K_d^p \log n)^{3/4}}{\gamma^{1/2} n^{1/4}} + B_{1d}^{1/3}\frac{(K_d^p \log n)^{2/3}}{\gamma^{1/3} n^{1/6}}$$

and $\Delta_{2n}(\gamma) = \gamma + n^{-1}\log n$.

Step 3. Now we restrict attention to the Gaussian supremum Ψ^*. By Corollary 2.2.8 in [46] and (8.17), we get

$$E\Psi^* \lesssim \int_0^2 \sqrt{\sup_Q \log N(\mathcal{F}, L_2(Q), \varepsilon)}\, d\varepsilon \lesssim \sqrt{p + \log d}.$$

Combined with Corollary 2.1 in [13], this implies for every $\varepsilon \geq 0$ that

$$\sup_{t\geq 0} P\big(|\Psi^* - t| \geq \varepsilon\big) \lesssim \varepsilon\sqrt{p + \log d}. \tag{7.31}$$

Together, (7.30) and (7.31) yield, for every $t \geq 0$,

$$P(\hat{\Psi}_{\max} \leq t) \leq P\{\Psi^* \leq t + C\Delta_{1n}(\gamma)\} + C\Delta_{2n}(\gamma)$$

$$\leq P(\Psi^* \leq t) + C\{\Delta_{1n}(\gamma)\sqrt{p + \log d} + \Delta_{2n}(\gamma)\}.$$

A similar argument leads to the reverse inequality. Finally, in view of (4.7), taking

$$\gamma = \gamma_n(p, d)$$
$$= \max\left\{ B_{2d}^{1/2} \frac{(p + \log n)^{1/4}}{n^{1/4}}, B_{1d}^{1/4} (\log n)^{1/2} \frac{p^{7/8}}{n^{1/8}}, B_{1d}^{1/3} (\log n)^{1/2} \frac{p^{5/6}}{n^{1/6}}, B_{1d}^{2/3} (\log n)^{2/3} \frac{p}{n^{1/3}} \right\}$$

completes the proof under the assumption $d \leq \min\{n, m, \exp(C_0 p)\}$.

7.4.3. *Proof of Proposition* 7.3

Throughout the proof, $\{e_i\}_{i=1}^n$ is a sequence of i.i.d. standard normal random variables and P_e denotes the probability measure induced by $\{e_i\}_{i=1}^n$ holding $\{X_i\}_{i=1}^n$ fixed. For every $(u, k) \in \mathcal{T} = \mathcal{S}^{p-1} \times [d]$, by Taylor expansion we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n e_i \psi_k(\hat{U}_i^u) = \frac{1}{\sqrt{n}} \sum_{i=1}^n e_i \psi_k(U_i^u) + \frac{1}{n^{3/2}} \sum_{i=1}^n \sum_{j=1}^n \bar{h}_{u,k}(\bar{X}_i, \bar{X}_j) + \hat{R}_{u,k}, \quad (7.32)$$

where $\bar{X}_i = (e_i, X_i^\mathsf{T})^\mathsf{T} \in \mathbb{R}^{p+1}$,

$$\bar{h}_{u,k}(\bar{x}_1, \bar{x}_2) = e_1 \psi_k' \circ F^u(u^\mathsf{T} x_1)\big[I\{u^\mathsf{T}(x_2 - x_1) \leq 0\} - F^u(u^\mathsf{T} x_1)\big], \qquad \bar{x}_\ell = (e_\ell, x_\ell^\mathsf{T})^\mathsf{T} \in \mathbb{R}^{p+1}$$

for $\ell = 1, 2$ and the remainder $\hat{R}_{u,k}$ is such that

$$|\hat{R}_{u,k}| \leq \frac{1}{2} B_{2d} \sqrt{n} \max_{i \in [n]} |e_i| \times \sup_{(u,i) \in \mathcal{S}^{p-1} \times [n]} (\hat{U}_i^u - U_i^u)^2. \quad (7.33)$$

Because e_i and X_i are independent, we have $E\{\bar{h}_{u,k}(\bar{X}_1, \bar{X}_2)|\bar{X}_1\} = E\{\bar{h}_{u,k}(\bar{X}_1, \bar{X}_2)|\bar{X}_2\} = 0$ so that $\{\sum_{i=1}^n \sum_{j=1}^n \bar{h}_{u,k}(\bar{X}_i, \bar{X}_j)\}_{(u,k) \in \mathcal{T}}$ forms a degenerate U-process. With slight abuse of notation, we rewrite the function $\bar{h}_{u,k}$ as $\bar{h}_{u,k}(\bar{x}_1, \bar{x}_2) = e_1 \cdot w_{u,k}(x_1, x_2)$, where

$$w_{u,k}(x_1, x_2) = \psi_k'(F^u(u^\mathsf{T} x_1))\big[I\{u^\mathsf{T}(x_2 - x_1) \leq 0\} - F^u(u^\mathsf{T} x_1)\big].$$

In this notation, we have $\bar{\mathcal{H}}_d^p := \{\bar{h}_{u,k} : (u, k) \in \mathcal{T}\} \subseteq \{e \mapsto e\} \cdot \mathcal{W}_d^p$ with $\mathcal{W}_d^p = \{w_{u,k} : (u, k) \in \mathcal{T}\}$. Arguments similar to those employed in the proof of Lemma 8.4 can be used to prove that the collection $\mathcal{W}_d^p$ is VC-type, and so is $\bar{\mathcal{H}}_d^p$ with envelop $\bar{H}$ given by $\bar{H}(\bar{x}) = \bar{H}(e, x) = B_{2d}|e|$, such that

$$\sup_Q N(\bar{\mathcal{H}}_d^p, L_2(Q), \varepsilon \|\bar{H}\|_{Q,2}) \leq d \cdot (A/\varepsilon)^{vp} \quad (7.34)$$

for some constants $A > 2e$ and $v \geq 2$. This uniform entropy bound, together with Theorem 6 in [39] yields

$$
E\left\{\sup_{(u,k)\in\mathcal{T}}\left|\sum_{i=1}^{n}\sum_{j=1}^{n}\bar{h}_{u,k}(\bar{X}_i,\bar{X}_j)\right|\right\}
$$
$$
\lesssim B_{2d}n\left\{\frac{1}{4}+\int_0^{1/4}\sup_{Q}\sqrt{\log N\big(\bar{\mathcal{H}}_d^p, L_2(Q), \varepsilon\|\bar{H}\|_{Q,2}\big)}\,d\varepsilon\right\} \lesssim B_{2d}n\sqrt{p+\log d} \tag{7.35}
$$

by following the same lines as in the proof of Proposition 7.1.

For $\hat{R}_{u,k}$, applying the Borell-TIS inequality gives

$$
P\left\{\max_{i\in[n]}|e_i| \leq E\left(\max_{i\in[n]}|e_i|\right)+t\right\} \leq \exp\left(-t^2/2\right)
$$

for every $t > 0$. A standard result on Gaussian maxima is that $E(\max_{i\in[n]}|e_i|) \leq 2\sqrt{\log n}$. Consequently, combining Proposition 7.2 and (7.33) implies that

$$
\sup_{(u,k)\in\mathcal{T}}|\hat{R}_{u,k}| \lesssim B_{2d}(\log n)^{1/2}\sqrt{\frac{p+\log(dn)}{n}} \tag{7.36}
$$

holds with probability at least $1-3n^{-1}$,

By (7.32), (7.35) and (7.36), a similar argument to that leading to (7.28) gives, on this occasion that for any $\gamma \in (0,1)$,

$$
P\left\{\left|\hat{\Psi}_{\max}^{\mathrm{MB}}-\Psi_0^{\dagger}\right| \gtrsim B_{2d}\big(\sqrt{\log n}\vee\gamma^{-1}\big)\sqrt{\frac{p+\log(dn)}{n}}\right\} \lesssim \gamma + n^{-1}, \tag{7.37}
$$

where

$$
\Psi_0^{\dagger} = \sup_{(u,k)\in\mathcal{S}^{p-1}\times[d]}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}e_i\psi_k\big(U_i^u\big)\right| = \sup_{f\in\mathcal{F}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}e_i f(X_i)\right| \tag{7.38}
$$

for $\mathcal{F} = \mathcal{F}_d^p$ as in (7.20).

Notice that $\Psi_0^{\dagger}$ is the supremum of a (conditional) Gaussian process $\mathbb{G}^{\dagger}$ indexed by $\mathcal{F}$ with covariance function $E_e\{(\mathbb{G}^{\dagger}f_{u,k})(\mathbb{G}^{\dagger}f_{v,\ell}) = n^{-1}\sum_{i=1}^{n}\psi_k(F^u(u^\mathsf{T}X_i))\psi_\ell(F^u(u^\mathsf{T}X_i))$. Next we use an approximation due to [13]. Let $\mathcal{X}_n = \{X_1,\ldots,X_n\}$ be a realization of the data. Theorem A.2 there shows that for every $\delta > 0$, there exists a subset Ω_n such that $P(\mathcal{X}_n \in \Omega_n) \geq 1-3n^{-1}$ and for every $\mathcal{X}_n \in \Omega_n$, one can construct on an enriched probability space a random variable $\Psi^{\dagger}$ such that $\Psi^{\dagger} =_d \|\mathbb{G}\|_{\mathcal{F}}$ for $\mathbb{G}$ as in Lemma 7.3 and that

$$
P\left\{\left|\Psi_0^{\dagger}-\Psi^{\dagger}\right| \gtrsim \delta + \sqrt{\frac{K_d^p\log n}{n}} + B_{1d}^{1/2}\frac{(K_d^p\log n)^{3/4}}{n^{1/4}}\,\bigg|\,\mathcal{X}_n\right\}
$$
$$
\lesssim B_{1d}^{1/2}\frac{(K_d^p\log n)^{3/4}}{\delta n^{1/4}} + n^{-1}, \tag{7.39}
$$

where $K_d^p = p + \log d$.

Finally, combining (7.31) with inequalities (7.37) and (7.39), and setting

$$\gamma = B_{2d}^{1/2}\{(p+\log n)/n\}^{1/4} \quad \text{and} \quad \delta = B_{1d}^{1/4}(\log n)^{3/8}(p/n)^{1/8}$$

complete the proof of (7.6) in view of (4.7) and (7.31).

8. Proof of technical lemmas

We provide proofs here for all the technical lemmas. Throughout, we use C and c to denote universal positive constants, which may take different values at each occurrence.

Lemma 8.1. *Let $\{\xi_i, i \geq 1\}$ be a sequence of independent random variables with zero means and finite variances. Put $S_n = \sum_{i=1}^{n} \xi_i$, $v_n^2 = \sum_{i=1}^{n} \xi_i^2$ and $b_n^2 = \sum_{i=1}^{n} E(\xi_i^2)$, then for any $x > 0$,*

$$P\{|S_n| \geq x(v_n + 4b_n)\} \leq 4\exp(-x^2/2) \quad \text{and} \tag{8.1}$$

$$E[S_n^2 I\{|S_n| \geq x(v_n + 4b_n)\}] \leq 23b_n^2 \exp(-x^2/4). \tag{8.2}$$

Proof. The proof is based on Theorem 2.16 in [16] and Lemma 3.2 in [33]. $\quad\square$

Lemma 8.2. *Assume that the conditions of Proposition 7.1 hold, then for every $k \geq 1$ and $t > 0$,*

$$P(\sqrt{nm}|R_{2k}| \geq C_1 \|\psi_k'\|_\infty t) \leq C_2 \exp(-t/4), \tag{8.3}$$

where $C_1, C_2 > 0$ are absolute constants.

Proof. Without loss of generality, we only prove the result for $t \geq 4$, otherwise we can simply adjust the constant C_2 so that $C_2 \exp(-t/4) \geq 1$ for $0 \leq t \leq 4$. For given $k \geq 1$, define $Q_i = \sum_{j=1}^{m} q_{ij}$ with $q_{ij} = q_{ij,k} = h_{0k}(X_i, Y_j)$ for h_{0k} as in (7.10). Put $\mathcal{F}_Y = \sigma\{Y_1, \ldots, Y_m\}$, such that given $\mathcal{F}_Y$, $\{Q_i\}_{i=1}^{n}$ forms a sequence of independent random variables with zero (conditional) means. Noting that $\sum_{i=1}^{n}\sum_{j=1}^{m} q_{ij} = \sum_{i=1}^{n} Q_i$, it follows from a conditional version of (8.1) that for any $t \geq 4$,

$$P\left(\left|\sum_{i=1}^{n}\sum_{j=1}^{m} q_{ij}\right| \geq t\left[\left(\sum_{i=1}^{n} Q_i^2\right)^{1/2} + 4\left\{\sum_{i=1}^{n} E(Q_i^2|\mathcal{F}_Y)\right\}^{1/2}\right]\Big|\mathcal{F}_Y\right) \leq 4\exp(-t^2/2). \tag{8.4}$$

We study the tail behaviors of $\sum_{i=1}^{n} Q_i^2$ and $\sum_{i=1}^{n} E(Q_i^2|\mathcal{F}_Y)$ separately, starting with $\sum_{i=1}^{n} Q_i^2$. Observe that given X_i, Q_i is a sum of independent random variables with zero means. Put $V_i^2 = \sum_{j=1}^{m} q_{ij}^2$ and $B_i^2 = \sum_{j=1}^{m} E(q_{ij}^2|X_i)$. A direct consequence of (8.2) is that for every $t > 0$, $E[Q_i^2 I\{|Q_i| \geq t(V_i + 4B_i)\}|X_i] \leq 23B_i^2 \exp(-t^2/4)$. This implies by taking expectations on both sides that

$$E[Q_i^2 I\{|Q_i| \geq t(V_i + 4B_i)\}] \leq 23E(B_i^2)\exp(-t^2/4), \tag{8.5}$$

where $E(B_i^2) = \sum_{j=1}^m E(q_{ij}^2) \le mE\{h_k(X,Y)^2\} = m\sigma_k^2$ for σ_k^2 as in (7.9). Together, (8.5) and Lemma 7.2 in [44] imply, for $t \ge 4$,

$$P\left[\sum_{i=1}^n Q_i^2 \ge t^2\left\{nm\sigma_k^2 + \sum_{i=1}^n (V_i + 4B_i)^2\right\}\right] \le 92t^{-4}\exp(-t^2/4) \le (1/2)\exp(-t^2/4). \quad (8.6)$$

We consider next $\sum_{i=1}^n E(Q_i^2|\mathcal{F}_Y)$, which can be decomposed as

$$
\begin{aligned}
E&\left(Q_i^2|\mathcal{F}_Y\right) \\
&= E\left[Q_i^2 I\{|Q_i| \le t(V_i + 4B_i)\}|\mathcal{F}_Y\right] + E\left[Q_i^2 I\{|Q_i| > t(V_i + 4B_i)\}|\mathcal{F}_Y\right] \\
&\le t^2 E\left[(V_i + 4B_i)^2|\mathcal{F}_Y\right] + E\left[Q_i^2 I\{|Q_i| > t(V_i + 4B_i)\}|\mathcal{F}_Y\right] \\
&\le 17t^2 m\sigma_k^2 + 17t^2 \sum_{j=1}^m E\left(q_{ij}^2|Y_j\right) + E\left[Q_i^2 I\{|Q_i| > t(V_i + 4B_i)\}|\mathcal{F}_Y\right].
\end{aligned}
$$

Hence, it follows from Markov's inequality and (8.5) that

$$
\begin{aligned}
P&\left[\sum_{i=1}^n E\left(Q_i^2|\mathcal{F}_Y\right) \ge 18t^2\left\{nm\sigma_k^2 + \sum_{i=1}^n\sum_{j=1}^m E\left(q_{ij}^2|Y_j\right)\right\}\right] \\
&\le P\left(\sum_{i=1}^n E\left[Q_i^2 I\{|Q_i| > t(V_i + 4B_i)\}|\mathcal{F}_Y\right] \ge t^2 nm\sigma_k^2\right) \\
&\le t^{-2}\left(nm\sigma_k^2\right)^{-1}\sum_{i=1}^n E\left[Q_i^2 I\{Q_i^2 > t^2(V_i + 4B_i)^2\}\right] \\
&\le (3/2)\exp(-t^2/4).
\end{aligned}
\qquad (8.7)
$$

By (7.9), we have $\|h_{0k}\|_\infty \le 2b_k$. Then combining (8.4), (8.6) and (8.7) gives, for $t \ge 4$,

$$P\left\{\left|\frac{1}{\sqrt{nm}}\sum_{i=1}^n\sum_{j=1}^m q_{ij}\right| \ge C_1(\sigma_k + b_k)t\right\} \le 6\exp(-t/4).$$

This completes the proof of Lemma 8.2. $\qquad\square$

Lemma 8.3. *Assume that the conditions of Theorem* 4.2 *are fulfilled, then for all sufficiently large* n,

$$\max_{k\in[d]}|\vartheta_k - \theta_k| \lesssim B_{0d}^2 d^{2\tau}\frac{\log d}{n}, \qquad (8.8)$$

where $\vartheta_k := E_{H_1^d}\{\psi_k(V)\}$ *with* $V = F(Y)$.

Proof. Under the alternative H_1^d, the density of $V = F(Y)$ is of the form $\rho_\theta(z) = C_d(\theta) \times \exp\{\theta^\top \boldsymbol{\psi}(z)\}$, where $\{C_d(\theta)\}^{-1} = \int_0^1 \exp\{\theta^\top \boldsymbol{\psi}(z)\} \, dz$ and $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_d)^\top$. In this notation, we have $\vartheta_k = C_d(\theta) \int_0^1 \psi_k(z) \exp\{\theta^\top \boldsymbol{\psi}(z)\} \, dz$. Note that

$$\left| \theta^\top \boldsymbol{\psi}(z) \right| = \left| \sum_{k=1}^d \theta_k \psi_k(z) \right| \leq B_{0d} d^\tau \max_{k \in [d]} |\theta_k| = \lambda B_{0d} d^\tau \sqrt{\frac{\log d}{\bar{n}}}.$$

Consequently, using the inequality $|e^t - 1 - t| \leq \frac{1}{2} t^2 \exp(t \vee 0)$ which holds for every $t \in \mathbb{R}$ to $t = |\theta^\top \boldsymbol{\psi}(z)|$ yields

$$\int_0^1 \psi_k(z) \exp\{\theta^\top \boldsymbol{\psi}(z)\} \, dz = \int_0^1 \psi_k(z) \{1 + \theta^\top \boldsymbol{\psi}(z)\} \, dz + O(1) \int_0^1 |\psi_k(z)| \{\theta^\top \boldsymbol{\psi}(z)\}^2 \, dz$$

$$= \sum_{\ell=1}^d \theta_\ell \int_0^1 \psi_k(z) \psi_\ell(z) \, dz + O(1) B_{0d}^2 d^{2\tau} \frac{\log d}{n}$$

$$= \theta_k + O(1) B_{0d}^2 d^{2\tau} \frac{\log d}{n}$$

uniformly over $k \in [d]$. Similarly, it can be proved that

$$\left| \int_0^1 \exp\{\theta^\top \boldsymbol{\psi}(z)\} \, dz - 1 \right| \lesssim B_{0d}^2 d^{2\tau} \frac{\log d}{n},$$

which implies $C_d(\theta) = 1 + o(1)$ as $d, n \to \infty$. Combining the above calculations proves (8.8). $\qquad \square$

Lemma 8.4. *Under the null hypothesis $H_0 : F = G$, the class $\mathcal{H}_0$ of degenerate kernels $\mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbb{R}$, to which an envelop $\equiv 2B_{2d}$ is attached, is VC-type; that is, there are constants $A > 2e$ and $v \geq 2$ such that*

$$\sup_{Q \text{ discrete}} N\big(\mathcal{H}_0, L_2(Q), 2\varepsilon B_{2d}\big) \leq d \cdot (A/\varepsilon)^{vp}, \tag{8.9}$$

where the supremum ranges over all finitely discrete Borel probability measures on $\mathbb{R}^p \times \mathbb{R}^p$.

Proof. First, we prove that the class $\mathcal{H}$ of kernels is VC-type. Note that $\mathcal{H}$ has envelop $\equiv B_{2d}$ and admits the partition $\mathcal{H} = \bigcup_{k=1}^d \mathcal{H}_k$, where for each $k \in [d]$, the class $\mathcal{H}_k = \{h_{u,k} \in \mathcal{H} : u \in \mathcal{S}^{p-1}\}$ has an envelop $\leq B_{2d}$. This implies

$$\sup_{Q \text{ discrete}} N\big(\mathcal{H}, L_2(Q), \varepsilon B_{2d}\big) \leq \sum_{k=1}^d \sup_{Q \text{ discrete}} N\big(\mathcal{H}_k, L_2(Q), \varepsilon B_{2d}\big), \tag{8.10}$$

where the supremum ranges over all finitely discrete Borel probability measures on $\mathbb{S}_2 := \mathbb{R}^p \times \mathbb{R}^p$. Therefore, it suffices to restrict attention to the class $\mathcal{H}_k$ with a fixed $k \in [d]$. For every

$u \in \mathcal{S}^{p-1}$, observe that $h_{u,k}(x, y) = \psi'_k \circ f_u(y) \cdot I_u(x, y)$. Regarding each element of $\psi'_k(\mathcal{F}^p) :=$ $\{y \mapsto \psi'_k \circ f_u(y) : f_u \in \mathcal{F}^p\}$ as a measurable function on $\mathbb{S}_2$, i.e. $(x, y) \mapsto \psi'_k \circ f_u(y)$, we have $\mathcal{H}_k \subset \psi'_k(\mathcal{F}^p) \cdot \mathcal{I}^p$ for $\mathcal{F}^p$ and $\mathcal{I}^p$ given in (7.21). Since both the classes $\mathcal{F}^p$ and $\mathcal{I}^p$ have envelop $\equiv 1$ and the function ψ'_k is Lipschitz continuous, it follows from Lemma A.6 and Corollary A.1 in [12] that, for any $0 < \varepsilon \le 1$,

$$\sup_{Q \text{ discrete}} N\big(\psi'_k(\mathcal{F}^p), L_2(Q), \varepsilon B_{2d}\big) \le \sup_{Q \text{ discrete}} N\big(\mathcal{F}^p, L_2(Q), \varepsilon\big) \tag{8.11}$$

and

$$\sup_{Q \text{ discrete}} N\big(\mathcal{H}_k, L_2(Q), 2\varepsilon B_{2d}\big)$$
$$\le \sup_{Q \text{ discrete}} N\big(\psi'_k(\mathcal{F}^p), L_2(Q), \varepsilon B_{2d}\big) \sup_{Q \text{ discrete}} N\big(\mathcal{I}^p, L_2(Q), \varepsilon\big), \tag{8.12}$$

where the suprema appeared above are taken over all finitely discrete Borel probability measures on $\mathbb{S}_2$.

In view of (8.11) and (8.12), it remains to focus on the classes $\mathcal{F}^p$ and $\mathcal{I}^p$. Arguments similar to those in [45] can be used to control the entropies of $\mathcal{I}^p$. To see this, define $\mathcal{V} = \{v(\cdot, \cdot, \cdot; u) : u \in \mathbb{R}^p\}$ and $\mathcal{W} = \{w(\cdot, \cdot, \cdot; \gamma) : \gamma \in \mathbb{R}\}$, where $v(x, y, t; u) = u^{\mathsf{T}} x - u^{\mathsf{T}} y$ and $w(x, y, t; \gamma) = \gamma t$ for $x, y \in \mathbb{R}^p$ and $t \in \mathbb{R}$. Note that $\mathcal{V}$ (resp. $\mathcal{W}$) is a p-dimensional (resp. 1-dimensional) vector space of real-valued functions on $\mathbb{S}_2 \times \mathbb{R}$. By Theorem 4.6 in [18], the class of sets of the form $\{z : v(z) > s\}$ or $\{z : v(z) \ge s\}$ with $v \in \mathcal{V}$ for some $s \in \mathbb{R}$ fixed is a VC class with index $p + 1$. For every $u \in \mathcal{S}^{p-1}$,

$$\begin{aligned}
\text{graph}(I_u) &= \big\{(x, y, t) \in \mathbb{S}_2 \times \mathbb{R} : 0 < t < I_u(x, y)\big\} \\
&= \big\{u^{\mathsf{T}} x - u^{\mathsf{T}} y \le 0\big\} \cap \{t > 0\} \cap \{t \ge 1\}^{\mathrm{c}} \\
&= \{v_1 > 0\}^{\mathrm{c}} \cap \{w_1 > 0\} \cap \{w_2 \ge 1\}^{\mathrm{c}},
\end{aligned}$$

where $v_1 \in \mathcal{V}$ and $w_1, w_2 \in \mathcal{W}$. Together with Lemma 9.7 in [32], this implies that $\{\text{graph}(I_u) : I_u \in \mathcal{I}\}$ forms a VC class with index $\le p + 3$. Consequently, by Theorem 9.3 in [32], there exist constants $a > 2e$ and $c \ge 2$ such that

$$\sup_Q N\big(\mathcal{I}^p, L_2(Q), \varepsilon\big) \le (a/\varepsilon)^{cp} \tag{8.13}$$

for any $0 < \varepsilon \le 1$, where the supremum is taken over all Borel probability measures on $\mathbb{S}_2$. For $\mathcal{F}^p$, applying Lemma A.2 in [23] combined with (8.13) gives

$$\sup_Q N\big(\mathcal{F}^p, L_2(Q), 2\varepsilon\big) \le \sup_Q N\big(\mathcal{I}^p, L_2(P_X \times Q), \varepsilon^2\big) \le (\sqrt{a}/\varepsilon)^{2cp}, \tag{8.14}$$

where the supremum ranges over all Borel probability measures on $\mathbb{R}^p$.

Together, (8.10)–(8.14) imply the VC-type property of the class $\mathcal{H}$.

We consider next the class $\mathcal{H}_0$ of degenerate kernels under H_0, which admits a partition similar to (8.10), i.e. $\mathcal{H}_0 = \bigcup_{k=1}^{d} \mathcal{H}_{0k}$. Observe that for each $(u, k) \in \mathcal{T}$,

$$h_{u,k,0}(x, y) = h_{u,k}(x, y) + \psi_k \circ f^u(x) + \phi_k \circ f^u(y), \qquad x, y \in \mathbb{R}^p.$$

where $h_{u,k} \in \mathcal{H}_k \subset \mathcal{H}$, $f^u \in \mathcal{F}$ and $\phi_k(s) := -s\psi_k'(s)$ for $0 \le s \le 1$. For any $u, v \in \mathcal{S}^{p-1}$ and $k \in [d]$, we have $|\phi_k \circ f_u(y) - \phi_k \circ f_v(y)| \le 2B_{2d}|f_u(x) - f_v(y)|$. This, together with Lemma A.6 in [12] yields

$$\sup_{Q \text{ discrete}} N\big(\phi_k\big(\mathcal{F}^p\big), L_2(Q), 2\varepsilon B_{2d}\big) \le \sup_{Q \text{ discrete}} N\big(\mathcal{F}^p, L_2(Q), \varepsilon\big). \tag{8.15}$$

On combing (8.11), (8.12) and (8.15), and recalling the permanence of the uniform entropy bound under summation that is implied by Lemma A.6 in [12], we obtain

$$\sup_{Q \text{ discrete}} N\big(\mathcal{H}_{0k}, L_2(Q), 3\varepsilon B_{2d}\big)$$

$$\le \sup_{Q \text{ discrete}} N\big(\mathcal{H}_k, L_2(Q), \varepsilon B_{2d}\big)$$

$$\times \sup_{Q \text{ discrete}} N\big(\psi_k\big(\mathcal{F}^p\big), L_2(Q), \varepsilon B_{2d}\big) \sup_{Q \text{ discrete}} N\big(\phi_k\big(\mathcal{F}^p\big), L_2(Q), \varepsilon B_{2d}\big)$$

$$\le \sup_{Q \text{ discrete}} N\big(\mathcal{I}^p, L_2(Q), \varepsilon/2\big)\Big\{ \sup_{Q \text{ discrete}} N\big(\mathcal{F}^p, L_2(Q), \varepsilon/2\big)\Big\}^3.$$

This completes the proof of (8.9) in view of (8.13) and (8.14). $\qquad\square$

8.1. Proof of Lemma 7.1

Observe that $\{U_{m,n}(h_0)\}_{h_0 \in \mathcal{H}_0}$ forms a degenerate two-sample U-process indexed by $\mathcal{H}_0$ and by Lemma 8.4, $\mathcal{H}_0$ is VC-type with envelop $\equiv 2B_{2d}$. The entropy bound given in (8.9) now allows us to apply Lemma 2.4 in [37], yielding

$$E\big(\|U_{n,m}\|_{\mathcal{H}_0}\big) \lesssim B_{2d}\sqrt{nm}\left\{ \frac{1}{4} + \int_0^{1/4} \sup_{Q \text{ discrete}} \sqrt{\log N\big(\mathcal{H}_0, L_2(Q), 2\varepsilon B_{2d}\big)}\, d\varepsilon \right\}$$

$$\lesssim B_{2d}\sqrt{nm}\left\{ \frac{1}{4} + \int_0^{1/4} \sqrt{\log d + vp\log(A/\varepsilon)}\, d\varepsilon \right\} \tag{8.16}$$

$$\lesssim B_{2d}\sqrt{nm}\left\{ \frac{1}{4}(1 + \sqrt{\log d}) + \sqrt{vp}\int_{4A}^{\infty} t^{-2}\sqrt{\log t}\, dt \right\}.$$

For any $a > e$, it follows from integration by parts that

$$\int_a^{\infty} t^{-2}\sqrt{\log t}\, dt$$

$$= a^{-1}\sqrt{\log a} + \frac{1}{2}\int_a^{\infty} t^{-2}(\log t)^{-1/2}\, dt$$

$$\leq a^{-1}\sqrt{\log a} + \frac{1}{2\log a}\int_a^\infty t^{-2}\sqrt{\log t}\,dt$$

$$\leq a^{-1}\sqrt{\log a} + \frac{1}{2}\int_a^\infty t^{-2}\sqrt{\log t}\,dt.$$

Substituting this into (8.16) proves (7.25).

8.2. Proof of Lemma 7.2

Define the class $\mathcal{G} = \{x \mapsto g_{u,t}(x) = I(u^\mathsf{T} x \leq t) : (u,t) \in \mathcal{S}^{p-1} \times \mathbb{R}\}$ of indicator functions on closed half-spaces in $\mathbb{R}^d$, such that

$$D_n(\mathcal{G}) := \sup_{(u,t)\in\mathcal{S}^{p-1}\times\mathbb{R}} \left|F_n^u(t) - F^u(t)\right| = \sup_{g\in\mathcal{G}} \left|\frac{1}{n}\sum_{i=1}^n \{g(X_i) - P_X g\}\right|,$$

where $P_X g := E\{g(X)\}$. Note that, for every $(u,t) \in \mathcal{T}$, $\mathrm{var}\{g_{u,t}(X)\} = F^u(t)\{1 - F^u(t)\} \leq 1/4$. A direct consequence of Theorem 7.3 in [7] is that, for every $t \geq 0$,

$$P\left(D_n(\mathcal{G}) \geq E\{D_n(\mathcal{G})\} + \left[\frac{1}{2} + 4E\{D_n(\mathcal{G})\}\right]^{1/2}\sqrt{\frac{t}{n}} + \frac{t}{3n}\right) \leq 2e^{-t}.$$

To control the expectation $E\{D_n(\mathcal{G})\}$, first it follows from Theorem B in [17] that the class $\mathcal{G}$ is a VC-subgraph class with index $p+2$, such that for any probability measure Q on $\mathbb{R}^p$ and any $0 < \varepsilon < 1$, $N(\mathcal{G}, L_2(Q), \varepsilon) \leq (C\varepsilon^{-1})^{2(p+1)}$. This, together with Proposition 3 in [24] gives

$$E\{D_n(\mathcal{G})\} \lesssim \sqrt{\frac{p}{n}} + \frac{p}{n}.$$

Since $D_n(\mathcal{G}) \leq 1$, the last three displays together complete the proof of (7.26).

8.3. Proof of Lemma 7.3

To prove (7.29), a new coupling inequality for the suprema of empirical processes in [12] plays an important role in our analysis. Recall in the proof of Lemma 8.4 that the collection $\mathcal{F}^p$ is VC-type, from which we obtain

$$\sup_{Q \text{ discrete}} N\big(\mathcal{F}, L_2(Q), \varepsilon B_{1d}\big) \leq \sum_{k=1}^d \sup_{Q \text{ discrete}} N\big(\psi_k(\mathcal{F}^p), L_2(Q), \varepsilon B_{1d}\big) \leq d \cdot (A/\varepsilon)^{\upsilon p} \quad (8.17)$$

for some constants $A > 2e$ and $\upsilon \geq 2$, where $\mathcal{F} = \mathcal{F}_d^p$. This implies by Lemma 2.1 in [12] that the collection $\mathcal{F}$ is a VC-type pre-Gaussian class with a constant envelop $\equiv B_{1d}$. Therefore,

there exists a centered, tight Gaussian process $\mathbb{G}$ defined on $\mathcal{F}$ with covariance function (7.24). Moreover, for any integer $k \geq 2$,

$$\sup_{f \in \mathcal{F}} P_X |f|^k \leq B_{0d}^{k-2} \sup_{(u,k) \in \mathcal{S}^{p-1} \times [d]} E\{\psi_k(U^u)^2\} = B_{0d}^{k-2}, \qquad (8.18)$$

where we used the fact that $U^u = F^u(u^\intercal X) =_d \text{Unif}(0, 1)$ and hence $E\{\psi_k(U^u)^2\} = 1$ for all $(u, k) \in \mathcal{S}^{p-1} \times [d]$.

The entropy bound (8.17) and the moment inequality (8.18) now allow to apply Corollary 2.2 in [12], yielding (7.29).

Acknowledgements

References

[1] Baringhaus, L. and Franz, C. (2004). On a new multivariate two-sample test. *J. Multivariate Anal.* **88** 190–206. MR2021870

[2] Baringhaus, L. and Franz, C. (2010). Rigid motion invariant two-sample tests. *Statist. Sinica* **20** 1333–1361. MR2777328

[3] Barrett, G.F. and Donald, S.G. (2003). Consistent tests for stochastic dominance. *Econometrica* **71** 71–104. MR1956856

[4] Bera, A.K. and Ghosh, A. (2002). Neyman's smooth test and its applications in econometrics. In *Handbook of Applied Econometrics and Statistical Inference* (A. Ullah, A.T.K. Wan and A. Chaturvedi, eds.). *Statist. Textbooks Monogr.* **165** 177–230. New York: Dekker. MR1893336

[5] Bera, A.K., Ghosh, A. and Xiao, Z. (2013). A smooth test for the equality of distributions. *Econometric Theory* **29** 419–446. MR3042761

[6] Biswas, M. and Ghosh, A.K. (2014). A nonparametric two-sample test applicable to high dimensional data. *J. Multivariate Anal.* **123** 160–171. MR3130427

[7] Bousquet, O. (2003). Concentration inequalities for sub-additive functions using the entropy method. In *Stochastic Inequalities and Applications*. *Progress in Probability* **56** 213–247. Basel: Birkhäuser. MR2073435

[8] Cai, T.T., Liu, W. and Xia, Y. (2014). Two-sample test of high dimensional means under dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 349–372. MR3164870

[9] Cattaneo, M.D. and Farrell, M.H. (2013). Optimal convergence rates, Bahadur representation, and asymptotic normality of partitioning estimators. *J. Econometrics* **174** 127–143. MR3045024

[10] Chang, J., Zhou, W. and Zhou, W.-X. (2014). Simulation-based hypothesis testing of high dimensional means under covariance heterogeneity. Available at arXiv:1406.1939.

[11] Chernozhukov, V., Chetverikov, D. and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41** 2786–2819. MR3161448

[12] Chernozhukov, V., Chetverikov, D. and Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *Ann. Statist.* **42** 1564–1597. MR3262461

[13] Chernozhukov, V., Chetverikov, D. and Kato, K. (2014). Anti-concentration and honest, adaptive confidence bands. *Ann. Statist.* **42** 1787–1818. MR3262468

[14] Darling, D.A. (1957). The Kolmogorov–Smirnov, Cramér–von Mises tests. *Ann. Math. Statist* **28** 823–838. MR0093870

[15] de Boor, C. (1978). *A Practical Guide to Splines*. *Applied Mathematical Sciences* **27**. New York: Springer. MR0507062

[16] de la Peña, V.H., Lai, T.L. and Shao, Q.-M. (2009). *Self-Normalized Processes*: *Limit Theory and Statistical Applications*. *Probability and Its Applications* (*New York*). Berlin: Springer. MR2488094

[17] Dudley, R.M. (1979). Balls in $\mathbf{R}^k$ do not cut all subsets of $k + 2$ points. *Adv. in Math.* **31** 306–308. MR0532837

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

[19] Escanciano, J.C. (2009). On the lack of power of omnibus specification tests. *Econometric Theory* **25** 162–194. MR2472049

[20] Eubank, R.L. and LaRiccia, V.N. (1992). Asymptotic comparison of Cramér–von Mises and nonparametric function estimation techniques for testing goodness-of-fit. *Ann. Statist.* **20** 2071–2086. MR1193326

[21] Fan, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation. *J. Amer. Statist. Assoc.* **91** 674–688. MR1395735

[22] Friedman, J.H. and Rafsky, L.C. (1979). Multivariate generalizations of the Wald–Wolfowitz and Smirnov two-sample tests. *Ann. Statist.* **7** 697–717. MR0532236

[23] Ghosal, S., Sen, A. and van der Vaart, A.W. (2000). Testing monotonicity of regression. *Ann. Statist.* **28** 1054–1082. MR1810919

[24] Giné, E. and Nickl, R. (2009). An exponential inequality for the distribution function of the kernel density estimator, with applications to adaptive estimation. *Probab. Theory Related Fields* **143** 569–596. MR2475673

[25] Hall, P. and Tajvidi, N. (2002). Permutation tests for equality of distributions in high-dimensional settings. *Biometrika* **89** 359–374. MR1913964

[26] Hansen, B.E. (1996). Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica* **64** 413–430. MR1375740

[27] Henze, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *Ann. Statist.* **16** 772–783. MR0947577

[28] Heyde, C.C. (1963). On a property of the lognormal distribution. *J. Roy. Statist. Soc. Ser. B* **25** 392–393. MR0171336

[29] Horowitz, J.L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica* **60** 505–531. MR1162997

[30] Inglot, T., Kallenberg, W.C.M. and Ledwina, T. (1997). Data driven smooth tests for composite hypotheses. *Ann. Statist.* **25** 1222–1250. MR1447749

[31] Janssen, A. (2000). Global power functions of goodness of fit tests. *Ann. Statist.* **28** 239–253. MR1762910

[32] Kosorok, M.R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. *Springer Series in Statistics*. New York: Springer. MR2724368

[33] Lai, T.L., Shao, Q.-M. and Wang, Q. (2011). Cramér type moderate deviations for Studentized U-statistics. *ESAIM Probab. Stat.* **15** 168–179. MR2870510

[34] Lehmann, E.L. and Romano, J.P. (2005). *Testing Statistical Hypotheses*, 3rd ed. *Springer Texts in Statistics*. New York: Springer. MR2135927

[35] Mallat, S. (1998). *A Wavelet Tour of Signal Processing*. San Diego, CA: Academic Press. MR1614527

[36] Massart, P. (1990). The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *Ann. Probab.* **18** 1269–1283. MR1062069

[37] Neumeyer, N. (2004). A central limit theorem for two-sample U-processes. *Statist. Probab. Lett.* **67** 73–85. MR2039935

[38] Neyman, J. (1937). Smooth test for goodness of fit. *Skand. Aktuarietidskr.* **20** 150–199.

[39] Nolan, D. and Pollard, D. (1987). U-processes: Rates of convergence. *Ann. Statist.* **15** 780–799. MR0888439

[40] Rosenbaum, P.R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 515–530. MR2168202

[41] Sansone, G. (1959). *Orthogonal Functions*. New York: Interscience. MR0103368

[42] Schilling, M.F. (1986). Multivariate two-sample tests based on nearest neighbors. *J. Amer. Statist. Assoc.* **81** 799–806. MR0860514

[43] Shadrin, A. (1992). Interpolation with Lagrange polynomials. A simple proof of Markov inequality and some of its generalizations. *Approx. Theory Appl.* **8** 51–61. MR1195174

[44] Shao, Q.-M. and Zhou, W.-X. (2016). Cramér type moderate derivation theorems for self-normalized processes. *Bernoulli* **22** 2029–2079.

[45] Sherman, R.P. (1994). Maximal inequalities for degenerate U-processes with applications to optimization estimators. *Ann. Statist.* **22** 439–459. MR1272092

[46] 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

[47] Weiss, L. (1960). Two-sample tests for multivariate distributions. *Ann. Math. Statist.* **31** 159–164. MR0119305