Information and Inference: A Journal of the IMA (2024) **13**, iaae017
https://doi.org/10.1093/imaiai/iaae017

Generalized kernel distance covariance in high dimensions: non-null CLTs and power universality

QIYANG HAN
Department of Statistics, Rutgers University, Piscataway, NJ 08854, USA

AND

YANDI SHEN*
Department of Statistics, University of Chicago, Chicago, IL 60637, USA
*Corresponding author: Email: ydshen@uchicago.edu

[Received on 10 January 2024; accepted on 26 June 2024]

Distance covariance is a popular dependence measure for two random vectors X and Y of possibly different dimensions and types. Recent years have witnessed concentrated efforts in the literature to understand the distributional properties of the sample distance covariance in a high-dimensional setting, with an exclusive emphasis on the null case that X and Y are independent. This paper derives the first non-null central limit theorem for the sample distance covariance, and the more general sample (Hilbert–Schmidt) kernel distance covariance in high dimensions, in the distributional class of (X, Y) with a separable covariance structure. The new non-null central limit theorem yields an asymptotically exact first-order power formula for the widely used generalized kernel distance correlation test of independence between X and Y. The power formula in particular unveils an interesting universality phenomenon: the power of the generalized kernel distance correlation test is completely determined by $n \cdot \mathrm{dCor}^2(X, Y)/\sqrt{2}$ in the high-dimensional limit, regardless of a wide range of choices of the kernels and bandwidth parameters. Furthermore, this separation rate is also shown to be optimal in a minimax sense. The key step in the proof of the non-null central limit theorem is a precise expansion of the mean and variance of the sample distance covariance in high dimensions, which shows, among other things, that the non-null Gaussian approximation of the sample distance covariance involves a rather subtle interplay between the dimension-to-sample ratio and the dependence between X and Y.

Keywords: central limit theorem; distance covariance; independent test; non-null analysis; power analysis; Poincaré inequalities.

1. Introduction

1.1 *Overview*

Given samples from a random vector (X, Y) in $\mathbb{R}^{p+q}$, it is of fundamental statistical interest to test whether X and Y are independent. The long history of this problem has given rise to a large number of dependence measures targeting at different types of dependence structure. Notable examples include the classical Pearson correlation coefficient [32], rank-based correlation coefficients [5,15,21,25,35,40,49], Cramér-von Mises-type measures [47], measure based on characteristic functions [41,44], kernel-based measures [18,19] and sign covariance [4,46]. We refer to the classical textbooks [1, Chapter 9] and [31, Chapter 11] for a systematic exposition on this topic.

Among the plentiful dependence measures for such a purpose, the distance covariance metric and its generalizations [43,44] have attracted much attention in recent years. In one of its many equivalent

forms, the distance covariance between X and Y can be defined as (cf. [41, Theorems 7 and 8])

$$\begin{aligned}\mathrm{dCov}^2(X,Y) \equiv{}& \mathbb{E}\big(\|X_1 - X_2\|\|Y_1 - Y_2\|\big) - 2\mathbb{E}\big(\|X_1 - X_2\|\|Y_1 - Y_3\|\big) \\ &+ \mathbb{E}\big(\|X_1 - X_2\|\big)\mathbb{E}\big(\|Y_1 - Y_2\|\big). \end{aligned} \tag{1.1}$$

Here (X_i, Y_i), $i = 1, 2, 3$ are independent copies from the joint distribution of (X, Y), and $\|\cdot\|$ is the Euclidean norm. The distance covariance metric $\mathrm{dCov}^2(X, Y)$ is particularly appealing for several nice features. First, X and Y are independent if and only if $\mathrm{dCov}^2(X, Y) = 0$. Second, $\mathrm{dCov}^2(X, Y)$ can be used in cases where X and Y are of different dimensions and data type (discrete, continuous or mixed). Third, several estimators of $\mathrm{dCov}^2(X, Y)$ are known to allow for efficient calculation. Due to these reasons, the distance covariance has been utilized in a wide range of both methodological and applied contexts, see e.g. [27,28,30,39,50,51] for an incomplete list of references.

An estimator of $\mathrm{dCov}^2(X, Y)$ based on n i.i.d. samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ from the distribution of (X, Y) is first proposed in [44], with its bias-corrected version proposed in [43], which is now known as the sample distance covariance $\mathrm{dCov}^2_*(\boldsymbol{X}, \boldsymbol{Y})$. The finite-sample distribution of $\mathrm{dCov}^2_*(\boldsymbol{X}, \boldsymbol{Y})$ is generally intractable; so the literature has focused on deriving its asymptotic distribution in different growth regimes of (n, p, q), cf. [17,22,42,52]. In the fixed-dimensional asymptotic regime when p, q are fixed and n diverges to infinity, [22] showed that $\mathrm{dCov}^2_*(\boldsymbol{X}, \boldsymbol{Y})$ converges in distribution to a mixture of chi-squared distributions. This is complemented by the result of [42], where a t-distribution limit was derived in the so-called 'high-dimensional low sample size' regime when n remains fixed and both p, q diverge to infinity. The high-dimensional regime where both the sample size n and the data dimension p, q diverge was recently studied in [17,52], where $\mathrm{dCov}^2_*(\boldsymbol{X}, \boldsymbol{Y})$ was shown to obey a central limit theorem (CLT). We also refer to [16,48] for some related distributional results in the problem of two-sample distribution testing. Except for some non-null results in [52, Proposition 2.2.2] under the fixed n regime, all these results are derived under the null scenario where X and Y are independent. This leaves open the more challenging but equally important issue of non-null limiting distributions of $\mathrm{dCov}^2_*(\boldsymbol{X}, \boldsymbol{Y})$, which are the key to a complete characterization of the power behaviour of distance covariance based tests. Bridging this significant theoretical gap is one of the main motivations of this paper.

1.2 *Non-null CLTs*

For the majority of the paper, we work with distributions of (X, Y) with a separable covariance structure (see Section 2.2 ahead for details), and perform an exact analysis of the distributional properties of the sample distance covariance $\mathrm{dCov}^2_*(\boldsymbol{X}, \boldsymbol{Y})$ in the following high-dimensional regime:

$$\min\{n, p, q\} \to \infty. \tag{1.2}$$

Our first main result is the following non-null CLT (see Theorem 2.2 below for the formal statement): Uniformly over the covariance matrix Σ of $(X^\top, Y^\top)^\top$ with a compact spectrum in $(0, \infty)$,

$$\frac{\mathrm{dCov}^2_*(\boldsymbol{X}, \boldsymbol{Y}) - \mathrm{dCov}^2(X, Y)}{\mathrm{Var}^{1/2}\big(\mathrm{dCov}^2_*(\boldsymbol{X}, \boldsymbol{Y})\big)} \text{ converges in distribution to } \mathcal{N}(0, 1) \tag{1.3}$$

in the regime (1.2). Here $\mathcal{N}(0,1)$ denotes the standard normal distribution. For simplicity, we have presented here the asymptotic version of the result; the more complete Theorem 2.2 below is non-asymptotic in nature and gives an error bound with explicit dependence on the problem parameters (n,p,q). Furthermore, we show that an analogue of (1.3) also holds for $\mathrm{dCov}_*^2(\boldsymbol{X})$, the 'marginal' analogous unbiased estimator of $\mathrm{dCov}^2(X)$. To the best of our knowledge, (1.3) is the first non-null CLT for $\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y})$ in the literature.

Let us now give some intuition why one would expect a non-null CLT (1.3) that holds for a general Σ in the regime (1.2). It is well known that the sample distance covariance $\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y})$ admits a U-statistics representation (cf. Proposition 2.1) with first-order degeneracy under the null. By classical theory in the fixed-dimensional asymptotics (i.e. p,q fixed with $n\to\infty$), a CLT holds for $\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y})$ under any fixed alternative $\Sigma \neq I_{p+q}$, while a non-Gaussian limit holds under the null $\Sigma = I_{p+q}$. In such fixed-dimensional asymptotics, the Gaussian limit is due to the *non-degeneracy* of $\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y})$ under the alternative, while the non-Gaussian limit is due to the *degeneracy* of $\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y})$ under the null. Now, as high dimensionality also enforces a Gaussian approximation of $\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y})$ under the null with degeneracy (cf. [17,52]), one would naturally expect the finite-sample distribution of the centred $\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y})$ under a general Σ, to be approximately a 'mixture' of a centred Gaussian component due to non-degeneracy and another centred Gaussian component due to degeneracy, which is again Gaussian. The non-null CLT (1.3) formalizes this intuition in the regime (1.2).

To formally implement the above intuition, an important step in the proof of (1.3) is to obtain precise mean and variance expansions for the sample distance covariance $\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y})$ in the regime (1.2). In particular, we show in Theorem 8.4 that the mean can be expanded as

$$\mathrm{dCov}^2(X,Y) = \frac{\|\Sigma_{XY}\|_F^2}{2\sqrt{\mathrm{tr}(\Sigma_X)\,\mathrm{tr}(\Sigma_Y)}}\big(1+\mathfrak{o}(1)\big), \tag{1.4}$$

and in Theorem 9.12 that the variance *under the null* σ_{null}^2 can be expanded as

$$\sigma_{\mathrm{null}}^2 = \frac{\|\Sigma_X\|_F^2\|\Sigma_Y\|_F^2}{2n(n-1)\,\mathrm{tr}(\Sigma_X)\,\mathrm{tr}(\Sigma_Y)}\big(1+\mathfrak{o}(1)\big). \tag{1.5}$$

Here Σ_X, Σ_Y and Σ_{XY} are sub-blocks of the covariance matrix $\Sigma = [\Sigma_X, \Sigma_{XY}; \Sigma_{YX}, \Sigma_Y]$; $\mathrm{tr}(\cdot)$ denotes the trace and $\|\cdot\|_F$ denotes the matrix Frobenious norm, and $\mathfrak{o}(1)$ is the standard small-o notation representing a vanishing term under the asymptotics (1.2). The variance formula for general Σ (explicit form see Theorem 9.12) is rather complicated so is not presented here, but as explained above, it is expected to contain two parts that are contributed individually by the non-degenerate and the degenerate components of $\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y})$. Notably, the contributions of these two components to the Gaussian approximation in (1.3) depend on the dimension-to-sample ratio in a fairly subtle way. In 'very high dimensions', the non-null CLT is entirely driven by the degeneracy of $\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y})$ regardless of the degree of dependence between X and Y. On the other hand, in 'moderate high dimensions', dependence between X and Y plays a critical role in determining the contributions of the (non-)degeneracy in the non-null CLT. See the discussion after Theorem 2.2 for details.

1.3 *Independent test via distance covariance: power asymptotics*

A major application of the non-null CLT derived in (1.3) is a precise power formula for the following popular distance correlation test of independence between X and Y, first considered in [42]:

$$\Psi(\boldsymbol{X},\boldsymbol{Y};\alpha) \equiv \mathbf{1}\bigg(\bigg|\frac{n\cdot \mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y})}{\sqrt{2\,\mathrm{dCov}_*^2(\boldsymbol{X})\cdot \mathrm{dCov}_*^2(\boldsymbol{Y})}}\bigg| > z_{\alpha/2}\bigg). \tag{1.6}$$

The above independence test and the null part of (1.3) is connected by the mean and variance expansions in (1.4) and (1.5). In fact, as will be detailed in Section 3, the above test is asymptotically (in the regime (1.2)) equivalent to the (infeasible) z-test built from the null part of (1.3). As a direct consequence, $\Psi(\boldsymbol{X},\boldsymbol{Y};\alpha)$ will also have an asymptotic size of α. The null behaviour of (a variant of) $\Psi(\boldsymbol{X},\boldsymbol{Y};\alpha)$ was first studied in [42] in the regime of fixed n and $\min\{p,q\}\to\infty$, and then in [17] in a high-dimensional regime slightly broader than ours (1.2).

Having understood the behaviour of the test (1.6) under the null, we now turn to the more challenging question of its behaviour under a generic alternative covariance Σ. Using again the non-null CLT in (1.3), we show that the test statistic in (1.6) is asymptotically normal with a mean shift (formal statement see Theorem 3.1):

$$\mathbb{E}_{\Sigma}\Psi(\boldsymbol{X},\boldsymbol{Y};\alpha) = \mathbb{P}(|\mathcal{N}(m_n(\Sigma),1)| > z_{\alpha/2}) + \mathfrak{o}(1). \tag{1.7}$$

Here $\mathbb{E}_{\Sigma}$ denotes expectation under the data distribution with covariance Σ so the left side is the power of the test $\Psi(\boldsymbol{X},\boldsymbol{Y};\alpha)$, and the mean shift parameter $m_n(\Sigma)$ can be either

$$\frac{n\cdot \mathrm{dCor}^2(X,Y)}{\sqrt{2}} \equiv \frac{n\,\mathrm{dCov}^2(X,Y)}{\sqrt{2\,\mathrm{dCov}^2(X)\,\mathrm{dCov}^2(Y)}} \quad \text{or} \quad \frac{n\|\Sigma_{XY}\|_F^2}{\sqrt{2}\|\Sigma_X\|_F\|\Sigma_Y\|_F}.$$

Here the (rescaled) left side is known as the *distance correlation* between X and Y, and its asymptotic equivalence to the right side follows again from the mean expansion in (1.4). It follows directly from (1.7) that if the spectra of Σ_X and Σ_Y are appropriately bounded, $\Psi(\boldsymbol{X},\boldsymbol{Y};\alpha)$ has asymptotically full power if and only if $n\cdot \mathrm{dCor}^2(X,Y)\to\infty$. A complementary minimax lower bound in Theorem 3.3 shows that this separation rate cannot be further improved from an information theoretic point of view.

Power results for tests based on distance covariance (correlation) are scarce, particularly in high dimensions when both n and p and/or q diverge to infinity. [52] gives a relatively complete power characterization for a related studentized test in the regime of fixed n and $p\wedge q\to\infty$, followed by some partial results in the regime $\min\{p,q\}/n^2\to\infty$. The same test $\Psi(\boldsymbol{X},\boldsymbol{Y};\alpha)$ as in (1.6) is recently analysed in [17] in the slightly broader regime $\min\{n,\max\{p,q\}\}\to\infty$, but their analysis requires a much stronger condition $\sqrt{n}\cdot \mathrm{dCor}^2(X,Y)\to\infty$ for power consistency (see their theorem 5 and subsequent discussion). In contrast, under the distributional Assumption A, (1.7) gives a much more precise characterization of the power behaviour of the distance correlation test (1.6), even when consistency does not hold.

1.4 *Kernel generalizations and power universality*

Following [18,19], the distance covariance in (1.1) can be naturally generalized to the so-called *Hilbert–Schmidt covariance*:

$$\begin{aligned}\mathrm{dCov}^2(X,Y;f,\gamma) \equiv{}& \mathbb{E}\Big[f_X\big(\|X_1-X_2\|/\gamma_X\big)f_Y\big(\|Y_1-Y_2\|/\gamma_Y\big)\Big]\\ &-2\mathbb{E}\Big[f_X\big(\|X_1-X_2\|/\gamma_X\big)f_Y\big(\|Y_1-Y_3\|/\gamma_X\big)\Big]\\ &+\mathbb{E}\Big[f_X\big(\|X_1-X_2\|/\gamma_X\big)\Big]\mathbb{E}\Big[f_Y\big(\|Y_1-Y_2\|/\gamma_Y\big)\Big].\end{aligned} \tag{1.8}$$

Here $f=(f_X,f_Y)$ are kernel functions, and $\gamma=(\gamma_X,\gamma_Y)\in\mathbb{R}^2_{\geq 0}$ are the bandwidth parameters for X and Y, respectively. $\mathrm{dCov}^2(X;f,\gamma)$ and $\mathrm{dCov}^2(Y;f,\gamma)$ are defined analogously. The above definition reduces to the (rescaled) standard distance covariance in (1.1) when the kernel is taken to be the identity function. Let $\mathrm{dCov}^2_*(\boldsymbol{X},\boldsymbol{Y};f,\gamma)$, $\mathrm{dCov}^2(\boldsymbol{Y};f,\gamma)$, $\mathrm{dCov}^2_*(\boldsymbol{X};f,\gamma)$ be the sample kernel distance covariance. As a key step of the universality results presented below, we show that these quantities can be related to the standard sample distance covariance: under mild conditions on kernels $f=(f_X,f_Y)$ and bandwidths $\gamma=(\gamma_X,\gamma_Y)$,

$$\mathrm{dCov}^2_*(\boldsymbol{X},\boldsymbol{Y};f,\gamma)=\varrho(\gamma)\,\mathrm{dCov}^2_*(\boldsymbol{X},\boldsymbol{Y})\big(1+\mathfrak{o}_{\mathbf{P}}(1)\big). \tag{1.9}$$

Here $\mathfrak{o}_{\mathbf{P}}(1)$ is again under the asymptotics (1.2) and $\varrho(\gamma)$, whose exact definition is given in (2.13) below, depends on the kernels f, bandwidths γ and population covariance Σ. Similar expansions hold for the marginal quantities $\mathrm{dCov}^2(X;f,\gamma)$ and $\mathrm{dCov}^2(Y;f,\gamma)$.

Relation (1.9) implies that as long as the scaling factor $\varrho(\gamma)$ stabilizes away from zero and infinity, $\mathrm{dCov}^2_*(\boldsymbol{X},\boldsymbol{Y};f,\gamma)$ (up to a scaling) shares the same limiting distribution as $\mathrm{dCov}^2_*(\boldsymbol{X},\boldsymbol{Y})$, which has been studied in detail in the previous subsection. In particular, both the non-null CLT in (1.3) and the power expansion in (1.7) hold for the kernelized distance covariance as well, upon changing the test (1.6) to its kernelized version in the latter result; see Theorems 2.6 and 3.1 for formal statements. In other words, the power behaviour in (1.7) exhibits *universality* with respect to both the choice of kernels and bandwidth parameters; see Section 4 for numerical evidence.

1.5 *Organization*

The rest of the paper is organized as follows. Section 2 starts with some background knowledge of the distance covariance metric and then states the main non-null CLTs for both the canonical sample distance covariance and its kernel generalizations. Section 3 studies the power behaviour of a class of generalized kernel distance correlation tests and discusses their minimax optimality. Some numerical simulations for the main results in the preceding two sections are presented in Section 4, with some concluding remarks in Section 5. Section 6 is devoted to a proof outline for the non-null CLTs. Details of the important steps are then presented in Sections 7–10, followed by the main proof in Section 11. The rest of the technical proofs are deferred to the supplement.

1.6 *Notation*

For any positive integer n, let $[n]$ denote the set $\{1,\dots,n\}$. For $a,b\in\mathbb{R}$, $a\vee b\equiv\max\{a,b\}$ and $a\wedge b\equiv\min\{a,b\}$. For $a\in\mathbb{R}$, let $a_+\equiv a\vee 0$ and $a_-\equiv(-a)\vee 0$. For $x\in\mathbb{R}^n$, let $\|x\|_p=\|x\|_{\ell_p(\mathbb{R}^n)}$ denote its

p-norm ($0 \le p \le \infty$) with $\|x\|_2$ abbreviated as $\|x\|$. Let $B_p(r;x) \equiv \{z \in \mathbb{R}^p : \|z-x\| \le r\}$ be the unit ℓ_2 ball in $\mathbb{R}^p$. By $\mathbf{1}_n$ we denote the vector of all ones in $\mathbb{R}^n$. For a matrix $M \in \mathbb{R}^{n\times n}$, let $\|M\|_{\mathrm{op}}$ and $\|M\|_F$ denote the spectral and Frobenius norms of M, respectively. For two matrices M, N of the same size, let $M \circ N$ denote their Hadamard product. We use $\{e_j\}$ to denote the canonical basis, whose dimension should be self-clear from the context.

We use C_x to denote a generic constant that depends only on x, whose numeric value may change from line to line unless otherwise specified. Notations $a \lesssim_x b$ and $a \gtrsim_x b$ mean $a \le C_x b$ and $a \ge C_x b$, respectively, and $a \asymp_x b$ means $a \lesssim_x b$ and $a \gtrsim_x b$. The symbol $a \lesssim b$ means $a \le Cb$ for some absolute constant C. For two non-negative sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \ll b_n$ (respectively, $a_n \gg b_n$) if $\lim_{n\to\infty}(a_n/b_n) = 0$ (respectively, $\lim_{n\to\infty}(a_n/b_n) = \infty$). We write $a_n \sim b_n$ if $\lim_{n\to\infty}(a_n/b_n) = 1$. We follow the convention that $0/0 = 0$.

Let φ, Φ be the density and the cumulative distribution function of a standard normal random variable. For any $\alpha \in (0,1)$, let z_α be the normal quantile defined by $\mathbb{P}(\mathcal{N}(0,1) > z_\alpha) = \alpha$. For two random variables X, Y on $\mathbb{R}$, we use $d_{\mathrm{Kol}}(X,Y)$ to denote their Kolmogorov distance defined by

$$d_{\mathrm{Kol}}(X,Y) \equiv \sup_{t\in\mathbb{R}} \big|\mathbb{P}\big(X \le t\big) - \mathbb{P}\big(Y \le t\big)\big|. \tag{1.10}$$

Here $\mathscr{B}(\mathbb{R})$ denotes the Borel σ-algebra of $\mathbb{R}$.

2. Non-null CLTs

2.1 *Distance covariance: a review*

We start with a review for the distance covariance (correlation). For two random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, the squared distance covariance [44] is originally defined by

$$\mathrm{dCov}^2(X,Y) \equiv \int_{\mathbb{R}^{p+q}} \frac{|\varphi_{(X,Y)}(t,s) - \varphi_X(t)\varphi_Y(s)|^2}{c_p c_q \|t\|^{p+1}\|s\|^{q+1}}\,\mathrm{d}t\mathrm{d}s.$$

Here $c_p = \pi^{(p+1)/2}/\Gamma\big((p+1)/2\big)$ with $\Gamma(\cdot)$ denoting the gamma function, and $\varphi(\cdot)$ is the characteristic function. The marginal quantities $\mathrm{dCov}^2(X,X)$ and $\mathrm{dCov}^2(Y,Y)$ are defined analogously, and we will shorthand them as $\mathrm{dCov}^2(X)$ and $\mathrm{dCov}^2(Y)$ in the sequel. It is well known that X and Y are independent if and only if $\mathrm{dCov}(X,Y) = 0$, hence $\mathrm{dCov}^2(X,Y)$ captures any kind of dependence between X and Y including nonlinear and non-monotone ones. Analogous to the standard notion of covariance and correlation, the squared distance correlation is defined by

$$\mathrm{dCor}^2(X,Y) \equiv \frac{\mathrm{dCov}^2(X,Y)}{\sqrt{\mathrm{dCov}^2(X)\,\mathrm{dCov}^2(Y)}},$$

with convention $\mathrm{dCor}^2(X,Y) \equiv 0$ if $\mathrm{dCov}^2(X)\,\mathrm{dCov}^2(Y) = 0$.

The distance covariance can be equivalently characterized in a number of different ways. In addition to (1.1), another useful representation that will be particularly relevant for our purpose is through the

double-centred distances:

$$U(x_1,x_2) \equiv \|x_1-x_2\| - \mathbb{E}\|x_1-X\| - \mathbb{E}\|X-x_2\| + \mathbb{E}\|X-X'\|,$$
$$V(y_1,y_2) \equiv \|y_1-y_2\| - \mathbb{E}\|y_1-Y\| - \mathbb{E}\|Y-y_2\| + \mathbb{E}\|Y-Y'\|. \tag{2.1}$$

Then by (1.1) or [29, pp. 3287], we have the identity

$$\mathrm{dCov}^2(X,Y) = \mathbb{E}U(X_1,X_2)V(Y_1,Y_2). \tag{2.2}$$

Now we define the sample distance covariance. For n copies of i.i.d. observations $(X_1,Y_1),\dots,(X_n,Y_n)$, define two symmetric matrices $A,B \in \mathbb{R}^{n\times n}$ entrywise by

$$A_{k\ell} \equiv \|X_k - X_\ell\|, \quad B_{kl} \equiv \|Y_k - Y_\ell\|, \quad 1 \le k,\ell \le n. \tag{2.3}$$

Following [43], the bias-corrected sample distance covariance is defined by

$$\mathrm{dCov}^2_*(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{n(n-3)}\sum_{k\neq\ell} A^*_{k\ell}B^*_{k\ell}, \tag{2.4}$$

where $A^*, B^* \in \mathbb{R}^{n\times n}$ are U-centred versions of A,B defined by

$$A^* = A - \frac{\mathbf{1}\mathbf{1}^\top A + A\mathbf{1}\mathbf{1}^\top}{n-2} + \frac{\mathbf{1}^\top A\mathbf{1}\mathbf{1}^\top}{(n-1)(n-2)},$$
$$B^* = B - \frac{\mathbf{1}\mathbf{1}^\top B + B\mathbf{1}\mathbf{1}^\top}{n-2} + \frac{\mathbf{1}^\top B\mathbf{1}\mathbf{1}^\top}{(n-1)(n-2)}. \tag{2.5}$$

Marginal quantities $\mathrm{dCov}^2_*(\boldsymbol{X},\boldsymbol{X})$ and $\mathrm{dCov}^2_*(\boldsymbol{Y},\boldsymbol{Y})$ are defined analogously, and will be shorthanded as $\mathrm{dCov}^2_*(\boldsymbol{X})$ and $\mathrm{dCov}^2_*(\boldsymbol{Y})$ in the sequel.

The definition of the sample distance covariance $\mathrm{dCov}^2_*(\boldsymbol{X},\boldsymbol{Y})$ in (2.4) above looks a bit mysterious at first sight, but the following representation via a fourth-order U-statistic makes it clear why the definition is indeed natural. Recall the definitions of U,V in (2.1).

PROPOSITION 2.1. ([17, 50]). The following holds:

$$\mathrm{dCov}^2_*(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{\binom{n}{4}}\sum_{i_1<\cdots<i_4} k\big(Z_{i_1},Z_{i_2},Z_{i_3},Z_{i_4}\big),$$

where the symmetric kernel can be either

$$k(Z_1,Z_2,Z_3,Z_4) = \frac{1}{4!}\sum_{(i_1,\dots,i_4)\in\sigma(1,2,3,4)}\Big[\|X_{i_1}-X_{i_2}\|\|Y_{i_1}-Y_{i_2}\|$$
$$+ \|X_{i_1}-X_{i_2}\|\|Y_{i_3}-Y_{i_4}\| - 2\|X_{i_1}-X_{i_2}\|\|Y_{i_1}-Y_{i_3}\|\Big], \tag{2.6}$$

or

$$k(Z_1,Z_2,Z_3,Z_4)=\frac{1}{4!}\sum_{(i_1,\dots,i_4)\in\sigma(1,2,3,4)}\Big[U(X_{i_1},X_{i_2})V(Y_{i_1},Y_{i_2})$$
$$+U(X_{i_1},X_{i_2})V(Y_{i_3},Y_{i_4})-2U(X_{i_1},X_{i_2})V(Y_{i_1},Y_{i_3})\Big]. \tag{2.7}$$

Here $Z_i=(X_i,Y_i)$ for $i\in\mathbb{N}$, and $\sigma(1,2,3,4)$ denotes the set of all ordered permutation of $\{1,2,3,4\}$.

It is a direct consequence of the above result that $\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y})$ is an unbiased estimator for $\mathrm{dCov}^2(X,Y)$. The fact that $\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y})$ can be represented as a U-statistic is first validated in [22, Section 3.2]. The kernel representation (2.6) (proved in e.g. [50, Lemma 2.1]) is quite natural in that it gives an unbiased estimate for the population in the form (1.1). The double-centred version (2.7), which turns out to be more convenient and useful for the purpose of theoretical developments in this paper, is essentially proved in [17, Lemma 5] in a different form. For the convenience of the reader, we provide a self-contained proof in Appendix A.

2.2 *General non-null CLTs I: distance covariance*

Throughout the paper, we work with the following distributional family of (X,Y) with a *separable covariance* structure:

Assumption A. Suppose

$$(X^\top,Y^\top)^\top\overset{d}{=}\Sigma^{1/2}Z. \tag{2.8}$$

Here Σ, in its block form $[\Sigma_X,\Sigma_{XY};\Sigma_{YX},\Sigma_Y]$, is a covariance matrix in $\mathbb{R}^{(p+q)\times(p+q)}$, $Z\in\mathbb{R}^{p+q}$ has i.i.d. components with mean 0, variance 1 and satisfies the following:

(A1) Z_1 is symmetric around 0 with excess kurtosis $\kappa\equiv\mathbb{E}Z_1^4-3$.

(A2) Z_1 satisfies a Poincaré inequality: for some $c_*>0$, we have $\mathrm{Var}f(Z_1)\le c_*\mathbb{E}(f'(Z_1))^2$ for any absolutely continuous f such that $\mathbb{E}(f'(Z_1))^2<\infty$.

(A3) Z_1 has a Lebesgue density $f_Z(\cdot)$ with $\sup_{x\in(-\varepsilon,\varepsilon)}f_Z(x)\le C_\varepsilon$ for some small $\varepsilon>0$ and positive C_ε. For future purpose, let $\varepsilon_0\equiv\sup\{\varepsilon>0:2\pi\varepsilon\cdot\sup_{x\in(-\varepsilon,\varepsilon)}f_Z(x)\le 1\}$.

The distribution class of (X,Y) with separable covariance is quite common in the literature, in particular in the study of non-null behaviour of statistics related to large random matrices; the readers are referred to the recent papers [11,12,26,45] and monographs [2,14] for more backgrounds and results under separable covariance in this direction.

The major assumption on the distribution of Z_1 is the requirement of a Poincaré inequality in condition (A2). It is well known that the existence of a Poincaré inequality as in (A2) is equivalent to exponential mixing of a Markov semigroup with stationary distribution Z_1 and Dirichlet form $\mathscr{E}(f,g)=\mathbb{E}f'(Z_1)g'(Z_1)$, cf. [3]. An important example fulfilling Assumption A is the family of symmetric, strongly log-concave distributions, which are known to satisfy a Poincaré inequality (cf. [6,36]) and contain the

Gaussian distribution as a special case. It is easy to further weaken condition (A2) to a *weighted Poincaré inequality* as in [7]; we shall not pursue these formal refinements here. Condition (A3) above is purely technical, and can be further weakened at the cost of a more involved mathematical expression. We choose to work under this condition for clean presentation.

We mention two important implications of Assumption A: (i) since Z_1 has a Lebesgue density, $\kappa \geq c_0 - 2$ for some $c_0 > 0$ depending only on the distribution of Z_1; (ii) by [7, Theorem 4.1], Z_1 has moments of any order with $(\mathbb{E}|Z_1|^p)^{1/p} \leq p \cdot \sqrt{c_*/2}$ for any $p \geq 1$.

Some notation that will be used throughout the paper:

$$\tau_X^2 \equiv \mathbb{E}\|X - X'\|^2 = 2\,\mathrm{tr}(\Sigma_X), \quad \tau_Y^2 \equiv \mathbb{E}\|Y - Y'\|^2 = 2\,\mathrm{tr}(\Sigma_Y). \tag{2.9}$$

We also reserve κ for the excess kurtosis of Z_1:

$$\kappa \equiv \mathbb{E}Z_1^4 - 3. \tag{2.10}$$

Since Z_1 has a Lebesgue density, we have $\kappa \geq c_0 - 2$ for some $c_0 > 0$ that only depends on the distribution of Z_1.

Let $I_{[ij]} \equiv (\mathbf{1}_{(i',j')=(i,j)})_{1\leq i',j'\leq 2}$ be the indicator of the block matrix, and $\Sigma_{[ij]} \equiv (\Sigma_{(i'j')}\mathbf{1}_{(i',j')=(i,j)})_{1\leq i',j'\leq 2}$. Let

$$G_{[ij]} \equiv \Sigma^{1/2}\Sigma_{[ij]}\Sigma^{1/2}, \quad H_{[ij]} \equiv \Sigma^{1/2}I_{[ij]}\Sigma^{1/2}. \tag{2.11}$$

Then $G_{[ij]}^\top = G_{[ji]}$ and $H_{[ij]}^\top = H_{[ji]}$. Let $\bar{G}_{[1,2]} \equiv (G_{[12]} + G_{[21]})/2$.

The following non-null CLT is the first main result of this paper; its proof can be found in Section 11.

THEOREM 2.2. Suppose that Assumption A holds, and that the spectrum of Σ is contained in $[1/M, M]$ for some $M > 1$. Then there exists some $C = C(M, Z_1) > 0$ such that

$$d_{\mathrm{Kol}}\left(\frac{\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y}) - \mathrm{dCov}^2(X,Y)}{\sigma_n(X,Y)}, \mathcal{N}(0,1)\right) \leq \frac{C}{(n\wedge p\wedge q)^{1/6}}.$$

Here $\sigma_n(X,Y)$ can be either $\mathrm{Var}^{1/2}(\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y}))$ or $\bar{\sigma}_n(X,Y)$, where

$$\bar{\sigma}_n^2(X,Y) \equiv \bar{\sigma}_{n,1}^2(X,Y) + \bar{\sigma}_{n,2}^2(X,Y),$$

with

$$\begin{aligned}\bar{\sigma}_{n,1}^2(X,Y) \equiv \frac{4}{n\tau_X^2\tau_Y^2}\Bigg[&\|\Sigma_{XY}\Sigma_{YX}\|_F^2 + \operatorname{tr}(\Sigma_{XY}\Sigma_Y\Sigma_{YX}\Sigma_X) + \frac{\|\Sigma_{XY}\|_F^4\|\Sigma_X\|_F^2}{2\tau_X^4} \\ &+ \frac{\|\Sigma_{XY}\|_F^4\|\Sigma_Y\|_F^2}{2\tau_Y^4} - \frac{2\|\Sigma_{XY}\|_F^2}{\tau_X^2}\operatorname{tr}(\Sigma_{XY}\Sigma_{YX}\Sigma_X) \\ &- \frac{2\|\Sigma_{XY}\|_F^2}{\tau_Y^2}\operatorname{tr}(\Sigma_{YX}\Sigma_{XY}\Sigma_Y) + \frac{\|\Sigma_{XY}\|_F^6}{\tau_X^2\tau_Y^2} + \kappa\cdot\Big(\operatorname{tr}(G_{[12]}\circ G_{[12]}) \\ &+ \frac{\|\Sigma_{XY}\|_F^4}{4\tau_X^4}\operatorname{tr}(H_{[11]}\circ H_{[11]}) + \frac{\|\Sigma_{XY}\|_F^4}{4\tau_Y^4}\operatorname{tr}(H_{[22]}\circ H_{[22]}) \\ &- \frac{\|\Sigma_{XY}\|_F^2}{2\tau_X^2}\operatorname{tr}(H_{[11]}\circ G_{[12]}) - \frac{\|\Sigma_{XY}\|_F^2}{2\tau_Y^2}\operatorname{tr}(H_{[22]}\circ G_{[12]}) \\ &+ \frac{\|\Sigma_{XY}\|_F^4}{4\tau_X^2\tau_Y^2}\operatorname{tr}(H_{[11]}\circ H_{[22]})\Big)\Bigg], \\ \bar{\sigma}_{n,2}^2(X,Y) \equiv \frac{2}{n(n-1)\tau_X^2\tau_Y^2}&\big(\|\Sigma_X\|_F^2\|\Sigma_Y\|_F^2 + \|\Sigma_{XY}\|_F^4\big).\end{aligned}$$

Here $\circ$ is the Hadamard product and $H_{[..]}$ and $G_{[..]}$ are given in (2.11) above.

REMARK 2.3. (Variance formula). The variance formula $\operatorname{Var}\big(\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y})\big) = \big(1+\mathfrak{o}(1)\big)\bar{\sigma}_n^2(X,Y)$ is valid in the high-dimensional limit $n\wedge p\wedge q\to\infty$; see Section 9 ahead for details. The same is true for the variance formula in Theorem 2.5 below.

REMARK 2.4. (Convergence rate). The convergence rate $(n\wedge p\wedge q)^{-1/6}$ results from the application of Chatterjee's second-order Poincaré inequality [9] (see Section 10 for details) and is likely to be sub-optimal. Such rate, however, is common in the literature. For example, [17, Proposition 3] assumed (X,Y) to be jointly Gaussian and employed a martingale CLT theorem to prove the normal limit under the null, but the convergence rate is also only $(pq)^{-1/5}\vee n^{-1/5}$. It is an interesting but challenging task to pin down the optimal Gaussian approximation rate in terms of (n,p,q).

As mentioned in the introduction, most of the CLTs in the literature so far are derived under the null case that X and Y are independent, and to the best of our knowledge, Theorem 2.2 is the first non-null CLT that applies to a general class of alternatives. Due to the challenging nature of non-null analysis, the proof of the above theorem requires several technically involved and intertwined steps, so an outline will be provided in Section 6 that discusses the relevance of the groundwork laid in Sections 7–10, which culminates in the proof of Theorem 2.2 in Section 11.

Let us examine the variance structure in more detail. Roughly speaking, in the high-dimensional regime $n\wedge p\wedge q\to\infty$, the variance $\bar{\sigma}_n^2(X,Y)$ of $\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y})$ only contains two possibly different sources—the first part $\bar{\sigma}_{n,1}^2(X,Y)$ comes from the contribution of the non-degenerate first-order kernel, while the second part $\bar{\sigma}_{n,2}^2(X,Y)$ comes from the contribution of the degenerate second-order kernel, in the Hoeffding decomposition of $\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y})$ to be detailed in Section 9 ahead.

One notable complication of $\bar{\sigma}_n^2(X, Y)$ is the existence of terms with negative signs in the first-order variance $\bar{\sigma}_{n,1}^2(X, Y)$. These terms may contribute to the same order of the leading quantities $\|\Sigma_{XY}\Sigma_{YX}\|_F^2$ and $\mathrm{tr}(\Sigma_{XY}\Sigma_Y\Sigma_{YX}\Sigma_X)$, but a lower bound in Lemma 9.5 ahead shows that their contributions do not lead to 'cancellations' of the main terms. In fact, under the spectrum condition of Theorem 2.2, the second claim of Lemma 9.5 indicates the order of $\bar{\sigma}_n^2(X, Y)$ with terms of positive signs only:

$$\bar{\sigma}_n^2(X, Y) \asymp_M \max\left\{\frac{\|\Sigma_{XY}\|_F^2}{npq}, \frac{1}{n^2}\right\}.$$

Here the first term in the above maximum is contributed by $\bar{\sigma}_{n,1}^2(X, Y)$ and the second term is contributed by $\bar{\sigma}_{n,2}^2(X, Y)$. Now we consider two regimes:

- (*Ultra high-dimensional regime* $\sqrt{pq} \gg n$). In this regime, as $\|\Sigma_{XY}\|_F^2 \lesssim_M p \wedge q \le \sqrt{pq}$ via Lemma G.5 in the appendix, the variance $\bar{\sigma}_n^2(X, Y)$ in this ultra high-dimensional regime is completely determined by the contribution from the degenerate second-order kernel $\bar{\sigma}_{n,2}^2(X, Y)$:

$$\bar{\sigma}_n^2(X, Y) = \big(1 + \mathfrak{o}(1)\big)\bar{\sigma}_{n,2}^2(X, Y).$$

- (*Moderate high-dimensional regime* $\sqrt{pq} \lesssim n$). In this regime, there are three possibilities:
  - If $\|\Sigma_{XY}\|_F^2 \ll (pq)/n$, which includes the null $\Sigma_{XY} = 0$ as a special case, the variance $\bar{\sigma}_n^2(X, Y)$ is again completely determined by the degenerate second-order kernel $\bar{\sigma}_{n,2}^2(X, Y)$.
  - If $\|\Sigma_{XY}\|_F^2 \gg (pq)/n$, then the variance $\bar{\sigma}_n^2(X, Y)$ is completely determined by the non-degenerate first-order kernel $\bar{\sigma}_{n,1}^2(X, Y)$:

$$\bar{\sigma}_n^2(X, Y) = \big(1 + \mathfrak{o}(1)\big)\bar{\sigma}_{n,1}^2(X, Y).$$

    If furthermore $\|\Sigma_{XY}\|_F^2 \ll p \wedge q$ (i.e. excluding the critical regime $\|\Sigma_{XY}\|_F^2 \asymp p \wedge q$), then the first-order variance $\bar{\sigma}_{n,1}^2(X, Y)$ can be simplified to be

$$\bar{\sigma}_n^2(X, Y) = \frac{4\big(1 + \mathfrak{o}(1)\big)}{n\tau_X^2\tau_Y^2}\Big[\|\Sigma_{XY}\Sigma_{YX}\|_F^2 + \mathrm{tr}(\Sigma_{XY}\Sigma_Y\Sigma_{YX}\Sigma_X)\Big].$$

  - If $\|\Sigma_{XY}\|_F^2 \asymp (pq)/n$, the variance $\bar{\sigma}_n^2(X, Y)$ is contributed by both the non-degenerate first-order kernel $\bar{\sigma}_{n,1}^2(X, Y)$ and the degenerate second-order kernel $\bar{\sigma}_{n,2}^2(X, Y)$ so the general variance expression in Theorem 2.2 cannot be further simplified.

The smallest eigenvalue condition in Theorem 2.2 excludes the case $X = Y$, but a slight variation of the proof can cover this case as well. We record formally the result below.

THEOREM 2.5. Suppose that Assumption A holds, and that the spectrum of Σ_X is contained in $[1/M, M]$ for some $M > 1$. Then there exists some $C = C(M, Z_1) > 0$ such that

$$d_{\mathrm{Kol}}\left(\frac{\mathrm{dCov}_*^2(\boldsymbol{X}) - \mathrm{dCov}^2(X)}{\sigma_n(X)}, \mathscr{N}(0,1)\right) \le \frac{C}{(n \wedge p)^{1/6}}.$$

Here $\sigma_n(X)$ can be either $\mathrm{Var}^{1/2}\left(\mathrm{dCov}_*^2(\boldsymbol{X})\right)$ or $\bar{\sigma}_n(X, X)$, where

$$\bar{\sigma}_n^2(X) \equiv \bar{\sigma}_{n,1}^2(X) + \bar{\sigma}_{n,2}^2(X),$$

with

$$\begin{aligned}\bar{\sigma}_{n,1}^2(X) \equiv \frac{1}{n\,\mathrm{tr}^2(\Sigma_X)}\Bigg[&2\|\Sigma_X^2\|_F^2 + \frac{\|\Sigma_X\|_F^6}{2\,\mathrm{tr}^2(\Sigma_X)} - \frac{2\|\Sigma_X\|_F^2\,\mathrm{tr}(\Sigma_X^3)}{\mathrm{tr}(\Sigma_X)} \\ &+ \kappa \cdot \left(\mathrm{tr}(\Sigma_X^2 \circ \Sigma_X^2) + \frac{3\|\Sigma_X\|_F^4}{4\tau_X^4}\,\mathrm{tr}(\Sigma_X \circ \Sigma_X) - \frac{\|\Sigma_X\|_F^2}{\tau_X^2}\,\mathrm{tr}(\Sigma_X^2 \circ \Sigma_X)\right)\Bigg], \\ \bar{\sigma}_{n,2}^2(X) \equiv \frac{\|\Sigma_X\|_F^4}{n(n-1)\,\mathrm{tr}^2(\Sigma_X)}&.\end{aligned}$$

The variance $\bar{\sigma}_n^2(X)$ in Theorem 2.5 is simpler than that in Theorem 2.2. In fact, a similar consideration using the variance lower bound in Lemma 9.5, we may obtain the order of $\bar{\sigma}_n^2(X)$:

$$\bar{\sigma}_n^2(X) \asymp_M \max\left\{\frac{1}{np}, \frac{1}{n^2}\right\}.$$

Through the Hoeffding decomposition of $\mathrm{dCov}_*^2(\boldsymbol{X})$, the first and second terms in the maximum are contributed by the variance of the non-degenerate first-order kernel and the the degenerate second-order kernel, respectively. Therefore,

- In the ultra high-dimensional regime $p \gg n$, the variance $\bar{\sigma}_n^2(X)$ is completely determined by the degenerate second-order kernel $\bar{\sigma}_{n,2}^2(X)$.
- In the strictly moderate high-dimensional regime $p \ll n$, the variance $\bar{\sigma}_n^2(X)$ is completely determined by the non-degenerate first-order kernel $\bar{\sigma}_{n,1}^2(X)$.
- In the critical regime $p \asymp n$, the variance $\bar{\sigma}_n^2(X)$ is determined jointly by the first- and second-order kernels $\bar{\sigma}_{n,1}^2(X)$, $\bar{\sigma}_{n,2}^2(X)$ so cannot be in general simplified.

2.3 *General non-null CLTs II: generalized kernel distance covariance*

The sample distance covariance $\mathrm{dCov}_*^2(\boldsymbol{X}, \boldsymbol{Y})$ can be generalized using kernel functions as follows. Given functions $f_X, f_Y : \mathbb{R}_{\ge 0} \to \mathbb{R}$, and bandwidth parameters $\gamma_X, \gamma_Y > 0$, let for $1 \le k, \ell \le n$

$$A_{k\ell}(f_X, \gamma_X) \equiv f_X\big(\|X_k - X_\ell\|/\gamma_X\big)\mathbf{1}_{k\neq\ell}, \quad B_{k\ell}(f_Y, \gamma_Y) \equiv f_Y\big(\|Y_k - Y_\ell\|/\gamma_Y\big)\mathbf{1}_{k\neq\ell}.$$

It is essential to set the diagonal terms $\{A_{kk}(f_X,\gamma_X)\}_k, \{B_{kk}(f_Y,\gamma_Y)\}_k$ to be 0, so that the generalized kernel distance covariance to be introduced below can be analysed in a unified manner; see Proposition F.1 in the appendix for details. Now with $A^*_{k\ell}(f_X,\gamma_X), B^*_{k\ell}(f_Y,\gamma_Y)$ defined similarly as in (2.5) by replacing $A_{k\ell}, B_{k\ell}$ with $A_{k\ell}(f_X,\gamma_X), B_{k\ell}(f_Y,\gamma_Y)$, we may define the generalized sample distance covariance with kernels $f=(f_X,f_Y)$ and bandwidth parameters $\gamma=(\gamma_X,\gamma_Y)\in\mathbb{R}^2_{>0}$ by

$$\mathrm{dCov}^2_*(\boldsymbol{X},\boldsymbol{Y};f,\gamma)\equiv\frac{1}{n(n-3)}\sum_{k\neq\ell}A^*_{k\ell}(f_X,\gamma_X)B^*_{k\ell}(f_Y,\gamma_Y), \tag{2.12}$$

and its population version $\mathrm{dCov}^2(X,Y;f,\gamma)$ as in (1.8). Marginal quantities $\mathrm{dCov}^2(X;f,\gamma)$ and $\mathrm{dCov}^2(Y;f,\gamma)$ are defined analogously, and similar to distance correlation, the kernelized distance correlation is defined as

$$\mathrm{dCor}^2(X,Y;f,\gamma)\equiv\frac{\mathrm{dCov}^2(X,Y;f,\gamma)}{\sqrt{\mathrm{dCov}^2(X;f,\gamma)\,\mathrm{dCov}^2(Y;f,\gamma)}},$$

with convention $\mathrm{dCor}^2(X,Y;f,\gamma)\equiv 0$ if $\mathrm{dCov}^2(X;f,\gamma)\,\mathrm{dCov}^2(Y;f,\gamma)=0$.

A more general formulation, when replacing $f_X(\|X_\ell-X_k\|/\gamma_X)$ (resp. $f_Y(\|Y_\ell-Y_k\|/\gamma_X)$) by some generic bivariate kernel $k_X(X_\ell,X_k)$ (resp. $k_Y(Y_\ell,Y_k)$), is also known as the *Hilbert–Schmidt independence criteria*, see e.g. [18,19], which can in fact be written as the maximum mean discrepancy between the joint distribution and the marginal distributions of X and Y; see e.g. [37, Section 3.3] for an in-depth discussion. Two particular important choices for f are the Laplace and Gaussian kernels:

- (Laplace kernel) $f(w)=e^{-w}$;
- (Gaussian kernel) $f(w)=e^{-w^2/2}$.

These kernels have been considered in, e.g. [20,52].

Assumption B. (Conditions on the kernel f)Suppose that $f\in\{f_X,f_Y\}$ is four times differentiable on $(0,\infty)$ such that:

1. f is bounded on $[0,M)$ for any $M>0$.
2. For any $\varepsilon>0$, $\max_{1\le\ell\le4}\sup_{x\ge\varepsilon}|f^{(\ell)}(x)|\le C_\varepsilon$ for some $C_\varepsilon>0$.
3. For any $\varepsilon>0$, there exists some $c_\varepsilon>0$ such that $\inf_{x\in(\varepsilon,\varepsilon^{-1})}|f'(x)|\ge c_\varepsilon$.
4. There exists some $\mathfrak{q}>0$ such that $\limsup_{x\downarrow0}\max_{1\le\ell\le4}x^{\mathfrak{q}}|f^{(\ell)}(x)|<\infty$.

In words, Assumption B-(1)(2) require the kernel functions f_X,f_Y and its derivatives to be appropriately bounded, (3) requires the first derivative to be bounded from below on any compacta in $(0,\infty)$ and finally (4) regulates that the derivatives of f_X,f_Y up to the fourth order can only blow up at 0 with at most a polynomial rate of divergence. It is easy to check that both the Laplace/Gaussian kernels, and the canonical choice $f(x)=x$ that recovers the distance covariance (up to a scaling factor) all satisfy Assumption B.

Let $\rho_X \equiv \tau_X/\gamma_X$, $\rho_Y \equiv \tau_Y/\gamma_Y$ and

$$\varrho(\gamma) \equiv \frac{f_X'(\rho_X)f_Y'(\rho_Y)}{\gamma_X\gamma_Y}. \tag{2.13}$$

We are now ready to state the following non-null CLT for the generalized kernel distance covariance $\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y};f,\gamma)$; its proof can be found in Appendix F.

THEOREM 2.6. Suppose that Assumptions A and B hold, and that (i) the spectrum of $\boldsymbol{\Sigma}$ (ii) ρ_X, ρ_Y are contained in $[1/M, M]$ for some $M > 1$. Then there exists some $C = C(f, M, Z_1) > 0$ such that

$$d_{\mathrm{Kol}}\left(\frac{\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y};f,\gamma) - \mathrm{dCov}^2(X,Y;f,\gamma)}{\sigma_n(X,Y;f,\gamma)}, \mathcal{N}(0,1)\right) \le \frac{C}{(n\wedge p\wedge q)^{1/6}}.$$

Here $\sigma_n(X,Y;f,\gamma) = \varrho(\gamma)\sigma_n(X,Y)$, where $\sigma_n(X,Y)$ is defined in Theorem 2.2. In the case $X = Y$, the conclusion continues to hold if (i) is replaced by (i') the spectrum of $\boldsymbol{\Sigma}_X$ is contained in $[1/M, M]$ for some $M > 1$.

We note that the conditions posed in the above theorem are not the weakest possible; for instance one may relax all the conditions on the kernels $f = (f_X, f_Y)$ and (ρ_X, ρ_Y) to some growth conditions involving n, p, q at the cost of a more involved error bound, but we have stated the current formulation to avoid unnecessary digressions.

The key step in the proof of Theorem 2.6 is to reduce the analysis of $\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y};f,\gamma)$ with general kernels f to that of the canonical $\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y})$. Analogous to the quantities U, V defined in (2.1) for the canonical distance covariance, let

$$\begin{aligned} U_{f_X,\gamma_X}(x_1,x_2) \equiv{}& f_X\big(\|x_1 - x_2\|/\gamma_X\big) - \mathbb{E}f_X\big(\|x_1 - X\|/\gamma_X\big) \\ &- \mathbb{E}f_X\big(\|X - x_2\|/\gamma_X\big) + \mathbb{E}f_X\big(\|X - X'\|/\gamma_X\big), \\ V_{f_Y,\gamma_Y}(y_1,y_2) \equiv{}& f_Y\big(\|y_1 - y_2\|/\gamma_Y\big) - \mathbb{E}f_Y\big(\|y_1 - Y\|/\gamma_Y\big) \\ &- \mathbb{E}f_Y\big(\|Y - y_2\|/\gamma_Y\big) + \mathbb{E}f_Y\big(\|Y - Y'\|/\gamma_Y\big). \end{aligned}$$

Then it is shown in Lemma F.5 that

$$U_{f_X,\gamma_X} \approx -\frac{f_X'(\rho_X)}{\gamma_X}U, \quad V_{f_Y,\gamma_Y} \approx -\frac{f_Y'(\rho_Y)}{\gamma_Y}V,$$

hence with appropriate control on the remainder terms, it follows from the U-statistic representation in (2.7) that

$$\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y};f,\gamma) \approx \frac{f_X'(\rho_X)}{\gamma_X}\frac{f_Y'(\rho_Y)}{\gamma_Y}\,\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y}) = \varrho(\gamma)\,\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y}).$$

Following this line of arguments, we are then able to study the asymptotics of $\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y};f,\gamma)$ and $\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y})$ in a unified manner; see Appendix F for detailed arguments.

2.4 *Local CLTs*

As a corollary of the non-null CLT in Theorem 2.2, we state below a local CLT that will be important to obtain the power formula for the distance correlation test introduced in (1.6). Its proof is presented in Section 12.

THEOREM 2.7. Suppose that Assumption A holds, and that the spectrum of Σ is contained in $[1/M, M]$ for some $M > 1$. Let

$$A(\Sigma) \equiv \frac{n\|\Sigma_{XY}\|_F^2}{\|\Sigma_X\|_F\|\Sigma_Y\|_F}. \tag{2.14}$$

Then there exists some constant $C = C(M, Z_1) > 0$ such that

$$d_{\mathrm{Kol}}\left(\frac{n\big(\tau_X\tau_Y\,\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y}) - \|\Sigma_{XY}\|_F^2\big)}{\sqrt{2}\|\Sigma_X\|_F\|\Sigma_Y\|_F}, \mathcal{N}(0,1)\right) \le C\left[1 \bigwedge \left(\frac{1 \vee A(\Sigma)^2}{n \wedge p \wedge q}\right)^{1/6}\right].$$

If Assumption B holds and ρ_X, ρ_Y are contained in $[1/M, M]$ for some $M > 1$, then the above conclusion holds with $\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y})$ replaced by $\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y};f,\gamma)/\varrho(\gamma)$ and C replaced by $C' = C'(M, f, Z_1)$.

The definition of the local (contiguity) parameter $A(\Sigma)$ is motivated by the critical parameter in the power expansion formula of the (generalized kernel) distance correlation test in Theorem 3.1 below. The interesting phenomenon in the local CLT above is that in the local (contiguity) regime $\limsup A(\Sigma) < \infty$, a CLT holds for $\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y})$ and $\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y};f,\gamma)/\varrho(\gamma)$ with the *null variance* σ_{null}^2 in (1.5), i.e. the variance under $\Sigma_{XY} = 0$. Of course, this necessarily implies that (recall $\bar{\sigma}_n^2 \equiv \bar{\sigma}_n^2(X, Y)$ defined in Theorem 2.2)

$$\frac{\bar{\sigma}_n^2}{\sigma_{\mathrm{null}}^2} \to 1, \qquad \text{if} \quad \limsup A(\Sigma) < \infty, \tag{2.15}$$

which can be verified via elementary calculations (see e.g. (12.1) ahead). This fact will be crucial in Theorem 3.1 ahead, where we obtain the asymptotic exact power formula for the distance correlation test using the distance correlation itself (or equivalently, $A(\Sigma)$) as the critical parameter.

3. Generalized kernel distance correlation tests

In this section, we study the performance of the distance correlation test $\Psi(\boldsymbol{X},\boldsymbol{Y};\alpha)$ in (1.6) and its kernel generalizations for the null hypothesis H_0 : $X$is independent of$Y$, or equivalently under our Gaussian assumption, $\Sigma_{XY} = 0$.

Let us start with a motivation for the test $\Psi(\boldsymbol{X},\boldsymbol{Y};\alpha)$ in (1.6) by explaining its connection to the non-null CLT derived in Theorem 2.2 for the sample distance covariance $\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y})$. The null part of Theorem 2.2 (i.e. the case of independent X and Y) motivates the following 'oracle' independence test: for any prescribed size $\alpha \in (0, 1)$,

$$\widetilde{\Psi}(\boldsymbol{X},\boldsymbol{Y};\alpha) \equiv \mathbf{1}\left(\left|\frac{\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y})}{\sigma_{\mathrm{null}}}\right| > z_{\alpha/2}\right), \tag{3.1}$$

where σ^2_{null} is the variance of $\mathrm{dCov}^2_*(\boldsymbol{X},\boldsymbol{Y})$ under the null in (1.5). Since $\mathrm{dCov}^2(X,Y)=0$ under the null, Theorem 2.2 implies immediately that the above test has an asymptotic size of α. The test $\widetilde{\Psi}(\boldsymbol{X},\boldsymbol{Y};\alpha)$, however, is not practical because even under the null, σ^2_{null} might still depend on the unknown marginal distributions of X and Y. To see the connection between $\widetilde{\Psi}(\boldsymbol{X},\boldsymbol{Y};\alpha)$ in (3.1) and $\Psi(\boldsymbol{X},\boldsymbol{Y};\alpha)$ in (1.6), note that by some preliminary variance bounds, $\mathrm{dCov}^2_*(\boldsymbol{X})$ and $\mathrm{dCov}^2_*(\boldsymbol{Y})$ appearing in the denominator of $\Psi(\boldsymbol{X},\boldsymbol{Y};\alpha)$ in (1.6) will concentrate around their mean values $\mathrm{dCov}^2(X)$ and $\mathrm{dCov}^2(Y)$, respectively (cf. Lemma 13.1). Furthermore, by the mean and variance formula in (1.4) and (1.5) (cf. Theorems 8.4 and 9.12),

$$\sigma^2_{\text{null}}=\frac{2\big(1+\mathfrak{o}(1)\big)}{n^2}\,\mathrm{dCov}^2(X)\,\mathrm{dCov}^2(Y).$$

The above identity implies the asymptotic equivalence between $\widetilde{\Psi}(\boldsymbol{X},\boldsymbol{Y};\alpha)$ in (3.1) and $\Psi(\boldsymbol{X},\boldsymbol{Y};\alpha)$ in (1.6) under the null, showing in particular that $\Psi(\boldsymbol{X},\boldsymbol{Y};\alpha)$ will also have an asymptotic size of α. The rest of the section is devoted to studying the power asymptotics of $\Psi(\boldsymbol{X},\boldsymbol{Y};\alpha)$ and its kernel generalizations.

3.1 *Power universality*

Recall the generalized kernel distance covariance in (2.12) with kernel functions $f=(f_X,f_Y)$ and bandwidth parameters $\gamma=(\gamma_X,\gamma_Y)$. Let the generalized kernel distance correlation test, i.e. a kernelized version of $\Psi(\boldsymbol{X},\boldsymbol{Y};\alpha)$, be defined by

$$\Psi_{f,\gamma}(\boldsymbol{X},\boldsymbol{Y};\alpha)\equiv\mathbf{1}\bigg(\bigg|\frac{n\cdot\mathrm{dCov}^2_*(\boldsymbol{X},\boldsymbol{Y};f,\gamma)}{\sqrt{2\,\mathrm{dCov}^2_*(\boldsymbol{X};f,\gamma)\cdot\mathrm{dCov}^2_*(\boldsymbol{Y};f,\gamma)}}\bigg|>z_{\alpha/2}\bigg).\tag{3.2}$$

The factor n in the definition above is sometimes replaced by $\sqrt{n(n-1)}$ (e.g. [17]), but this will make no difference in the theory below. Using the (local) CLTs derived in Theorem 2.7, the following result gives a unified power expansion formula for the distance correlation test $\Psi(\boldsymbol{X},\boldsymbol{Y};\alpha)$ and the generalized kernel distance correlation test $\Psi_{f,\gamma}(\boldsymbol{X},\boldsymbol{Y};\alpha)$. Its proof can be found in Section 13.

THEOREM 3.1. Suppose that Assumption A holds, and that the spectrum of Σ is contained in $[1/M,M]$ for some $M>1$. Then there exists some constant $C=C(\alpha,M,Z_1)>0$ such that

$$\Big|\mathbb{E}_\Sigma\Psi(\boldsymbol{X},\boldsymbol{Y};\alpha)-\mathbb{P}\big(\big|\mathcal{N}\big(m_n(\Sigma),1\big)\big|>z_{\alpha/2}\big)\Big|\le\frac{C}{(n\wedge p\wedge q)^{1/7}}.$$

Here $m_n(\Sigma)$ can be either

$$\frac{n\,\mathrm{dCov}^2(X,Y)}{\sqrt{2\,\mathrm{dCov}^2(X,X)\,\mathrm{dCov}^2(Y,Y)}}=\frac{n\,\mathrm{dCor}^2(X,Y)}{\sqrt{2}}\text{or}\frac{n\|\Sigma_{XY}\|_F^2}{\sqrt{2}\|\Sigma_X\|_F\|\Sigma_Y\|_F}=\frac{A(\Sigma)}{\sqrt{2}}.$$

If Assumption B holds and ρ_X,ρ_Y are contained in $[1/M,M]$ for some $M>1$, then the above conclusion holds with $\Psi(\boldsymbol{X},\boldsymbol{Y};\alpha)$ replaced by $\Psi_{f,\gamma}(\boldsymbol{X},\boldsymbol{Y};\alpha)$ and C replaced by $C'=C'(\alpha,M,f,Z_1)>0$.

A direct message of the above theorem is that, interestingly, for a large class of kernels $f=(f_X,f_Y)$ and bandwidth parameters $\gamma=(\gamma_X,\gamma_Y)$, the generalized kernel distance correlation test $\Psi_{f,\gamma}(\boldsymbol{X},\boldsymbol{Y};\alpha)$

in (3.2) exhibits exactly the same power behaviour with the distance correlation test $\Psi(\boldsymbol{X}, \boldsymbol{Y}; \alpha)$ in (1.6) in the high-dimensional limit $n \wedge p \wedge q \to \infty$. Here we have focused on deterministic choices of γ_X, γ_Y merely for simplicity of exposition, but following [52], analogous results for data driven choices of γ_X, γ_Y can also be proved with further concentration arguments, for instance for the popular choice

$$\gamma_X \equiv \text{median}\big\{\|X_s - X_t\| : s \neq t\big\}, \quad \gamma_Y \equiv \text{median}\big\{\|Y_s - Y_t\| : s \neq t\big\}.$$

We omit here formal developments along these lines.

The proof of Theorem 3.1 crucially depends on the local CLT in Theorem 2.7. An interesting feature of Theorem 3.1 is that although one may expect that the power formula of the distance correlation test $\Psi(\boldsymbol{X}, \boldsymbol{Y}; \alpha)$ in (1.6) and the generalized kernel distance correlation test $\Psi_{f,\gamma}(\boldsymbol{X}, \boldsymbol{Y}; \alpha)$ in (3.2) involves the complicated expression of the variance $\bar{\sigma}_n^2$ in Theorem 2.2, in fact only the null variance plays a role as in Theorem 2.7. The main reason for this phenomenon to occur is due to the fact that *the regime in which the local central theorem in Theorem 2.7 with the null variance holds covers the entire local contiguity regime* $\limsup A(\Sigma) < \infty$. In other words:

- In the contiguity regime $\limsup A(\Sigma) < \infty$, the ratio of the non-null variance and the null variance is asymptotically 1, cf. (2.15).
- In the large departure regime $A(\Sigma) \to \infty$, both the distance correlation test $\Psi(\boldsymbol{X}, \boldsymbol{Y}; \alpha)$ in (1.6) and the generalized kernel distance correlation test $\Psi_{f,\gamma}(\boldsymbol{X}, \boldsymbol{Y}; \alpha)$ in (3.2) achieve asymptotically full power.

As a result, the 'driving parameter' $n\,\mathrm{dCor}^2(X, Y)/\sqrt{2}$ (or equivalently $A(\Sigma)/\sqrt{2}$) in the power formula for the distance correlation test $\Psi(\boldsymbol{X}, \boldsymbol{Y}; \alpha)$ in (1.6) and its kernelized version $\Psi_{f,\gamma}(\boldsymbol{X}, \boldsymbol{Y}; \alpha)$ in (3.2), inherited from the local CLT in Theorem 2.7 is in a similar form of the test itself, although its proof to reach such a conclusion is far from being obvious.

3.2 *Minimax optimality*

Theorems 3.1 directly implies a separation rate for the (generalized) distance correlation test in the Frobenius norm $\|\cdot\|_F$ in a minimax framework. To formulate this, for any $\zeta > 0$, $M > 1$ and $\Sigma_0 \equiv \mathrm{diag}(\Sigma_X, \Sigma_Y)$, consider the alternative class

$$\Theta(\zeta, \Sigma_0; M) \equiv \Big\{\Sigma = \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{pmatrix} \in \mathbb{R}^{(p+q)\times(p+q)} : \|\Sigma_{XY}\|_F^2 \geq \zeta\sqrt{pq}/n, M^{-1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq M\Big\}.$$

A direct consequence of Theorem 3.1 is the following (for simplicity we only state the result for the distance correlation test $\Psi(\boldsymbol{X}, \boldsymbol{Y}; \alpha)$ in (1.6)).

COROLLARY 3.2. Fix $\alpha \in (0, 1)$. Suppose that Assumption A holds. Then there exists some constant $C = C(\alpha, M, Z_1) > 0$ such that the distance correlation test (1.6) satisfies

$$\sup_{\Sigma \in \Theta(\zeta, \Sigma_0; M)} \big(\mathbb{E}_{\Sigma_0}\Psi(\boldsymbol{X}, \boldsymbol{Y}; \alpha) + \mathbb{E}_{\Sigma}(1 - \Psi(\boldsymbol{X}, \boldsymbol{Y}; \alpha))\big) \leq \alpha + C\bigg[e^{-\zeta^2/C} + \frac{1}{(n \wedge p \wedge q)^{1/7}}\bigg].$$

In particular, the above corollary shows that the distance correlation test (1.6) gives a separation rate in $\|\cdot\|_F$ of order $(pq)^{1/4}/n^{1/2}$, i.e. the testing error (Type I + Type II error) on the left side is

bounded by any prescribed α for $\zeta \to \infty$ in the regime (1.2). In view of the power universality derived in Theorem 3.1, the above results continues to hold when the distance correlation test $\Psi(\boldsymbol{X}, \boldsymbol{Y}; \alpha)$ in (1.6) is replaced by the generalized kernel distance correlation test $\Psi_{f,\gamma}(\boldsymbol{X}, \boldsymbol{Y}; \alpha)$ in (3.2) under the assumption that Assumption B holds and ρ_X, ρ_Y are contained in $[1/M, M]$ for some $M > 1$.

The separation rate $(pq)^{1/4}/n^{1/2}$ in $\|\cdot\|_F$, as will be shown in the following theorem, cannot be improved in a minimax sense. While previous covariance testing literature has mostly focused on the likelihood-ratio test [10,13,23,24,33], this implies that the (generalized) distance correlation tests (1.6) and (3.2) are rate-optimal in this minimax sense. We prove the lower bound in the special case of Gaussian distribution in (2.8).

THEOREM 3.3. Suppose that $Z_1 \overset{d}{=} \mathcal{N}(0,1)$, and $\sqrt{pq}/n \le M$ for some $M > 1$. Then for any small $\delta \in (0,1)$, there exists some positive constant $\zeta = \zeta(\delta, M)$ such that

$$\inf_{\psi} \sup_{\Sigma \in \Theta(\zeta, \Sigma_0; M)} \big(\mathbb{E}_{\Sigma_0} \psi(\boldsymbol{X}, \boldsymbol{Y}) + \mathbb{E}_{\Sigma}(1 - \psi(\boldsymbol{X}, \boldsymbol{Y}))\big) \ge 1 - \delta,$$

where the infimum is taken over all measurable test functions.

The above theorem improves [34, Theorem 1] by requiring $\sqrt{pq}/n \lesssim 1$ rather than $(p+q)/n \lesssim 1$ therein. Note that this improvement in terms of a single condition on $\sqrt{pq}/n$ is particularly compatible with alternative class $\Theta(\zeta, \Sigma_0, M)$ defined above.

The proof of Theorem 3.3 follows a standard minimax reduction in that we only need to find a prior Π on $\Theta(\zeta, \Sigma_0; M)$ with sufficient separation from Σ_0, while at the same time the chi-squared divergence between the posterior density corresponding to Π and the density corresponding to Σ_0 is small. For $\Sigma_0 = I$, the prior Π we construct takes the form

$$\Sigma_{u,v}(a) = \begin{pmatrix} I_p & a\widetilde{u}\widetilde{v}^\top \\ a\widetilde{v}\widetilde{u}^\top & I_q \end{pmatrix},$$

with component-wise independent priors $\widetilde{u}_i \sim_{\text{i.i.d.}} \sqrt{q} \cdot \text{Unif}\{\pm 1\}$ for $\widetilde{u} \in \mathbb{R}^p$ and $\widetilde{v}_j \sim_{\text{i.i.d.}} \sqrt{p} \cdot \text{Unif}\{\pm 1\}$ for $\widetilde{v} \in \mathbb{R}^q$, and some $a > 0$ to be chosen in the end. The calculations of the chi-squared divergence require an exact evaluation of the eigenvalues of certain inverse of $\Sigma_{u,v}(a)$, which eventually leads to a bound of order $a^4 n^2 p^3 q^3$. So under the constraint that the chi-squared distance is bounded by some sufficiently small constant, the maximal choice $a = a_* \asymp n^{-1/2} p^{-3/4} q^{-3/4}$ leads to a minimax separation rate in $\|\cdot\|_F^2$ of the order $\|\Sigma_{u,v}(a_*) - I\|_F^2 \asymp \|a_* \widetilde{u}\widetilde{v}^\top\|_F^2 = a_*^2 \|\widetilde{u}\|^2 \|\widetilde{v}\|^2 = a_*^2 p^2 q^2 \asymp (pq)^{1/2}/n$. Details of the arguments can be found in Section 14 in the supplement.

REMARK 3.4. In Theorem 3.3 above, the growth condition $\sqrt{pq}/n \le M$ for some $M > 1$ is similar to the condition '$p/n \le M$' in the covariance testing literature (e.g. [8]), under which the lower bound construction mentioned above is valid. Whether this condition can be removed (and similarly the condition in [8]) remains open.

4. Simulation studies

In this section, we perform a small-scale simulation study to validate the theoretical results established in previous sections. We consider the balanced case $p = q$ under the following data-generating scheme:

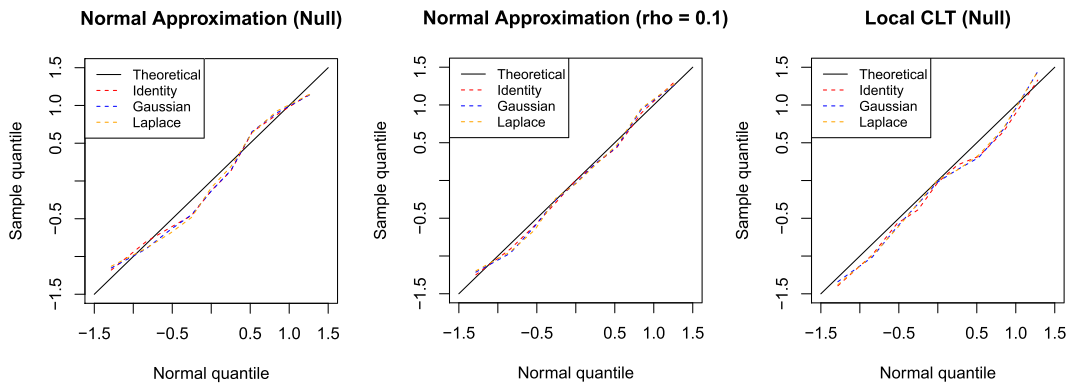

FIG. 1. Verification of CLTs. The solid lines correspond to the standard normal quantiles, and the dashed lines correspond to sample quantiles with the identity, Gaussian and Laplace kernels, respectively. Simulation parameters: $(n, p, q) = (1000, 100, 100)$, $B = 200$ replications, bandwidth choices $\rho_X = \rho_Y = \sqrt{2}$ for both Gaussian and Laplace kernels.

i.i.d. across $j \in [p]$,

$$(X_j, Y_j) \overset{d}{=} (\sqrt{\rho}Z_1 + \sqrt{1-\rho}Z_2, \sqrt{\rho}Z_1 + \sqrt{1-\rho}Z_3), \tag{4.1}$$

where $\rho \in (0, 1)$ is the dependence parameter, and Z_1, Z_2, Z_3 are independent variables with mean zero and variance one. We carry out the simulation primarily in the case where Z_1-Z_3 are standard normal, which is a special case of (2.8) with $\Sigma_X = \Sigma_Y = I_p$ and $\Sigma_{XY} = \rho I_p$. Non-Gaussianity is examined in Fig. 3.

We start by verifying the CLTs derived in Theorems 2.2 and 2.7 in Fig. 1. We take $\rho = 0$ for the null case and $\rho = 0.1$ for the non-null case, and compare the normal quantiles with the corresponding sample quantiles. Normal approximation appears to be accurate in all three cases.

Figure 2 verifies power universality demonstrated via Theorem 3.1 in two aspects: (i) the choice of kernel; (ii) the choice of bandwidth parameters γ_X, γ_Y when using the Gaussian and Laplace kernels. The first two figures illustrates the second point, where the Gaussian and Laplace kernels are used with different bandwidth parameters $\rho_X = \rho_Y \in \{0.5, 1, \sqrt{2}, 5\}$. The third figure uses a fixed bandwidth $\rho_X = \rho_Y = \sqrt{2}$ for both Gaussian and Laplace kernels and compares the performances of different kernels.

Finally we examine the robustness of our theory for non-Gaussian data. We take two choices for Z_1-Z_3 in the set-up (4.1): (i) uniform distribution on $[-\sqrt{3}, \sqrt{3}]$; (ii) t-distribution with four degrees of freedom scaled by $\sqrt{2}$. These parameters are chosen such that Z_1-Z_3 have mean zero and variance one. Normal approximation and power universality are examined in Fig. 3 for the uniform distribution and the (rescaled) t-distribution. These figures suggest that our theory continues to hold for a broader class of data distributions.

5. Concluding remarks and open questions

In this paper, we establish in Theorem 2.2 a general non-null CLT for the sample distance covariance $\mathrm{dCov}^2_*(\boldsymbol{X}, \boldsymbol{Y})$ in the high-dimensional regime $n \wedge p \wedge q \to \infty$ under a separable covariance structure of (X, Y) and a spectral condition on its covariance Σ. The non-null CLT then applies to obtain a first-order

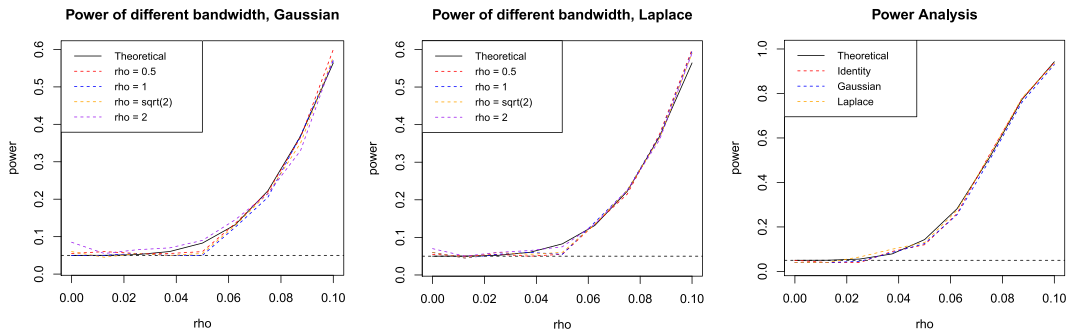

FIG. 2. Verification of power universality in choice of bandwidth parameter (left and middle) and choice of kernel (right). The solid lines correspond to the standard normal quantiles, and the dashed lines correspond to sample quantiles.

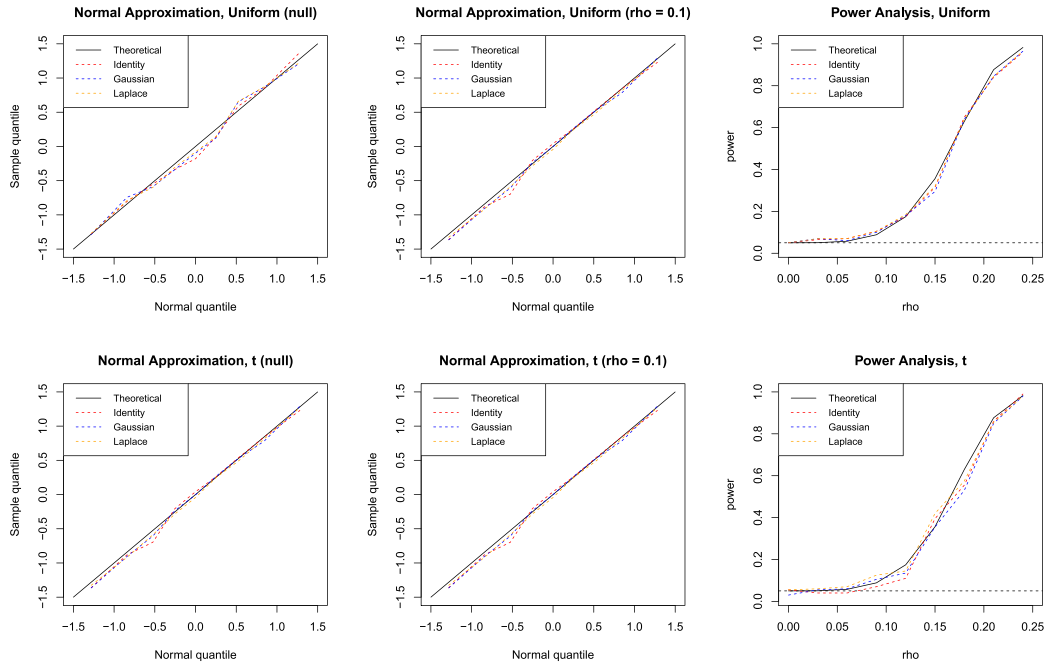

FIG. 3. Verification of CLTs and power expansion for uniform (top three figures) and t- (bottom three figures) distributed data. Simulation parameters: $(n, p, q) = (100, 100, 100)$, $B = 200$ replications, bandwidth choices $\rho_X = \rho_Y = \sqrt{2}$ for both Gaussian and Laplace kernels.

power expansion for the distance correlation test $\Psi(\boldsymbol{X}, \boldsymbol{Y}; \alpha)$ in (1.6):

$$\mathbb{E}\Psi(\boldsymbol{X}, \boldsymbol{Y}; \alpha) \sim \mathbb{P}\left(\left|\mathcal{N}\left(\frac{n\,\mathrm{dCor}^2(X, Y)}{\sqrt{2}}, 1\right)\right| > z_{\alpha/2}\right). \tag{5.1}$$

The non-null CLT and the power expansion (5.1) are also established for a more general class of Hilbert–Schmidt kernel distance covariance, and the associated generalized kernel distance correlation

test $\Psi_{f,\gamma}(\boldsymbol{X},\boldsymbol{Y};\alpha)$ in (3.2), under mild conditions on the kernels and the bandwidth parameters. This result in particular implies that the generalized kernel distance correlation test admits a universal power behaviour with respect to a wide range of choices of kernels and bandwidth parameters.

An important open question is the universality of the power expansion formula (5.1) and the non-null CLT with respect to more general distributions of (X, Y). The first question of power universality, recorded below, is motivated by the fact that the power expansion formula in the form of (5.1) does not explicitly involve any specific form of (X, Y) assumed in (2.8).

PROBLEM 5.1. Establish the universality of (5.1) for general data distributions.

Note that in this paper we established (5.1) by a precise mean and variance expansion for the sample distance covariance $\mathrm{dCov}^2_*(X, Y)$. This is the place where the specific form of the data generating distribution in (2.8) is crucially used. As a preliminary step towards both power asymptotics and non-null CLT, it is therefore of great interest to investigate:

PROBLEM 5.2. Obtain (asymptotic) mean and variance formulae for $\mathrm{dCov}^2_*(\boldsymbol{X},\boldsymbol{Y})$ for general data distributions in the *entire* high-dimensional regime $n \wedge p \wedge q \to \infty$.

In principle, one can obtain 'some' mean and variance formulae for general data distributions by expanding the square root $\|X_\ell - X_k\|, \|Y_\ell - Y_k\|$ to further sufficient many 'higher order terms' in (6.1) below. However, it seems likely this approach will suffer from significant deficiencies in certain regimes within $n \wedge p \wedge q \to \infty$ as a cost of handling the 'residual term' of the highest order. In fact, even with the current distributional form in (2.8), it is already a fairly complicated task to handle the residual terms sharply enough to allow a mean and variance expansion in the entire regime $n \wedge p \wedge q \to \infty$; see Section 6 below for an outline of the complications involved. A new approach may be needed for this problem.

6. Proof road-map of Theorem 2.2

We give a road-map for the proof of the main Theorem 2.2. The basic strategy is to identify the 'main terms' in the Hoeffding decomposition of the fourth-order U-statistics representation in Proposition 2.1. An immediate problem is that the U, V functions in (2.1) involve the square root of the squared ℓ_2 norm which causes differentiability problems. A simple idea is to use the expansion

$$\begin{aligned}\frac{\|X_1 - X_2\|}{\tau_X} &= \left(1 + \frac{\|X_1 - X_2\|^2 - \tau_X^2}{\tau_X^2}\right)^{1/2} \approx \frac{\|X_1 - X_2\|^2 - \tau_X^2}{2\tau_X^2},\\ \frac{\|Y_1 - Y_2\|}{\tau_Y} &= \left(1 + \frac{\|Y_1 - Y_2\|^2 - \tau_Y^2}{\tau_Y^2}\right)^{1/2} \approx \frac{\|Y_1 - Y_2\|^2 - \tau_Y^2}{2\tau_Y^2},\end{aligned} \tag{6.1}$$

as the fluctuation of $\|X_1 - X_2\|^2$ (resp. $\|Y_1 - Y_2\|^2$) around τ_X^2 (resp. τ_Y^2) is expected to be of smaller order than τ_X^2 (resp. τ_Y^2) in high dimensions. Proceeding with this heuristic, with some book-keeping calculations, we may obtain the following approximation of U, V functions:

$$U(x_1, x_2) \approx -x_1^\top x_2/\tau_X, \quad V(y_1, y_2) \approx -y_1^\top y_2/\tau_Y. \tag{6.2}$$

Now if we replace U, V in the k function defined in (2.7) with the above approximation (6.2), the approximate first- and second-order kernels $g_{1,*}, g_{2,*}$ associated with the k function may be computed

as follows:

$$g_{1,*}(x_1, y_1) \equiv \frac{1}{2\tau_X \tau_Y}\Big[\big(x_1^\top \Sigma_{XY} y_1 - \|\Sigma_{XY}\|_F^2\big)\Big],$$
$$g_{2,*}\big((x_1, y_1), (x_2, y_2)\big) \equiv \frac{1}{6\tau_X \tau_Y}\Big[\Big(x_1^\top x_2 y_1^\top y_2 - x_1^\top \Sigma_{XY} y_1 - x_2^\top \Sigma_{XY} y_2 + \|\Sigma_{XY}\|_F^2\Big) - \Big(x_1^\top \Sigma_{XY} y_2 + x_2^\top \Sigma_{XY} y_1\Big)\Big].$$

Although the heuristic so far seems plausible, it turns out that the above approximation falls short of fully capturing the behaviour of the sample distance covariance, even pretending that the effect of higher order kernels can be neglected. In fact, the approximation (6.2) is not good enough, in a somewhat subtle way, in the entire high-dimensional regime $p \wedge q \to \infty$: The first-order kernel $g_{1,*}$ requires the following correction:

$$\widetilde{g}_{1,*}(x_1, y_1) \equiv g_{1,*}(x_1, y_1) - \frac{1}{2\tau_X \tau_Y}\Big[\frac{\|\Sigma_{XY}\|_F^2}{2\tau_X^2}(\|x_1\|^2 - \mathrm{tr}(\Sigma_X)) + \frac{\|\Sigma_{XY}\|_F^2}{2\tau_Y^2}(\|y_1\|^2 - \mathrm{tr}(\Sigma_Y))\Big],$$

whereas, interestingly, no correction is required for the second-order kernel $g_{2,*}$. The underlying reason for the correction terms in the first-order kernel appears to be non-negligible interaction of the approximation of U, V in (6.1), while such interaction is of a strict smaller order in the second-order kernel approximation. In fact, the correction terms in $\widetilde{g}_{1,*}(x_1, y_1)$ above contributes to the difficult terms of negative signs in the first-order variance $\bar{\sigma}_{n,1}^2(X, Y)$ in Theorem 2.2. As a consequence, the variance of $g_{1,*}$ and $\widetilde{g}_{1,*}$ are of the same order but not asymptotically equivalent. Of course, at this point there is no apriori reason to explain why the correction terms must take this form—they come out of exact calculations.

From here, a road-map of the proof of Theorem 2.2 can be outlined:

1. Derive sharp enough estimates for the approximation errors of U, V in (6.2) and their interactions. This will be detailed in Section 8. These sharp enough estimates will immediately give a mean expansion for the sample distance covariance in Theorem 8.4.
2. Using the estimates in (1), validate that the corrected first-order kernel $\widetilde{g}_{1,*}$ and the vanilla second-order kernel $g_{2,*}$ are indeed 'good enough main terms' to approximate the sample distance covariance. This is done via variance considerations detailed in Section 9. As a result, a sharp variance expansion of the sample distance covariance is obtained in Theorem 9.12.
3. Using the mean and variance expansion established in (1)–(2), we establish a non-null CLT for the 'good enough main terms' involving the kernels $\widetilde{g}_{1,*}$ and $g_{2,*}$. The main tool is Chatterjee's discrete second-order Poincaré inequality [9]. This is accomplished in Section 10.

Finally Section 11 assemblies all these steps to complete the proof for Theorem 2.2. In the Section 7, we record some further notations and preliminary results that will be used throughout the proofs.

7. Proof preliminaries

7.1 $G_{[\cdot\cdot]}$ and $H_{[\cdot\cdot]}$

Recall the definition of the matrices $G_{[\cdot\cdot]}, H_{[\cdot\cdot]}$ in (2.11). We summarize some basic properties of these matrices below.

LEMMA 7.1. The following hold.

1. $H_{[11]}^2 = G_{[11]}, H_{[22]}^2 = G_{[22]}, H_{[11]}H_{[22]} = G_{[12]}, H_{[22]}H_{[11]} = G_{[21]}$.
2. $\|G_{[11]}\|_F^2 = \|\Sigma_X^2\|_F^2, \|G_{[22]}\|_F^2 = \|\Sigma_Y^2\|_F^2, \|G_{[12]}\|_F^2 = \|G_{[21]}\|_F^2 = \|\Sigma_{XY}\Sigma_{YX}\|_F^2$.
3. $\mathrm{tr}(G_{[11]}) = \|\Sigma_X\|_F^2, \mathrm{tr}(G_{[22]}) = \|\Sigma_Y\|_F^2, \mathrm{tr}(G_{[12]}) = \mathrm{tr}(G_{[21]}) = \|\Sigma_{XY}\|_F^2$.
4. $\|H_{[11]}\|_F^2 = \|\Sigma_X\|_F^2, \|H_{[22]}\|_F^2 = \|\Sigma_Y\|_F^2, \|H_{[12]}\|_F^2 = \|H_{[21]}\|_F^2 = \|\Sigma_{XY}\|_F^2$.
5. $\mathrm{tr}(H_{[11]}) = \mathrm{tr}(\Sigma_X), \mathrm{tr}(H_{[22]}) = \mathrm{tr}(\Sigma_Y), \mathrm{tr}(G_{[12]}) = \mathrm{tr}(G_{[21]}) = \mathrm{tr}(\Sigma_{XY})$.
6. $\mathrm{tr}(G_{[11]}G_{[22]}) = \mathrm{tr}(G_{[12]}G_{[21]}) = \mathrm{tr}(\Sigma_X\Sigma_{XY}\Sigma_Y\Sigma_{YX})$.

Proof. These claims follow from direct calculations so we omit the details. □

The following lemma will be useful in the second moment part of Proposition 8.2 ahead.

LEMMA 7.2. Suppose that the spectrum of Σ is contained in $[M^{-1}, M]$ for some $M > 1$, then the following hold.

1. $\mathrm{tr}(G_{[12]} \circ G_{[12]}) \vee \mathrm{tr}(H_{[11]}^2 \circ H_{[22]}^2) \vee \|H_{[11]} \circ H_{[22]}\|_F^2 \lesssim_M \|\Sigma_X\|_F^2\|\Sigma_Y\|_F^2/(\tau_X \wedge \tau_Y)^2$.
2. $\|G_{[11]} \circ G_{[22]}\|_F \lesssim_M \|\Sigma_{XY}\|_F^2$.

Proof. (Proof of Lemma 7.2)See Appendix B. □

7.2 *The function h*

Let for $u > -1$

$$h(u) \equiv \sqrt{1+u} - 1 - \frac{u}{2} = -\frac{u^2}{4}\int_0^1 \frac{(1-s)}{(1+su)^{3/2}}\,\mathrm{d}s. \tag{7.1}$$

We summarize below some basic properties of h.

LEMMA 7.3. We have $|h(u)| \lesssim u^2$ and $|h'(u)| \lesssim |u|/(1+u)^{1/2}$. Furthermore,

$$\begin{aligned} h(u) &= -\frac{u^2}{8} + u^3\int_0^1 \frac{3(1-s)^2}{16(1+su)^{5/2}}\,\mathrm{d}s \\ &= -\frac{u^2}{8} + \frac{u^3}{16} - u^4\int_0^1 \frac{5(1-s)^3}{32(1+su)^{7/2}}\,\mathrm{d}s \equiv -\frac{u^2}{8} + h_3(u). \end{aligned}$$

The proof of the above lemma can be found in Appendix B.

7.3 L_X, L_Y and R_X, R_Y

Let

$$L_X(x_1,x_2) \equiv \frac{\|x_1-x_2\|^2-\tau_X^2}{\tau_X^2} \geq -1, \quad R_X(x_1,x_2) \equiv h\big(L_X(x_1,x_2)\big)$$
$$L_Y(y_1,y_2) \equiv \frac{\|y_1-y_2\|^2-\tau_Y^2}{\tau_Y^2} \geq -1, \quad R_Y(y_1,y_2) \equiv h\big(L_Y(y_1,y_2)\big), \tag{7.2}$$

and the double-centred quantities

$$\bar{R}_X(x_1,x_2) \equiv R_X(x_1,x_2) - \mathbb{E}\big[R_X(x_1,X)\big] - \mathbb{E}\big[R_X(X,x_2)\big] + \mathbb{E}\big[R_X(X,X')\big],$$
$$\bar{R}_Y(y_1,y_2) \equiv R_Y(y_1,y_2) - \mathbb{E}\big[R_Y(y_1,Y)\big] - \mathbb{E}\big[R_Y(Y,y_2)\big] + \mathbb{E}\big[R_Y(Y,Y')\big].$$

Using these quantities, we may represent the square root of the Euclidean distance as follows.

LEMMA 7.4. The following hold:

$$\frac{\|x_1-x_2\|}{\tau_X} \equiv 1 + \frac{L_X(x_1,x_2)}{2} + R_X(x_1,x_2) = 1 + \frac{L_X(x_1,x_2)}{2} + h\big(L_X(x_1,x_2)\big),$$
$$\frac{\|y_1-y_2\|}{\tau_Y} \equiv 1 + \frac{L_Y(y_1,y_2)}{2} + R_Y(y_1,y_2) = 1 + \frac{L_Y(y_1,y_2)}{2} + h\big(L_Y(y_1,y_2)\big),$$

and

$$U(x_1,x_2) = -\frac{1}{\tau_X}\Big(x_1^\top x_2 - \tau_X^2 \bar{R}_X(x_1,x_2)\Big),$$
$$V(y_1,y_2) = -\frac{1}{\tau_Y}\Big(y_1^\top y_2 - \tau_Y^2 \bar{R}_Y(y_1,y_2)\Big). \tag{7.3}$$

The following moment estimate will be used repeatedly.

LEMMA 7.5. Suppose that the spectrum of Σ lies in $[M^{-1}, M]$ for some $M > 1$. Fix any positive integer $s \in \mathbb{N}$, there exists some $C = C(s, M, Z_1) > 0$ such that the following moment estimates hold.

1. For any positive integer $s \in \mathbb{N}$,

$$\mathbb{E}L_X^s(X_1,X_2) \lesssim_s \tau_X^{-2s}\|\Sigma_X\|_F^s, \quad \mathbb{E}L_Y^s(Y_1,Y_2) \lesssim_s \tau_Y^{-2s}\|\Sigma_Y\|_F^s.$$

2. For any positive integer $s \in \mathbb{N}$,

$$\mathbb{E}R_X^s(X_1,X_2) \lesssim_s \tau_X^{-4s}\|\Sigma_X\|_F^{2s}, \quad \mathbb{E}R_Y^s(Y_1,Y_2) \lesssim_s \tau_Y^{-4s}\|\Sigma_Y\|_F^{2s}.$$

Consequently the same estimates hold with $\mathbb{E}R_X^s(X_1,X_2), \mathbb{E}R_Y^s(Y_1,Y_2)$ replaced by their double-centred analogues $\mathbb{E}\bar{R}_X^s(X_1,X_2), \mathbb{E}\bar{R}_Y^s(Y_1,Y_2)$.

3. Suppose the spectrum of Σ_X, Σ_Y is contained in $[1/M, M]$ for some $M > 1$. Then for any positive integer $s \in \mathbb{N}$, for $p \wedge q \geq 2s + 1$,

$$\mathbb{E}h'(L_X(X_1, X_2))^s \lesssim_{M,s} \tau_X^{-s}, \quad \mathbb{E}h'(L_Y(Y_1, Y_2))^s \lesssim_{M,s} \tau_Y^{-s}.$$

The proofs of the above lemmas can be found in Appendix B.

8. Residual estimates and mean expansion

8.1 *Residual estimates*

Let

$$\begin{aligned}
\psi_X(x_1, y_1) &\equiv \mathbb{E}_{X_2,Y_2}[\bar{R}_X(x_1, X_2)Y_2^\top y_1], \\
\psi_Y(x_1, y_1) &\equiv \mathbb{E}_{X_2,Y_2}[\bar{R}_Y(y_1, Y_2)X_2^\top x_1], \\
\psi_{X,Y}(x_1, y_1) &\equiv \mathbb{E}_{X_2,Y_2}[\bar{R}_X(x_1, X_2)\bar{R}_Y(y_1, Y_2)].
\end{aligned} \tag{8.1}$$

In view of Lemma 7.4, these terms appear naturally as the interaction error terms when $U(x_1, x_2)V(y_1, y_2)$ is approximated using (6.2). As mentioned in Section 6, sharply controlling these 'residual terms' constitutes the first crucial step in the proof of Theorem 2.2.

First, we have the following representation of $\psi_X(x_1, y_1), \psi_Y(x_1, y_1)$.

LEMMA 8.1. The following decomposition holds:

$$\psi_X(x_1, y_1) = A_{1,X}(x_1, y_1) + A_{2,X}(x_1, y_1),\ \psi_Y(x_1, y_1) = A_{1,Y}(x_1, y_1) + A_{2,Y}(x_1, y_1).$$

Here

$$\begin{aligned}
A_{1,X}(x_1, y_1) &= \frac{1}{2\tau_X^4}\Big[\big(\|x_1\|^2 - \operatorname{tr}(\Sigma_X)\big)x_1^\top \Sigma_{XY} y_1 + 2x_1^\top \Sigma_X \Sigma_{XY} y_1 + \kappa \operatorname{tr}(H_{[11]} \circ Q_X)\Big], \\
A_{1,Y}(x_1, y_1) &= \frac{1}{2\tau_Y^4}\Big[\big(\|y_1\|^2 - \operatorname{tr}(\Sigma_Y)\big)x_1^\top \Sigma_{XY} y_1 + 2x_1^\top \Sigma_{XY} \Sigma_Y y_1 + \kappa \operatorname{tr}(H_{[22]} \circ Q_Y)\Big],
\end{aligned}$$

and

$$A_{2,X}(x_1, y_1) \equiv \mathbb{E}\big[h_3\big(L_X(x_1, X)\big)(Y^\top y_1)\big],\ A_{2,Y}(x_1, y_1) \equiv \mathbb{E}\big[h_3\big(L_Y(y_1, Y)\big)(X^\top x_1)\big],$$

with h_3 defined in Lemma 7.3 and

$$Q_X = H_{[11]} z_1 z_1^\top H_{[22]}, \quad Q_Y = H_{[22]} z_1 z_1^\top H_{[11]}, \tag{8.2}$$

with $z_1 = (x_1^\top, y_1^\top)^\top$ and $H_{[\cdot\cdot]}$ given in (2.11).

PROPOSITION 8.2. Suppose that the spectrum of Σ is contained in $[1/M, M]$ for some $M > 1$, and that p, q are larger than a big enough absolute constant.

1. (First moments) The following hold:

$$\tau_X^4|\mathbb{E}\psi_X(X_1,Y_1)| \bigvee \tau_Y^4|\mathbb{E}\psi_Y(X_1,Y_1)| \bigvee \tau_X^2\tau_Y^2(\tau_X\wedge\tau_Y)|\mathbb{E}\psi_{X,Y}(X_1,Y_1)| \lesssim \|\Sigma_{XY}\|_F^2.$$

2. (Second moments) The following hold:

$$\tau_X^6\mathbb{E}\psi_X^2(X_1,Y_1) \bigvee \tau_Y^6\mathbb{E}\psi_Y^2(X_1,Y_1) \bigvee \tau_X^4\tau_Y^4(\tau_X\wedge\tau_Y)^2\mathbb{E}\psi_{X,Y}^2(X_1,Y_1) \lesssim \|\Sigma_{XY}\|_F^2\big(1\vee\|\Sigma_{XY}\|_F^2\big).$$

The constants in $\lesssim$ only depend on M and the distribution of Z_1 via its Poincaré constant c_* and ε_0 prescribed in Assumption A. The claims remain valid with $X = Y$ when the spectrum of $\Sigma_X = \Sigma_Y$ is contained in $[1/M, M]$ for some $M > 1$.

The role and sharpness of these bounds will be gradually clear in later sections. In particular, these bounds will be essential in the proof of the mean expansion Theorem 8.4 and the variance expansion Theorem 9.12 ahead.

Note that here the first moment bounds in Proposition 8.2 do not follow directly by the stated second moment bounds, as the 'first moments' here are obtained by first taking expectation followed by the absolute value. In fact, these first moment estimates are stronger by those derived directly from the second moment estimates, indicating the essential role of the order of taking expectation and absolute value in this setting.

An important feature of the bounds in Proposition 8.2 above is that when $\Sigma_{XY} = 0$, all estimates reduce to 0. Furthermore the exponent in $\|\Sigma_{XY}\|_F, \tau_X, \tau_Y$ also need be correct to allow precise mean and variance expansions in Theorems 8.4 and 9.12, and therefore the non-null CLT in Theorem 2.2, under the entire high-dimensional regime $n \wedge p \wedge q \to \infty$. It is for this reason that the proof of Proposition 8.2 is rather involved, the details of which can be found in Appendix C. The following lemma is representative in terms of an interpolation technique in proving Proposition 8.2 and may be of broader interest.

LEMMA 8.3. Suppose the spectrum of Σ is contained in $[1/M, M]$ for some $M > 1$. Let $\mathfrak{h}_X, \mathfrak{h}_Y : \mathbb{R} \to \mathbb{R}$ be smooth functions. For any $k, k', \ell, \ell' \in \{1, 2\}$, define

$$\psi_{\mathfrak{h}_X,\mathfrak{h}_Y}(\Sigma_{XY}) \equiv \mathbb{E}\Big[\mathfrak{h}_X\big(L_X(X_k, X_\ell)\big)\mathfrak{h}_Y\big(L_Y(Y_{k'}, Y_{\ell'})\big)\big|X_1, Y_1\Big].$$

Then

$$\mathbb{E}\big(\psi_{\mathfrak{h}_X,\mathfrak{h}_Y}(\Sigma_{XY}) - \psi_{\mathfrak{h}_X,\mathfrak{h}_Y}(0)\big)^2 \lesssim \|\Sigma_{XY}\|_F^2(1\vee\|\Sigma_{XY}\|_F^2) \times \bigg(\tau_X^{-4}\cdot\mathbb{E}^{1/4}(\mathfrak{h}_X'\circ L_X)^8\cdot\mathbb{E}^{1/4}(\mathfrak{h}_Y\circ L_Y)^8 + \tau_Y^{-4}\cdot\mathbb{E}^{1/4}(\mathfrak{h}_Y'\circ L_Y)^8\cdot\mathbb{E}^{1/4}(\mathfrak{h}_X\circ L_X)^8\bigg).$$

The constants in $\lesssim$ only depend on M. The claims remain valid with $X = Y$ when the spectrum of $\Sigma_X = \Sigma_Y$ is contained in $[1/M, M]$ for some $M > 1$.

Lemma 8.3 gives a general recipe of bounding the second moment, in terms of the dependence measure $\|\Sigma_{XY}\|_F$. Details see Appendix C.

8.2 *Mean expansion*

As a quick application of the residual estimates in Proposition 8.2, we may get the following mean expansion.

THEOREM 8.4. Suppose that the spectrum of Σ is contained in $[1/M, M]$ for some $M > 1$. Then the following expansion holds for the distance covariance:

$$m_\Sigma = \mathbb{E}_\Sigma \, \mathrm{dCov}_*^2(\boldsymbol{X}, \boldsymbol{Y}) = \mathrm{dCov}^2(X, Y) = \frac{\|\Sigma_{XY}\|_F^2}{\tau_X \tau_Y}\Big[1 + \mathscr{O}\big((\tau_X \wedge \tau_Y)^{-1}\big)\Big].$$

The constants in $\mathscr{O}$ only depend on M and the distribution of Z_1 via its Poincaré constant c_* and ε_0 prescribed in Assumption A. The claims remain valid with $X = Y$ when the spectrum of $\Sigma_X = \Sigma_Y$ is contained in $[1/M, M]$ for some $M > 1$.

Proof. Note that

$$\begin{aligned}
&\mathrm{dCov}^2(X, Y) = \mathbb{E}\big[U(X_1, X_2)V(Y_1, Y_2)\big] \\
&= \frac{1}{\tau_X \tau_Y}\Big[\|\Sigma_{XY}\|_F^2 - \tau_X^2 \mathbb{E}\big(\bar{R}_X(X_1, X_2) Y_1^\top Y_2\big) \\
&\qquad - \tau_X^2 \mathbb{E}\big(\bar{R}_Y(Y_1, Y_2) X_1^\top X_2\big) + \tau_X^2 \tau_Y^2 \mathbb{E}\big(\bar{R}_X(X_1, X_2)\bar{R}_Y(Y_1, Y_2)\big)\Big] \\
&\equiv \frac{1}{\tau_X \tau_Y}\Big[\|\Sigma_{XY}\|_F^2 - \tau_X^2 \mathbb{E}\psi_X(X_1, X_2) - \tau_Y^2 \mathbb{E}\psi_Y(Y_1, Y_2) + \tau_X^2 \tau_Y^2 \mathbb{E}\psi_{X,Y}(X_1, Y_1)\Big].
\end{aligned}$$

The claim now follows by invoking Proposition 8.2-(1). □

A stochastic version of the above theorem was previously derived in [52, Theorem 2.1.1], where the main term $\|\Sigma_{XY}\|_F^2/(\tau_X \tau_Y)$ was replaced by an unbiased estimator and the remainder term was controlled at the order $(p \wedge q)^{-1/2}$. In comparison, due to the sharp residual estimates in Proposition 8.2, our bound for the remainder term is much more refined in that it contains an important multiplicative factor $\|\Sigma_{XY}\|_F^2$, which makes it asymptotically negligible in the null case as well.

9. Hoeffding decomposition and variance expansion

We first review the basics of Hoeffding decomposition that will be relevant to our purpose. Following [38, Section 5.1.5, pp. 177], for a generic fourth-order U-statistic with symmetric kernel $k : \mathscr{Z}^4 \to \mathbb{R}$, let

$$k_c(z_1, \ldots, z_c) \equiv \mathbb{E}[k(z_1, \ldots, z_c, Z_{c+1}, \ldots, Z_4)], \quad z_1, \ldots, z_c \in \mathscr{Z},$$

and for any $z_1, z_2, z_3, z_4 \in \mathscr{Z}$,

$$\begin{aligned}
g_0 &\equiv \mathbb{E}k(\mathbf{Z}), \quad g_1(z_1) \equiv k_1(z_1) - \mathbb{E}k(\mathbf{Z}),\\
g_2(z_1, z_2) &\equiv k_2(z_1, z_2) - k_1(z_1) - k_2(z_2) + \mathbb{E}k(\mathbf{Z}),\\
g_3(z_1, z_2, z_3) &\equiv k_3(z_1, z_2, z_3) - \mathbb{E}k(\mathbf{Z}) - \sum_{\ell=1}^{3} g_1(z_\ell) - \sum_{1\le \ell_1<\ell_2\le 3} g_2\big(z_{\ell_1}, z_{\ell_2}\big),\\
g_4(z_1, z_2, z_3, z_4) &\equiv k_4\big(z_1, z_2, z_3, z_4\big) - \mathbb{E}k(\mathbf{Z}) - \sum_{\ell=1}^{3} g_1(z_\ell)\\
&\quad - \sum_{1\le \ell_1<\ell_2\le 3} g_2\big(z_{\ell_1}, z_{\ell_2}\big) - \sum_{1\le \ell_1<\ell_2<\ell_3\le 4} g_3\big(z_{\ell_1}, z_{\ell_2}, z_{\ell_3}\big). \qquad (9.1)
\end{aligned}$$

Then the Hoeffding decomposition says that

$$U_n(k) = \sum_{c=0}^{4} \binom{4}{c} U_n(g_c).$$

Here for a generic symmetric kernel $g : \mathscr{Z}^c \to \mathbb{R}$,

$$U_n(g) \equiv \begin{cases} \binom{n}{c}^{-1} \sum_{i_1<\ldots<i_c} g\big(z_{i_1}, \ldots, z_{i_c}\big), & c \ge 1;\\ g & c = 0. \end{cases}$$

For $c = 0$, g is understood as a real number. In what follows, we will take $\mathscr{Z} \equiv \mathscr{X} \times \mathscr{Y}$, and k as the kernel defined in Proposition 2.1. We will evaluate the variance of $\mathrm{dCov}_*^2(\boldsymbol{X}, \boldsymbol{Y}) = U_n(k)$ by evaluating the variance of g_1, g_2, g_3, g_4 associated with k as defined above.

9.1 *Hoeffding decomposition: first order*

The goal of this subsection is to prove the following variance expansion for the first-order kernel associated with k.

Proposition 9.1. Suppose the spectrum of Σ is contained in $[1/M, M]$ for some $M > 1$. Then for any $\varepsilon > 0$, the first-order variance is given by

$$\binom{4}{1}^2 \binom{n}{1}^{-1} \mathbb{E}g_1^2(X_1, Y_1) = (1 \pm \varepsilon) \cdot \bar{\sigma}_{n,1}^2(X, Y) \cdot \bigg[1 + \mathscr{O}\Big(\frac{1}{\varepsilon \cdot (\tau_X \wedge \tau_Y)}\Big)\bigg].$$

Here $\bar{\sigma}_{n,1}^2(X, Y)$ is as defined in Theorem 2.2, and the constants in $\mathscr{O}$ only depend on M and the distribution of Z_1 only via its Poincaré constant c_*, excess kurtosis κ and ε_0 prescribed in Assumption A. The claim remains valid with $X = Y$ when the spectrum of $\Sigma_X = \Sigma_Y$ is contained in $[1/M, M]$ for some $M > 1$.

The proof of the above proposition will be presented towards the end of this subsection. First, we may compute:

LEMMA 9.2. The first-order kernel is given by

$$k_1(z_1) = \mathbb{E}k(z_1, Z_2, Z_3, Z_4) = \frac{1}{2}\Big[\mathbb{E}U(x_1, X)V(y_1, Y) + \mathrm{dCov}^2(X, Y)\Big],$$
$$g_1(z_1) = k_1(z_1) - \mathbb{E}k(Z) = \frac{1}{2}\Big[\mathbb{E}U(x_1, X)V(y_1, Y) - \mathrm{dCov}^2(X, Y)\Big].$$

We will use the above lemma to devise an expansion for g_1. From the approximation of U, V in (6.2), one may hope that the main term for g_1 would be $2^{-1}\mathbb{E}U(x_1, X)V(y_1, Y) \approx (x_1^\top \Sigma_{XY} y_1 - \|\Sigma_{XY}\|_F^2)/2\tau_X\tau_Y$. As announced in Section 6, this is however not the case. Let the 'main term' be defined as

$$\bar{g}_1(x_1, y_1) \equiv \frac{1}{2\tau_X\tau_Y}\Big[\big(x_1^\top \Sigma_{XY} y_1 - \|\Sigma_{XY}\|_F^2\big) + \mathscr{A}_{1,X}(x_1, y_1) + \mathscr{A}_{1,Y}(x_1, y_1)\Big], \tag{9.2}$$

where

$$\mathscr{A}_{1,X}(x_1, y_1) \equiv -\frac{\|\Sigma_{XY}\|_F^2}{2\tau_X^2}\big(\|x_1\|^2 - \mathrm{tr}(\Sigma_X)\big), \quad \mathscr{A}_{1,Y}(x_1, y_1) \equiv -\frac{\|\Sigma_{XY}\|_F^2}{2\tau_Y^2}\big(\|y_1\|^2 - \mathrm{tr}(\Sigma_Y)\big).$$

The terms $\mathscr{A}_{1,X}(x_1, y_1), \mathscr{A}_{1,Y}(x_1, y_1)$ are essential to correct the naive approximation (6.2), in that these terms contribute to the somewhat difficult terms of negative sign in the variance expansion of $\bar{g}_1$ in Lemma 9.4, which cannot be neglected as they may have the same order as that of the leading terms.

With the main term defined above, let the 'residual term' be defined by

$$\bar{R}_1(x_1, y_1) \equiv -\tau_X^2\bar{\psi}_X(x_1, y_1) - \tau_Y^2\bar{\psi}_Y(x_1, y_1) + \tau_X^2\tau_Y^2\psi_{X,Y}(x_1, y_1), \tag{9.3}$$

where $\psi_{X,Y}$ is defined in (8.1), and

$$\begin{aligned}\bar{\psi}_X(x_1, y_1) \equiv \frac{1}{2\tau_X^4}\Big[&\big(\|x_1\|^2 - \mathrm{tr}(\Sigma_X)\big)\big(x_1^\top \Sigma_{XY} y_1 - \|\Sigma_{XY}\|_F^2\big) + 2x_1^\top \Sigma_X \Sigma_{XY} y_1 \\ &+ \kappa\,\mathrm{tr}(H_{[11]} \circ Q_X)\Big] + A_{2,X}(x_1, y_1), \\ \bar{\psi}_Y(x_1, y_1) \equiv \frac{1}{2\tau_Y^4}\Big[&\big(\|y_1\|^2 - \mathrm{tr}(\Sigma_Y)\big)\big(x_1^\top \Sigma_{XY} y_1 - \|\Sigma_{XY}\|_F^2\big) + 2x_1^\top \Sigma_{XY} \Sigma_Y y_1 \\ &+ \kappa\,\mathrm{tr}(H_{[22]} \circ Q_Y)\Big] A_{2,Y}(x_1, y_1).\end{aligned} \tag{9.4}$$

Here $Q_X, Q_Y, A_{2,X}, A_{2,Y}$ are defined in Lemma 8.1. Using $\bar{g}_1, \bar{R}_1$ defined above, we may expand g_1 into the sum of main and residual terms as follows.

LEMMA 9.3. The following expansion holds:

$$g_1(x_1,y_1)=\bar{g}_1(x_1,y_1)+\frac{1}{2\tau_X\tau_Y}\big(\bar{R}_1(x_1,y_1)-\mathbb{E}\bar{R}_1(X_1,Y_1)\big). \tag{9.5}$$

Now we will evaluate the variance of $\bar{g}_1$ and $\bar{R}_1$. The variance of $\bar{g}_1$ is given by the following.

LEMMA 9.4. We have $\mathbb{E}\bar{g}_1^2(X,Y)=4^{-2}n\cdot\bar{\sigma}_{n,1}^2$, where $\bar{\sigma}_{n,1}^2$ is given in Theorem 2.2.

As mentioned above, the variance of $\bar{g}_1$ as above involves terms with a negative sign that are contributed by the 'correction terms' $\mathscr{A}_{1,X}(x_1,y_1),\mathscr{A}_{1,Y}(x_1,y_1)$. These terms can be of the same order as the main terms. It is therefore important to have a lower bound on this quantity.

LEMMA 9.5.

1. Suppose $\|\Sigma^{-1}\|_{\mathrm{op}}\le M$ for some $M>1$. Then $\mathbb{E}\bar{g}_1^2(X,Y)\gtrsim\tau_X^{-2}\tau_Y^{-2}\|\Sigma_{XY}\|_F^2$. If furthermore $\|\Sigma\|_{\mathrm{op}}\le M$, then $\mathbb{E}\bar{g}_1^2(X,Y)\asymp\tau_X^{-2}\tau_Y^{-2}\|\Sigma_{XY}\|_F^2$.
2. Suppose $X=Y$, and $\|\Sigma_X^{-1}\|_{\mathrm{op}}\le M$ for some $M>1$. Then $\mathbb{E}\bar{g}_1^2(X,X)\gtrsim\tau_X^{-8}\cdot p\|\Sigma_X\|_F^4$. If furthermore $\|\Sigma_X\|_{\mathrm{op}}\le M$, then $\mathbb{E}\bar{g}_1^2(X,X)\asymp\tau_X^{-2}$.

The constants in $\gtrsim,\asymp$ only depend on M and the distribution of Z_1 via its excess kurtosis κ in Assumption A.

Lemma 9.5 above is an important result, showing that the negative contributions of the 'correction terms' $\mathscr{A}_{1,X}(x_1,y_1),\mathscr{A}_{1,Y}(x_1,y_1)$ will not affect the order the variance $\bar{g}_1$. In other words, these terms will contribute a non-vanishing but small proportion of the main terms.

Next to the variance of the main term $\bar{g}_1$, an important step to obtain variance bound for the residual term $\bar{R}_1$ is to obtain variance bounds for $\bar{\psi}_X,\bar{\psi}_Y$ defined in (9.4).

LEMMA 9.6. Suppose that the spectrum of Σ_X,Σ_Y is contained in $[1/M,M]$ for some $M>1$. Then

$$\tau_X^6\operatorname{Var}\big(\bar{\psi}_X(X,Y)\big)\bigvee\tau_Y^6\operatorname{Var}\big(\bar{\psi}_Y(X,Y)\big)\lesssim_{M,Z_1}\|\Sigma_{XY}\|_F^2.$$

Here the dependence of $\lesssim$ on Z_1 is via its Poincaré constant c_* and ε_0 prescribed by Assumption A.

This variance bound plays an important role to keep the residual terms small when $\|\Sigma_{XY}\|_F$ is large. In particular, if one uses the vanilla versions ψ_X,ψ_Y defined in (8.1), the right-hand side of the above display scales as $\|\Sigma_{XY}\|_F^4$ that would lead to essential difficulties in controlling the residuals. In other words, the reduction from $\|\Sigma_{XY}\|_F^4$ to $\|\Sigma_{XY}\|_F^2$ is made possible by the 'correction terms' $\mathscr{A}_{1,X}(x_1,y_1),\mathscr{A}_{1,Y}(x_1,y_1)$ that, in a certain sense, 'centre' the vanilla versions ψ_X,ψ_Y to reduce the variance.

Detailed proofs of Lemmas 9.2–9.6 are deferred to Appendix D. Now we are in a good position to prove Proposition 9.1.

Proof. (Proof of Proposition 9.1)By (9.3),

$$\begin{aligned}&\operatorname{Var}\big(\bar{R}_1(X_1,Y_1)\big)\\&\lesssim\tau_X^4\operatorname{Var}\big(\bar{\psi}_X(X_1,Y_1)\big)+\tau_Y^4\operatorname{Var}\big(\bar{\psi}_Y(X_1,Y_1)\big)+\tau_X^4\tau_Y^4\mathbb{E}\psi_{X,Y}^2(X_1,Y_1).\end{aligned}$$

The first two terms can be handled by Lemma 9.6, while the last term can be bounded by

$$\mathbb{E}\psi_{X,Y}^2(X_1,Y_1) \lesssim_M \tau_X^{-4}\tau_Y^{-4}\left[\frac{\|\Sigma_{XY}\|_F^2+\|\Sigma_{XY}\|_F^4}{(\tau_X\wedge\tau_Y)^2}\bigwedge 1\right].$$

This follows by Proposition 8.2-(2) and the simple bound

$$\mathbb{E}\psi_{X,Y}^2(X_1,Y_1) \le \mathbb{E}\bar{R}_X^2\cdot\mathbb{E}\bar{R}_Y^2 \lesssim_M \tau_X^{-4}\tau_Y^{-4}$$

using Lemma 7.5. Summarizing the estimates, we have

$$\operatorname{Var}\big(\bar{R}_1(X_1,Y_1)\big) \lesssim \frac{\|\Sigma_{XY}\|_F^2}{(\tau_X\wedge\tau_Y)^2}+\frac{\|\Sigma_{XY}\|_F^4}{(\tau_X\wedge\tau_Y)^2}\bigwedge 1.$$

As $\operatorname{Var}(g_1)=(1\pm\varepsilon)\operatorname{Var}(\bar{g}_1)+\mathscr{O}\big(\varepsilon^{-1}\cdot\tau_X^{-2}\tau_Y^{-2}\operatorname{Var}(\bar{R}_1(X,Y)\big)$ for any $\varepsilon>0$, the proof is now complete by noting that

$$\frac{\operatorname{Var}(\bar{R}_1(X,Y))}{\tau_X^2\tau_Y^2\operatorname{Var}(\bar{g}_1)} \lesssim_M \frac{\frac{\|\Sigma_{XY}\|_F^2}{(\tau_X\wedge\tau_Y)^2}+\frac{\|\Sigma_{XY}\|_F^4}{(\tau_X\wedge\tau_Y)^2}\bigwedge 1}{\|\Sigma_{XY}\|_F^2} \lesssim \frac{1}{\tau_X\wedge\tau_Y},$$

using Lemma 9.5 in the first inequality. □

9.2 *Hoeffding decomposition: second order*

The goal of this subsection is to prove the following variance expansion for the second-order kernel associated with k.

PROPOSITION 9.7. Suppose that the spectrum of Σ is contained in $[1/M,M]$ for some $M>1$. For any $\varepsilon>0$, the second-order variance is given by

$$\binom{4}{2}^2\binom{n}{2}^{-1}\mathbb{E}g_2^2\big((X_1,Y_1),(X_2,Y_2)\big)=(1\pm\varepsilon)\cdot\bar{\sigma}_{n,2}^2(X,Y)\cdot\Big[1+\mathscr{O}\Big(\frac{1}{\varepsilon\cdot(\tau_X\wedge\tau_Y)^2}\Big)\Big].$$

Here $\bar{\sigma}_{n,2}^2(X,Y)$ is as defined in Theorem 2.2, the constant in $\mathscr{O}$ depends on M and the distribution of Z_1 only via its Poincaré constant c_*, excess kurtosis κ and ε_0 prescribed by Assumption A. The claim remains valid with $X=Y$ when the spectrum of $\Sigma_X=\Sigma_Y$ is contained in $[1/M,M]$ for some $M>1$.

The prove Proposition 9.7, we will first get an expansion for g_2, which requires a calculation of k_2:

LEMMA 9.8. The second-order kernel is given by

$$\begin{aligned} k_2(z_1,z_2) &= \mathbb{E}k(z_1,z_2,Z_3,Z_4) \\ &= \frac{1}{6}\Big[U(x_1,x_2)V(y_1,y_2) + 2\mathbb{E}U(x_1,X)V(y_1,Y) + 2\mathbb{E}U(x_2,X)V(y_2,Y) + \mathrm{dCov}^2(X,Y) \\ &\qquad - \mathbb{E}U(x_1,X)V(y_2,Y) - \mathbb{E}U(x_2,X)V(y_1,Y)\Big]. \end{aligned}$$

We will use the above lemma to devise an expansion for g_2. In the first-order expansion in the previous subsection, we have seen that the approximation of U, V in (6.2) is *not* enough to get a precise variance expansion of g_1. Somewhat interestingly, as announced in Section 6 such approximation is good enough in the second-order expansion. Formally, let the 'main term' of g_2 be defined by

$$\begin{aligned} \bar{g}_2\big((x_1,y_1),(x_2,y_2)\big) \equiv \frac{1}{6\tau_X\tau_Y}\Big[&\Big(x_1^\top x_2 y_1^\top y_2 - x_1^\top \Sigma_{XY} y_1 - x_2^\top \Sigma_{XY} y_2 + \|\Sigma_{XY}\|_F^2\Big) \\ &- \Big(x_1^\top \Sigma_{XY} y_2 + x_2^\top \Sigma_{XY} y_1\Big)\Big], \end{aligned} \tag{9.6}$$

and the 'residual term' be defined by [recall the definitions of $\bar{R}_X, \bar{R}_Y$ after (7.2)]

$$\begin{aligned} &\bar{R}_2\big((x_1,y_1),(x_2,y_2)\big) \\ &= -\tau_Y^2 x_1^\top x_2 \bar{R}_Y(y_1,y_2) - \tau_X^2 y_1^\top y_2 \bar{R}_X(x_1,x_2) + \tau_X^2\tau_Y^2 \bar{R}_X(x_1,x_2)\bar{R}_Y(y_1,y_2) \\ &\quad - R_1(x_1,y_1) - R_1(x_2,y_2) - R_1(x_1,y_2) - R_1(x_2,y_1), \end{aligned}$$

with R_1 defined by [recall the definitions of $\psi_X, \psi_Y, \psi_{X,Y}$ in (8.1)]

$$R_1(x_1,y_1) \equiv -\tau_X^2\psi_X(x_1,y_1) - \tau_Y^2\psi_Y(x_1,y_1) + \tau_X^2\tau_Y^2\psi_{X,Y}(x_1,y_1). \tag{9.7}$$

The following lemma gives an expansion of g_2 into the sum of the main term $\bar{g}_2$ and the centred residual term $\bar{R}_2$.

LEMMA 9.9. The following expansion holds:

$$\begin{aligned} &g_2\big((x_1,y_1),(x_2,y_2)\big) \\ &= \bar{g}_2\big((x_1,y_1),(x_2,y_2)\big) + \frac{1}{6\tau_X\tau_Y}\Big[\bar{R}_2\big((x_1,y_1),(x_2,y_2)\big) - \mathbb{E}\bar{R}_2\big((X_1,Y_1),(X_2,Y_2)\big)\Big]. \end{aligned}$$

Proofs of the proceeding lemmas can be found in Appendix D. Using the above decomposition, we only need to compute the variance for the two terms $\bar{g}_2, \bar{R}_2$ on the right-hand side of the above display for the proof of Proposition 9.7. Clearly the variance of $\bar{g}_2$ can be evaluated by a book-keeping calculation, and the variance of $\bar{R}_2$ can be handled by the residual estimates in Proposition 8.2. The proof below illustrates the strength of the bounds obtained in Proposition 8.2.

Proof. (Proof of Proposition 9.7) First note that we may expand $36\tau_X^2\tau_Y^2\mathbb{E}\bar{g}_2^2\big((X_1,Y_1),(X_2,Y_2)\big)$ as

$$\mathbb{E}\Big[X_1^\top X_2 Y_1^\top Y_2 - X_1^\top \Sigma_{XY} Y_1 - X_2^\top \Sigma_{XY} Y_2 + \|\Sigma_{XY}\|_F^2\Big]^2 + \mathbb{E}\Big[X_1^\top \Sigma_{XY} Y_2 + X_2^\top \Sigma_{XY} Y_1\Big]^2 \equiv E_1 + E_2.$$

This follows as the cross term has expectation 0 due to the symmetry of the distribution of (X, Y). After expansion, we have

$$\begin{aligned} E_1 = {} & \mathbb{E}(X_1^\top X_2 Y_1^\top Y_2)^2 + 2\mathbb{E}(X_1^\top \Sigma_{XY} Y_1)^2 + \|\Sigma_{XY}\|_F^4 - 4\mathbb{E}(X_1^\top X_2 Y_1^\top Y_2 X_1^\top \Sigma_{XY} Y_1) \\ & + 2\|\Sigma_{XY}\|_F^4 + 2\mathbb{E}(X_1^\top \Sigma_{XY} Y_1 X_2^\top \Sigma_{XY} Y_2) - 4\|\Sigma_{XY}\|_F^4. \end{aligned}$$

The above terms can be calculated using Lemma G.3:

- The first term is

$$\begin{aligned} \mathbb{E}(X_1^\top X_2 Y_1^\top Y_2)^2 = {} & 2\Big(\|\Sigma_{XY}\|_F^4 + \operatorname{tr}(\Sigma_{XY}\Sigma_{YX}\Sigma_{XY}\Sigma_{YX}) + 2\operatorname{tr}(\Sigma_{XY}\Sigma_Y\Sigma_{YX}\Sigma_X)\Big) \\ & + \|\Sigma_X\|_F^2\|\Sigma_Y\|_F^2 + 2\kappa\cdot\big[2\operatorname{tr}(G_{[12]}\circ G_{[12]}) + \operatorname{tr}(H_{[11]}^2\circ H_{[22]}^2)\big] \\ & + \kappa^2\cdot\|H_{[11]}\circ H_{[22]}\|_F^2. \end{aligned}$$

- The second term is

$$2\mathbb{E}(X_1^\top \Sigma_{XY} Y_1)^2 = 2\Big(\|\Sigma_{XY}\|_F^4 + \|\Sigma_{XY}\Sigma_{YX}\|_F^2 + \operatorname{tr}(\Sigma_{XY}\Sigma_Y\Sigma_{YX}\Sigma_X) + \kappa\cdot\operatorname{tr}(G_{[12]}\circ G_{[12]})\Big).$$

- The fourth term is

$$\begin{aligned} & -4\mathbb{E}(X_1^\top X_2 Y_1^\top Y_2 X_1^\top \Sigma_{XY} Y_1) = -4\mathbb{E}(X_1^\top \Sigma_{XY} Y_1)^2 \\ & = -4\Big(\|\Sigma_{XY}\|_F^4 + \|\Sigma_{XY}\Sigma_{YX}\|_F^2 + \operatorname{tr}(\Sigma_{XY}\Sigma_Y\Sigma_{YX}\Sigma_X) + \kappa\cdot\operatorname{tr}(G_{[12]}\circ G_{[12]})\Big). \end{aligned}$$

- The sixth term is $2\mathbb{E}(X_1^\top \Sigma_{XY} Y_1 X_2^\top \Sigma_{XY} Y_2) = 2\|\Sigma_{XY}\|_F^4$.

  In summary, we have

$$\begin{aligned} E_1 = {} & \|\Sigma_X\|_F^2\|\Sigma_Y\|_F^2 + \|\Sigma_{XY}\|_F^4 + 2\operatorname{tr}(\Sigma_{XY}\Sigma_Y\Sigma_{YX}\Sigma_X) \\ & + 2\kappa\cdot\big(\operatorname{tr}(G_{[12]}\circ G_{[12]}) + \operatorname{tr}(H_{[11]}^2\circ H_{[22]}^2)\big) + \kappa^2\cdot\|H_{[11]}\circ H_{[22]}\|_F^2. \end{aligned}$$

Similarly, we have

$$E_2 = 2\mathbb{E}(X_1^\top \Sigma_{XY} Y_2)^2 + 2\mathbb{E}(X_1^\top \Sigma_{XY} Y_2 X_2^\top \Sigma_{XY} Y_1) = 2\operatorname{tr}(\Sigma_{XY}\Sigma_Y\Sigma_{YX}\Sigma_X) + 2\|\Sigma_{XY}\Sigma_{YX}\|_F^2.$$

Combining the identities and applying Lemmas 7.2 and G.5 yields that

$$\begin{aligned}
&\mathbb{E}\bar{g}_2^2\big((X_1,Y_1),(X_2,Y_2)\big)\\
&=\frac{1}{36\tau_X^2\tau_Y^2}\Big[\Big(\|\Sigma_X\|_F^2\|\Sigma_Y\|_F^2+4\operatorname{tr}(\Sigma_{XY}\Sigma_Y\Sigma_{YX}\Sigma_X)+\|\Sigma_{XY}\|_F^4+2\|\Sigma_{XY}\Sigma_{YX}\|_F^2\Big)\\
&\qquad+2\kappa\cdot\big(\operatorname{tr}(G_{[12]}\circ G_{[12]})+\operatorname{tr}(H_{[11]}^2\circ H_{[22]}^2)\big)+\kappa^2\cdot\|H_{[11]}\circ H_{[22]}\|_F^2\Big]\\
&=\frac{1}{36\tau_X^2\tau_Y^2}\Big(\|\Sigma_X\|_F^2\|\Sigma_Y\|_F^2+\|\Sigma_{XY}\|_F^4\Big)\Big[1+\mathscr{O}_{M,\kappa}\Big(\frac{1}{(\tau_X\wedge\tau_Y)^2}\Big)\Big].
\end{aligned}$$

For the residual term, it can be bounded as follows:

$$\begin{aligned}
\mathbb{E}\bar{R}_2^2\big((X_1,Y_1),(X_2,Y_2)\big)\lesssim\tau_Y^4\mathbb{E}^{1/2}(X_1^\top X_2)^4\cdot\mathbb{E}^{1/2}\bar{R}_Y^4+\tau_X^4\mathbb{E}^{1/2}(Y_1^\top Y_2)^4\cdot\mathbb{E}^{1/2}\bar{R}_X^4\\
+\tau_X^4\tau_Y^4\mathbb{E}^{1/2}\bar{R}_X^4\cdot\mathbb{E}^{1/2}\bar{R}_Y^4+\mathbb{E}R_1^2.
\end{aligned}$$

Using Proposition 8.2-(2), it follows that

$$\mathbb{E}R_1^2\lesssim\tau_X^4\mathbb{E}\psi_X^2+\tau_Y^4\mathbb{E}\psi_Y^2+\tau_X^4\tau_Y^4\mathbb{E}\psi_{X,Y}^2\lesssim_M\frac{\|\Sigma_{XY}\|_F^2(1\vee\|\Sigma_{XY}\|_F^2)}{(\tau_X\wedge\tau_Y)^2}.$$

This, combined with Lemma 7.5-(2) and an easy calculation that $\mathbb{E}(X_1^\top X_2)^4\lesssim\|\Sigma_X\|_F^4$ and $\mathbb{E}(Y_1^\top Y_2)^4\lesssim\|\Sigma_Y\|_F^4$, shows that

$$\begin{aligned}
&\mathbb{E}\bar{R}_2^2\big((X_1,Y_1),(X_2,Y_2)\big)\lesssim_M\tau_Y^4\cdot\|\Sigma_X\|_F^2\cdot\tau_Y^{-8}\|\Sigma_Y\|_F^4+\tau_X^4\cdot\|\Sigma_Y\|_F^2\cdot\tau_X^{-8}\|\Sigma_X\|_F^4\\
&\qquad+\tau_X^4\tau_Y^4\cdot\tau_Y^{-8}\|\Sigma_Y\|_F^4\cdot\tau_X^{-8}\|\Sigma_X\|_F^4+\frac{\|\Sigma_{XY}\|_F^2(1\vee\|\Sigma_{XY}\|_F^2)}{(\tau_X\wedge\tau_Y)^2}\\
&\lesssim_M\frac{\|\Sigma_X\|_F^2\|\Sigma_Y\|_F^2+\|\Sigma_{XY}\|_F^2(1\vee\|\Sigma_{XY}\|_F^2)}{(\tau_X\wedge\tau_Y)^2}.
\end{aligned}$$

As $\operatorname{Var}(g_2)=(1\pm\varepsilon)\operatorname{Var}(\bar{g}_2)+\mathscr{O}\big(\varepsilon^{-1}\cdot\tau_X^{-2}\tau_Y^{-2}\operatorname{Var}(\bar{R}_2(X,Y)\big)$ for any $\varepsilon>0$, the proof is now complete by noting that

$$\frac{\operatorname{Var}\big(\bar{R}_2(X,Y)\big)}{\tau_X^2\tau_Y^2\operatorname{Var}(\bar{g}_2)}\lesssim_M\frac{\frac{\|\Sigma_X\|_F^2\|\Sigma_Y\|_F^2+\|\Sigma_{XY}\|_F^2+\|\Sigma_{XY}\|_F^4}{(\tau_X\wedge\tau_Y)^2}}{\|\Sigma_X\|_F^2\|\Sigma_Y\|_F^2+\|\Sigma_{XY}\|_F^4}\lesssim_M\frac{1}{(\tau_X\wedge\tau_Y)^2},$$

as desired. □

9.3 *Hoeffding decomposition: higher orders*

The goal of this section is to prove the following.

PROPOSITION 9.10. Suppose that the spectrum of Σ is contained in $[1/M, M]$ for some $M > 1$. Then the third- and fourth-order variance are bounded by

$$\mathbb{E}g_3^2 + \mathbb{E}g_4^2 \lesssim \tau_X^{-2}\tau_Y^{-2}\Big(\|\Sigma_X\|_F^2\|\Sigma_Y\|_F^2 + \|\Sigma_{XY}\|_F^2 + \|\Sigma_{XY}\|_F^4\Big) \lesssim \mathbb{E}g_1^2 + \mathbb{E}g_2^2.$$

Here the constants in $\lesssim$ depend on M and the distribution of Z_1 only via its Poincaré constant c_*, excess kurtosis κ and ε_0 prescribed by Assumption A. The claims remain valid with $X = Y$ when the spectrum of $\Sigma_X = \Sigma_Y$ is contained in $[1/M, M]$ for some $M > 1$.

To prove this proposition, we need to evaluate k_3 and k_4. $k_4 = k$ is already given by Proposition 2.1, so we only need to compute k_3 as follows.

LEMMA 9.11. The third-order kernel is given by

$$\begin{aligned} k_3(z_1, z_2, z_3) &= \mathbb{E}k(z_1, z_2, z_3, Z_4) \\ &= \frac{1}{12}\Bigg[2\sum_{1\le i_1<i_2\le 3} U(x_{i_1}, x_{i_2})V(y_{i_1}, y_{i_2}) + 2\sum_{1\le i\le 3}\mathbb{E}U(X, x_i)V(Y, y_i) \\ &\quad - \sum_{(i_1,i_2,i_3)\in\sigma(1,2,3)} U(x_{i_1}, x_{i_2})V(y_{i_1}, y_{i_3}) - \sum_{1\le i_1\ne i_2\le 3}\mathbb{E}U(X, x_{i_1})V(Y, y_{i_2})\Bigg]. \end{aligned}$$

The proof of the above lemma can be found in Appendix D.

Proof. (Proof of Proposition 9.10)For the second moment of g_3, we each term in its definition (9.1) can be bounded as follows:

- First we have

$$\begin{aligned} &\mathbb{E}k_3^2\big((X_1, Y_1), (X_2, Y_2), (X_3, Y_3)\big) \\ &\lesssim \mathbb{E}^{1/2}U^4(X_1, X_2)\cdot\mathbb{E}^{1/2}V^4(Y_1, Y_2) \lesssim \tau_X^{-2}\tau_Y^{-2}\|\Sigma_X\|_F^2\|\Sigma_Y\|_F^2. \end{aligned}$$

  The last inequality follows as

$$\begin{aligned} \mathbb{E}U^4(X_1, X_2) &\lesssim \tau_X^{-4}\Big(\mathbb{E}(X_1^\top X_2)^4 + \tau_X^8\mathbb{E}\bar{R}_X^4(X_1, X_2)\Big) \\ &\overset{(*)}{\lesssim} \tau_X^{-4}\big(\|\Sigma_X\|_F^4 + \tau_X^{-8}\|\Sigma_X\|_F^8\big) \lesssim \tau_X^{-4}\|\Sigma_X\|_F^4, \end{aligned}$$

  and similarly $\mathbb{E}V^4(Y_1, Y_2) \lesssim \tau_Y^{-4}\|\Sigma_Y\|_F^4$, using Lemma 7.5 in $(*)$.

- By Theorem 8.4,

$$\big(\mathrm{dCov}^2(X, Y)\big)^2 \lesssim_M \tau_X^{-2}\tau_Y^{-2}\|\Sigma_{XY}\|_F^4.$$

- By Proposition 9.1 and Lemma G.5,

$$\mathbb{E}g_1^2(X_1, Y_1) \lesssim_M \tau_X^{-2}\tau_Y^{-2}\|\Sigma_{XY}\|_F^2.$$

- By Proposition 9.7,

$$\mathbb{E}g_2^2\big((X_1, Y_1), (X_2, Y_2)\big) \lesssim_M \tau_X^{-2}\tau_Y^{-2}\big(\|\Sigma_X\|_F^2\|\Sigma_Y\|_F^2 + \|\Sigma_{XY}\|_F^4\big).$$

Collecting the above bounds and using the variance lower bound in Lemma 9.5,

$$\mathbb{E}g_3^2 \lesssim_M \tau_X^{-2}\tau_Y^{-2}\Big(\|\Sigma_X\|_F^2\|\Sigma_Y\|_F^2 + \|\Sigma_{XY}\|_F^2 + \|\Sigma_{XY}\|_F^4\Big) \lesssim_M \mathbb{E}g_1^2 + \mathbb{E}g_2^2.$$

The second moment bound for g_4 can be obtained in a similar way so we omit the proof. □

9.4 *Variance expansion*

With the groundwork laid above, we are now able to prove the following variance expansion formula.

THEOREM 9.12. Suppose that the spectrum of Σ is contained in $[1/M, M]$ for some $M > 1$. Then

$$\left|\frac{\mathrm{Var}\left(\mathrm{dCov}_*^2(\boldsymbol{X}, \boldsymbol{Y})\right)}{\bar{\sigma}_n^2(X, Y)} - 1\right| \lesssim n^{-1/2} + (p \wedge q)^{-1/4}.$$

Here $\bar{\sigma}_n^2(X, Y)$ is defined in Theorem 2.2, and the constant in $\lesssim$ depends on M and the distribution of Z_1 only via its Poincaré constant c_*, excess kurtosis κ and ε_0 prescribed by Assumption A. The claims remain valid with $X = Y$ when the spectrum of $\Sigma_X = \Sigma_Y$ is contained in $[1/M, M]$ for some $M > 1$.

Proof. By Hoeffding decomposition $\binom{\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y})=\sum_{c=0}^4}{4cU_n(g_c)}$, so

$$\sigma_\Sigma^2 \equiv \mathrm{Var}_\Sigma\left(\mathrm{dCov}_*^2(\boldsymbol{X}, \boldsymbol{Y})\right) = \sum_{c=1}^4 \binom{4}{c}^2 \binom{n}{c}^{-1} \mathbb{E}g_c^2.$$

Now we may apply Propositions 9.1, 9.7 and 9.10 to conclude that the left-hand side of the desired inequality is bounded by

$$\inf_{\varepsilon>0}\left|(1+\varepsilon)\left(1 + \frac{\mathscr{O}_M(n^{-1} + (p \wedge q)^{-1/2})}{\varepsilon}\right) - 1\right| \asymp n^{-1/2} + (p \wedge q)^{-1/4}.$$

The claim follows. □

10. Normal approximation of truncated dCov_*^2

Let $\bar{g}_0 \equiv \mathbb{E}\,\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y}) = \mathrm{dCov}^2(X,Y)$. Recall $\bar{g}_1, \bar{g}_2$ defined in (9.2) and (9.6). Define the truncated sample distance covariance:

$$\bar{T}_n(\boldsymbol{X},\boldsymbol{Y}) = \sum_{c=0}^{2} \binom{4}{c} U_n(\bar{g}_c). \tag{10.1}$$

The goal of this section is to prove the following non-null CLT for $\bar{T}_n(\boldsymbol{X},\boldsymbol{Y})$.

THEOREM 10.1. Suppose that the spectrum of Σ lies in $[1/M, M]$ for some $M > 1$. Then there exists some $C = C(M, Z_1) > 0$ such that

$$\mathrm{err}_n \equiv d_{\mathrm{Kol}}\left(\frac{\bar{T}_n(\boldsymbol{X},\boldsymbol{Y}) - \mathbb{E}\bar{T}_n(\boldsymbol{X},\boldsymbol{Y})}{\mathrm{Var}^{1/2}(\bar{T}_n(\boldsymbol{X},\boldsymbol{Y}))}, \mathcal{N}(0,1)\right) \le C\left(\frac{1}{n} + \frac{1}{pq}\right)^{1/4}.$$

Here $\mathrm{Var}(\bar{T}_n(\boldsymbol{X},\boldsymbol{Y})) = \bar{\sigma}_n^2(X,Y)$ is defined in Theorem 2.2, and C depends on Z_1 only via its Poincaré constant c_*, excess kurtosis κ and ε_0 prescribed by Assumption A.

The major tool to prove the CLT in Theorem 10.1 is the following discrete second-order Poincaré inequality proved by Chatterjee [9].

LEMMA 10.2. (Discrete second-order Poincaré inequality). Let $X = (X_1, \ldots, X_n)$ be a vector of independent $\mathscr{X}$-valued random variables, and $X' = (X_1', \ldots, X_n')$ be an independent copy of X. For any $A \subset [n]$, define the random variable

$$X_i^A \equiv \begin{cases} X_i', & \text{if } i \in A, \\ X_i, & \text{if } i \notin A. \end{cases}$$

Define $\Delta_j f \equiv f(X) - f(X^{\{j\}})$, $T_A \equiv \sum_{j \notin A} \Delta_j f(X) \Delta_j f(X^A)$, and

$$T \equiv \frac{1}{2} \sum_{A \subsetneq [n]} \frac{T_A}{\binom{n}{|A|}(n - |A|)}.$$

Then with $W \equiv f(X)$ admitting finite variance σ^2,

$$d_{\mathrm{Kol}}\left(\frac{W - \mathbb{E}(W)}{\mathrm{Var}^{1/2}(W)}, \mathcal{N}(0,1)\right) \le 2\left[\frac{\mathrm{Var}^{1/2}\left(\mathbb{E}(T|W)\right)}{\sigma^2} + \frac{1}{2\sigma^3}\sum_{j=1}^{n} \mathbb{E}|\Delta_j f(X)|^3\right]^{1/2}.$$

Proof. This follows from [9, Theorem 2.2] and Lemma G.2. □

We start with the following decomposition of $\bar{T}_n(\boldsymbol{X},\boldsymbol{Y})$. Its proof will be presented in Appendix E.1.

LEMMA 10.3. Let

$$\psi_1(\boldsymbol{X},\boldsymbol{Y}) \equiv \sum_{I_n^2} \big(X_{i_1}^\top X_{i_2} Y_{i_1}^\top Y_{i_2} - \|\Sigma_{XY}\|_F^2\big),$$

$$\psi_2(\boldsymbol{X},\boldsymbol{Y}) \equiv \sum_{i\neq j} \big(X_i^\top \Sigma_{XY} Y_j + X_j^\top \Sigma_{XY} Y_i\big),$$

$$\psi_3(\boldsymbol{X},\boldsymbol{Y}) \equiv \sum_{i=1}^n \Big[\frac{\|\Sigma_{XY}\|_F^2}{\tau_X^2}\big(\|X_i\|^2 - \operatorname{tr}(\Sigma_X)\big) + \frac{\|\Sigma_{XY}\|_F^2}{\tau_Y^2}\big(\|Y_i\|^2 - \operatorname{tr}(\Sigma_Y)\big)\Big].$$

Then

$$\bar{T}_n(\boldsymbol{X},\boldsymbol{Y}) = \mathrm{dCov}^2(X,Y) + \frac{1}{\tau_X\tau_Y\cdot 2\binom{n}{2}}\Big(\psi_1(\boldsymbol{X},\boldsymbol{Y}) - \psi_2(\boldsymbol{X},\boldsymbol{Y})\Big) - \frac{2}{\tau_X\tau_Y n}\psi_3(\boldsymbol{X},\boldsymbol{Y}).$$

Proof. (Outline of the proof of Theorem 10.1) Define $T_{\psi_1}(\boldsymbol{X},\boldsymbol{Y})$-$T_{\psi_3}(\boldsymbol{X},\boldsymbol{Y})$ and $\Delta_i\psi_1(\boldsymbol{X},\boldsymbol{Y})$-$\Delta_i\psi_3(\boldsymbol{X},\boldsymbol{Y})$ as in the discrete second-order Poincaré inequality (cf. Lemma 10.2). The following three propositions give variance and third moment bounds for these quantities.

PROPOSITION 10.4. (Analysis of ψ_1). Assume the conditions in Theorem 10.1. Then the following hold:

1. (Variance bound)

$$\mathrm{Var}\big[\mathbb{E}(T_{\psi_1}|\boldsymbol{X},\boldsymbol{Y})\big] \lesssim n^3\cdot\|\Sigma_X\|_F^4\|\Sigma_Y\|_F^4 + n^4\cdot(1\vee\|\Sigma_{XY}\|_F^2)\|\Sigma_X\|_F^2\|\Sigma_Y\|_F^2 + n^5\cdot(1\vee\|\Sigma_{XY}\|_F^2)\|\Sigma_{XY}\|_F^2.$$

2. (Third moment bound)

$$\sum_{i=1}^n \mathbb{E}|\Delta_i\psi_1(\boldsymbol{X},\boldsymbol{Y})|^3 \lesssim n^{5/2}\operatorname{tr}^{3/2}(\Sigma_X)\operatorname{tr}^{3/2}(\Sigma_Y) + n^4\|\Sigma_{XY}\|_F^3.$$

The constants in $\lesssim$ depend on M and the distribution of Z_1 only.

PROPOSITION 10.5. (Analysis of ψ_2). Assume the conditions in Theorem 10.1. Then the following hold.

1. $\mathrm{Var}\big[\mathbb{E}(T_{\psi_2}|\boldsymbol{X},\boldsymbol{Y})\big] \lesssim n^4\|\Sigma_{XY}\|_F^4$.
2. $\sum_{i=1}^n \mathbb{E}|\Delta_j\psi_2(\boldsymbol{X},\boldsymbol{Y})|^3 \lesssim n^{5/2}\|\Sigma_{XY}\|_F^3$.

The constants in $\lesssim$ depend on M and the distribution of Z_1 only.

PROPOSITION 10.6. (Analysis of ψ_3). Assume the conditions in Theorem 10.1. Then the following hold.

1. $\mathrm{Var}\big[\mathbb{E}(T_{\psi_3}|\boldsymbol{X},\boldsymbol{Y})\big] \lesssim n\cdot\|\Sigma_{XY}\|_F^8(\tau_X^{-4}+\tau_Y^{-4})$.
2. $\sum_{i=1}^n \mathbb{E}|\Delta_i\psi_3(\boldsymbol{X},\boldsymbol{Y})|^3 \lesssim n\cdot\|\Sigma_{XY}\|_F^6(\tau_X^{-3}+\tau_Y^{-3})$.

The constants in $\lesssim$ depend on M and the distribution of Z_1 only.

The proofs of these propositions will be detailed in Appendix E. By the proceeding propositions and Lemma G.5, we have

$$\begin{aligned} D_1 &\equiv \frac{\mathrm{Var}^{1/2}\left(\mathbb{E}(T_{\psi_1}|\boldsymbol{X},\boldsymbol{Y})\right)+\mathrm{Var}^{1/2}\left(\mathbb{E}(T_{\psi_2}|\boldsymbol{X},\boldsymbol{Y})\right)}{n^4\tau_X^2\tau_Y^2}+\frac{\mathrm{Var}^{1/2}\left(\mathbb{E}(T_{\psi_3}|\boldsymbol{X},\boldsymbol{Y})\right)}{n^2\tau_X^2\tau_Y^2} \\ &\lesssim_M \left[\frac{1}{n^{5/2}}+\frac{1\vee\|\Sigma_{XY}\|_F}{n^2\tau_X\tau_Y}+\frac{(1\vee\|\Sigma_{XY}\|_F)\|\Sigma_{XY}\|_F}{n^{3/2}\tau_X^2\tau_Y^2}\right]+\frac{\|\Sigma_{XY}\|_F^4}{n^{3/2}\tau_X^2\tau_Y^2(\tau_X^2\wedge\tau_Y^2)} \\ &\asymp_M \frac{1}{n^2\tau_X^2\tau_Y^2}\Big[n^{-1/2}\tau_X^2\tau_Y^2+\tau_X\tau_Y(1\vee\|\Sigma_{XY}\|_F)+n^{1/2}(1\vee\|\Sigma_{XY}\|_F)\|\Sigma_{XY}\|_F\Big], \end{aligned}$$

and

$$\begin{aligned} D_2 &\equiv \frac{\sum_{i=1}^n\mathbb{E}|\Delta_i\psi_1(\boldsymbol{X},\boldsymbol{Y})|^3+\mathbb{E}|\Delta_i\psi_2(\boldsymbol{X},\boldsymbol{Y})|^3}{n^6\tau_X^3\tau_Y^3}+\frac{\sum_{i=1}^n\mathbb{E}|\Delta_i\psi_3(\boldsymbol{X},\boldsymbol{Y})|^3}{n^3\tau_X^3\tau_Y^3} \\ &\lesssim_M \left[\frac{1}{n^{7/2}}+\frac{\|\Sigma_{XY}\|_F^3}{n^2\tau_X^3\tau_Y^3}\right]+\frac{n\|\Sigma_{XY}\|_F^6(\tau_X^{-3}+\tau_Y^{-3})}{n^3\tau_X^3\tau_Y^3} \\ &\asymp_M \frac{1}{n^3\tau_X^3\tau_Y^3}\Big[n^{-1/2}\tau_X^3\tau_Y^3+n\|\Sigma_{XY}\|_F^3\Big]. \end{aligned}$$

Using Theorem 9.12 with the lower bound Lemma 9.5, we have

$$\bar{\sigma}_n^2=\mathrm{Var}\left(\bar{T}_n(\boldsymbol{X},\boldsymbol{Y})\right)\gtrsim_M \frac{1}{n^2\tau_X^2\tau_Y^2}\Big[n\|\Sigma_{XY}\|_F^2+\tau_X^2\tau_Y^2+\|\Sigma_{XY}\|_F^4\Big]. \tag{10.2}$$

This entails that

$$\begin{aligned} \frac{D_1}{\bar{\sigma}^2} &\lesssim_M \frac{n^{-1/2}\tau_X^2\tau_Y^2+\tau_X\tau_Y(1\vee\|\Sigma_{XY}\|_F)+n^{1/2}(1\vee\|\Sigma_{XY}\|_F)\|\Sigma_{XY}\|_F}{n\|\Sigma_{XY}\|_F^2+\tau_X^2\tau_Y^2+\|\Sigma_{XY}\|_F^4} \\ &\lesssim \frac{1}{n^{1/2}}+\frac{1}{\tau_X\tau_Y}+\frac{\tau_X\tau_Y\|\Sigma_{XY}\|_F}{n\|\Sigma_{XY}\|_F^2+\tau_X^2\tau_Y^2}+\frac{n^{1/2}\|\Sigma_{XY}\|_F}{n\|\Sigma_{XY}\|_F^2+\tau_X^2\tau_Y^2}\asymp\frac{1}{n^{1/2}}+\frac{1}{\tau_X\tau_Y}, \\ \frac{D_2}{\bar{\sigma}^3} &\lesssim_M \frac{n^{-1/2}\tau_X^3\tau_Y^3+n\|\Sigma_{XY}\|_F^3}{n^{3/2}\|\Sigma_{XY}\|_F^3+\tau_X^3\tau_Y^3+\|\Sigma_{XY}\|_F^6}\lesssim\frac{1}{n^{1/2}}. \end{aligned}$$

The claim follows using Lemma 10.2 in the form

$$\mathrm{err}_n\le C\cdot\left[\frac{D_1}{\bar{\sigma}^2}+\frac{D_2}{\bar{\sigma}^3}\right]^{1/2},$$

as desired. □

As outlined above, the major step in the proof is to obtain good enough variance and third moment bounds for $\mathbb{E}[T_{\psi_1}|\boldsymbol{X},\boldsymbol{Y}]$-$\mathbb{E}[T_{\psi_3}|\boldsymbol{X},\boldsymbol{Y}]$ and $\Delta_i\psi_1(\boldsymbol{X},\boldsymbol{Y})$-$\Delta_i\psi_3(\boldsymbol{X},\boldsymbol{Y})$, as claimed in Propositions 10.4–10.6. The proofs to these propositions are fairly delicate and involved. The most complicated case appears to be the control for $\mathbb{E}[T_{\psi_1}|\boldsymbol{X},\boldsymbol{Y}]$, $\Delta_i\psi_1(\boldsymbol{X},\boldsymbol{Y})$ associated with the first term $\psi_1(\boldsymbol{X},\boldsymbol{Y})$ due to its highest polynomial order by definition. The structure of the bounds in Propositions 10.4–10.6 also reveals a careful balance among the power in the terms $n, \|\Sigma_X\|_F\|\Sigma_Y\|_F, \|\Sigma_{XY}\|_F^2$. Such a balance turns out to be crucial to reach the announced error bound in Theorem 10.1 that requires no more than a bounded spectrum condition. See Appendix E for proof details.

11. Proof of Theorem 2.2

(**Step 1**) By definition of $\bar{T}_n(\boldsymbol{X},\boldsymbol{Y})$ in (10.1), we have

$$\Delta_n \equiv \mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y}) - \bar{T}_n(\boldsymbol{X},\boldsymbol{Y}) = \sum_{c=1}^{2} U_n(g_c - \bar{g}_c) + \sum_{c=3}^{4} U_n(g_c).$$

This means

$$\frac{\mathrm{Var}(\Delta_n)}{\bar{\sigma}_n^2(X,Y)} \lesssim \sum_{c=1}^{2} \frac{\mathrm{Var}\left[U_n(g_c - \bar{g}_c)\right]}{\bar{\sigma}_n^2(X,Y)} + \sum_{c=3}^{4} \frac{\mathrm{Var}\left[U_n(g_c)\right]}{\bar{\sigma}_n^2(X,Y)}.$$

For $c = 1$, by Lemma 9.3 and the proof of Proposition 9.1, we have

$$\frac{\mathrm{Var}\left[U_n(g_1 - \bar{g}_1)\right]}{\bar{\sigma}_n^2(X,Y)} \asymp \frac{\mathrm{Var}(\bar{R}_1(X,Y))}{n\tau_X^2\tau_Y^2\bar{\sigma}_n^2(X,Y)} \lesssim \frac{\mathrm{Var}(\bar{R}_1)}{\tau_X^2\tau_Y^2\,\mathrm{Var}(\bar{g}_1)} \lesssim_M \frac{1}{\tau_X \wedge \tau_Y}.$$

For $c = 2$, by Lemma 9.9 and the proof of Proposition 9.7, we have

$$\frac{\mathrm{Var}\left[U_n(g_2 - \bar{g}_2)\right]}{\bar{\sigma}_n^2(X,Y)} \lesssim_M \frac{\mathrm{Var}(\bar{R}_2)}{n^2\tau_X^2\tau_Y^2\bar{\sigma}_n^2(X,Y)} \lesssim_M \frac{\mathrm{Var}(\bar{R}_2)}{\tau_X^2\tau_Y^2\,\mathrm{Var}(\bar{g}_2)} \lesssim_M \frac{1}{(\tau_X \wedge \tau_Y)^2}.$$

For $c = 3, 4$, the proof of Proposition 9.10 yields that

$$\mathrm{Var}\left[U_n(g_3)\right] + \mathrm{Var}\left[U_n(g_4)\right] \lesssim_M \frac{1}{n^3\tau_X^2\tau_Y^2}\left(\|\Sigma_X\|_F^2\|\Sigma_Y\|_F^2 + \|\Sigma_{XY}\|_F^2 + \|\Sigma_{XY}\|_F^4\right),$$

so using the variance lower bound in (10.2), we have

$$\frac{\mathrm{Var}\left[U_n(g_3)\right] + \mathrm{Var}\left[U_n(g_4)\right]}{\bar{\sigma}_n^2(X,Y)} \lesssim_M \frac{1}{n}.$$

Collecting the bounds, we have

$$\frac{\mathrm{Var}(\Delta_n)}{\bar{\sigma}_n^2(X,Y)} \lesssim_M \frac{1}{n} + \frac{1}{\tau_X \wedge \tau_Y}. \tag{11.1}$$

(**Step 2**) First consider normalization by $\bar{\sigma}_n(X,Y) = \mathrm{Var}^{1/2}(\bar{T}_n(\boldsymbol{X},\boldsymbol{Y}))$. Using the decomposition

$$\bar{L}_n \equiv \frac{\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y}) - \mathrm{dCov}^2(X,Y)}{\bar{\sigma}_n(X,Y)} = \frac{\Delta_n}{\bar{\sigma}_n(X,Y)} + \frac{\bar{T}_n(\boldsymbol{X},\boldsymbol{Y}) - \mathbb{E}\bar{T}_n(\boldsymbol{X},\boldsymbol{Y})}{\mathrm{Var}^{1/2}(\bar{T}_n(\boldsymbol{X},\boldsymbol{Y}))},$$

by Lemma G.1 and Theorem 10.1 (note that $\mathbb{E}\Delta_n = 0$),

$$\begin{aligned} &d_{\mathrm{Kol}}\left(\frac{\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y}) - \mathrm{dCov}^2(X,Y)}{\bar{\sigma}_n(X,Y)}, \mathcal{N}(0,1)\right) \\ &\le d_{\mathrm{Kol}}\left(\frac{\bar{T}_n(\boldsymbol{X},\boldsymbol{Y}) - \mathbb{E}\bar{T}_n(\boldsymbol{X},\boldsymbol{Y})}{\mathrm{Var}^{1/2}(\bar{T}_n(\boldsymbol{X},\boldsymbol{Y}))}, \mathcal{N}(0,1)\right) + 2\left(\frac{\mathrm{Var}(\Delta_n)}{\bar{\sigma}_n^2(X,Y)}\right)^{1/3} \lesssim_M \left(\frac{1}{n \wedge p \wedge q}\right)^{1/6}. \end{aligned}$$

Next, with the normalization $\mathrm{Var}^{1/2}(\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y}))$, consider the decomposition

$$\frac{\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y}) - \mathrm{dCov}^2(X,Y)}{\mathrm{Var}^{1/2}(\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y}))} = \bar{L}_n + \bar{L}_n\left(\frac{\bar{\sigma}_n(X,Y)}{\mathrm{Var}^{1/2}(\mathrm{dCov}_*^2(\boldsymbol{X},\boldsymbol{Y}))} - 1\right) \equiv \bar{L}_n + \bar{\Delta}_n.$$

By Theorem 9.12, $\mathrm{Var}(\bar{\Delta}_n) \lesssim_M n^{-1} + (p \wedge q)^{-1/2}$. The claim now follows by invoking Lemma G.1. □

Funding

The research of Q. Han is partially supported by NSF grant DMS-2143468.

Data Availability Statement

No new data were generated or analysed in support of this review.

References

1. Anderson, T. W. (1958) *An introduction to multivariate statistical analysis* Wiley Publications in Statistics, John Wiley & Sons, Inc., New York. London: Chapman & Hall, Ltd.
2. Bai, Z. & Silverstein, J. W. (2010) *Spectral analysis of large dimensional random matrices*, second edn. Springer, New York: Springer Series in Statistics.
3. Bakry, D., Gentil, I. & Ledoux, M. (2014) *Analysis and geometry of Markov diffusion operators, Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*, vol. **348**. Cham: Springer.
4. Bergsma, W. & Dassios, A. (2014) A consistent test of independence based on a sign covariance related to Kendall's tau. *Bernoulli*, **20**, 1006–1028.
5. Blum, J. R., Kiefer, J. & Rosenblatt, M. (1961) Distribution free tests of independence based on the sample distribution function. *Ann. Math. Stat.*, **32**, 485–498.

6. Bobkov, S. G. (1999) Isoperimetric and analytic inequalities for log-concave probability measures. *Ann. Probab.*, **27**, 1903–1921.
7. Bobkov, S. G. & Ledoux, M. (2009) Weighted Poincaré-type inequalities for Cauchy and other convex measures. *Ann. Probab.*, **37**, 403–427.
8. Cai, T. T. & Zongming, M. A. (2013) Optimal hypothesis testing for high dimensional covariance matrices. *Bernoulli*, **19**, 2359–2388.
9. Chatterjee, S. (2008) A new method of normal approximation. *Ann. Probab.*, **36**, 1584–1610.
10. Dette, H. & Dörnemann, N. (2020) Likelihood ratio tests for many groups in high dimensions. *J. Multivar. Anal.*, **178**, 104605.
11. Ding, X. & Yang, F. (2018) A necessary and sufficient condition for edge universality at the largest singular values of covariance matrices. *Ann. Appl. Probab.*, **28**, 1679–1738.
12. Ding, X. & Yang, F. (2021) Spiked separable covariance matrices and principal components. *Ann. Stat.*, **49**, 1113–1138.
13. Dörnemann, N. (2023 Paper No. 105122) Likelihood ratio tests under model misspecification in high dimensions. *J. Multivar. Anal.*, **193**, 20.
14. Erdős, L. & Yau, H.-T. (2017) *A dynamical approach to random matrix theory, Courant Lecture Notes in Mathematics*, vol. **28** Courant Institute of Mathematical Sciences, New York. Providence, RI: American Mathematical Society.
15. Feuerverger, A. (1993) A consistent test for bivariate dependence. *Int. Stat. Rev.*, **61**, 419–433.
16. Gao, H. & Shao, X. (2023) Two sample testing in high dimension via maximum mean discrepancy. *Journal of Machine Learning Research (JMLR)*, **24**, 304.
17. Gao, L., Fan, Y., Lv, J. & Shao, Q.-M. (2021) Asymptotic distributions of high-dimensional distance correlation inference. *Ann. Stat.* to appear. Available at arXiv:1910.12970, **49**, 1999–2020.
18. Gretton, A., Herbrich, R., Smola, A., Bousquet, O. & Schölkopf, B. (2005) Kernel methods for measuring independence. *J. Mach. Learn. Res.*, **6**, 2075–2129.
19. Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B. & Smola, A. J. (2007) *A kernel statistical test of independence*. In: Platt, J., Koller, D., Singer, Y., & Roweis, S. (eds) *Advances in Neural Information Processing Systems*, vol. **20**. Citeseer: Curran Associates, Inc., pp. 585–592.
20. Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. & Smola, A. (2012) A kernel two-sample test. *J. Mach. Learn. Res.*, **13**, 723–773.
21. Hoeffding, W. (1948) A non-parametric test of independence. *Ann. Math. Stat.*, **19**, 546–557.
22. Huo, X. & Székely, G. J. (2016) Fast computing for distance covariance. *Technometrics*, **58**, 435–447.
23. Jiang, T. & Qi, Y. (2015) Likelihood ratio tests for high-dimensional normal distributions. *Scand. J. Stat.*, **42**, 988–1009.
24. Jiang, T. & Yang, F. (2013) Central limit theorems for classical likelihood ratio tests for high-dimensional normal distributions. *Ann. Stat.*, **41**, 2029–2074.
25. Kendall, M. G. (1938) A new measure of rank correlation. *Biometrika*, **30**, 81–93.
26. Knowles, A. & Yin, J. (2017) Anisotropic local laws for random matrices. *Probab. Theory Relat. Fields*, **169**, 257–352.
27. Kong, J., Klein, B. E. K., Klein, R., Lee, K. E. & Wahba, G. (2012) Using distance correlation and ss-anova to assess associations of familial relationships, lifestyle factors, diseases, and mortality. *Proc. Natl. Acad. Sci.*, **109**, 20352–20357.
28. Li, R., Zhong, W. & Zhu, L. (2012) Feature screening via distance correlation learning. *J. Am. Stat. Assoc.*, **107**, 1129–1139.
29. Lyons, R. (2013) Distance covariance in metric spaces. *Ann. Probab.*, **41**, 3284–3305.
30. Matteson, D. S. & Tsay, R. S. (2017) Independent component analysis via distance covariance. *J. Am. Stat. Assoc.*, **112**, 623–637.
31. Muirhead, R. J. (1982) *Aspects of multivariate statistical theory*. New York: John Wiley & Sons, Inc. Wiley Series in Probability and Mathematical Statistics.

32. Pearson, K. (1895) Notes on regression and inheritance in the case of two parents. *Proc. R. Soc. London*, **58**, 240–242.
33. Qi, Y., Wang, F. & Zhang, L. (2019) Limiting distributions of likelihood ratio test for independence of components for high-dimensional normal vectors. *Ann. Inst.Stat. Math.*, **71**, 911–946.
34. Ramdas Aaditya, Isenberg David, Singh Aarti, and Wasserman Larry, *Minimax lower bounds for linear independence testing*, *2016 IEEE international symposium on information theory (ISIT)*, *IEEE*, **2016**, pp. 965–969.
35. Rosenblatt, M. (1975) A quadratic measure of deviation of two-dimensional density estimates and a test of independence. *Ann. Stat.*, **3**, 1–14.
36. Saumard, A. & Wellner, J. A. (2014) Log-concavity and strong log-concavity: a review. *Stat. Surv.*, **8**, 45–114.
37. Sejdinovic, D., Sriperumbudur, B., Gretton, A. & Fukumizu, K. (2013) Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Stat.*, **41**, 2263–2291.
38. Serfling, R. J. (1980) *Approximation theorems of mathematical statistics*. New York: John Wiley & Sons, Inc. Wiley Series in Probability and Mathematical Statistics.
39. Shao, X. & Zhang, J. (2014) Martingale difference correlation and its use in high-dimensional variable screening. *J. Am. Stat. Assoc.*, **109**, 1302–1318.
40. Spearman, C. (1904) The proof and measurement of association between two things. *Am. J. Psychol.*, **15**, 72–101.
41. Székely, G. J. & Rizzo, M. L. (2009) Brownian distance covariance. *Ann. Appl. Stat.*, **3**, 1236–1265.
42. Székely, G. J. & Rizzo, M. L. (2013) The distance correlationt-test of independence in high dimension. *J. Multivar. Anal.*, **117**, 193–213.
43. Székely, G. J. & Rizzo, M. L. (2014) Partial distance correlation with methods for dissimilarities. *Ann. Stat.*, **42**, 2382–2412.
44. Székely, G. J., Rizzo, M. L. & Bakirov, N. K. (2007) Measuring and testing dependence by correlation of distances. *Ann. Stat.*, **35**, 2769–2794.
45. Tao, T. & Van, V. (2012) Random covariance matrices: universality of local statistics of eigenvalues. *Ann. Probab.*, **40**, 1285–1315.
46. Weihs, L., Drton, M. & Meinshausen, N. (2018) Symmetric rank covariances: a generalized framework for nonparametric measures of dependence. *Biometrika*, **105**, 547–562.
47. de Wet, T. (1980) Cramér-von Mises tests for independence. *J. Multivar. Anal.*, **10**, 38–50.
48. Yan, J. & Zhang, X. (2023) Kernel two-sample tests in high dimensions: interplay between moment discrepancy and dimension-and-sample orders. *Biometrika*, **110**, 411–430.
49. Yanagimoto, T. (1970) On measures of association and a related problem. *Ann. Inst. Stat. Math.*, **22**, 57–63.
50. Yao, S., Zhang, X. & Shao, X. (2018) Testing mutual independence in high dimension via distance covariance. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **80**, 455–480.
51. Zhang, X., Yao, S. & Shao, X. (2018) Conditional mean and quantile dependence testing in high dimension. *Ann. Stat.*, **46**, 219–246.
52. Zhu, C., Zhang, X., Yao, S. & Shao, X. (2020) Distance-based and RKHS-based dependence metrics in high dimension. *Ann. Stat.*, **48**, 3366–3394.