Robust Multivariate Nonparametric Tests via Projection-Pursuit

Ilmun Kim†  Sivaraman Balakrishnan†  Larry Wasserman†

Department of Statistics and Data Science†
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

In this work, we generalize the Cramér-von Mises statistic via projection pursuit to obtain a robust test for the multivariate two-sample problem. The proposed test is consistent against all fixed alternatives, robust to heavy-tailed data and minimax rate optimal. Our test statistic is completely free of tuning parameters and are computationally efficient even in high dimensions. When the dimension tends to infinity, the proposed test is shown to have identical power to that of the existing high-dimensional mean tests under certain location models. As a by-product of our approach, we introduce a new metric called the angular distance which can be thought of as a robust alternative to the Euclidean distance. Using the angular distance, we connect the proposed to the reproducing kernel Hilbert space approach. In addition to the Cramér-von Mises statistic, we show that the projection pursuit technique can be used to define robust, multivariate tests in many other problems.

1 Introduction

Let $X$ and $Y$ be random vectors in $\mathbb{R}^d$ with distribution functions $F_X$ and $G_Y$, respectively. Given two independent samples $X_m = \{X_1, \ldots, X_m\}$ and $Y_n = \{Y_1, \ldots, Y_n\}$ from $F_X$ and $G_Y$, we want to test

$$H_0 : F_X = G_Y \quad \text{versus} \quad H_1 : F_X \neq G_Y.$$  

This fundamental testing problem has received considerable attention in statistics with a wide range of applications (see, e.g., Thas, 2010, for a review). A common statistic for the univariate two-sample testing is the Cramér-von Mises (CvM) statistic (Anderson, 1962):

$$mn \int_{-\infty}^{\infty} \left( \hat{F}_X(t) - \hat{G}_Y(t) \right)^2 d\hat{H}(t),$$

where $\hat{F}_X(t)$ and $\hat{G}_Y(t)$ are the empirical distribution functions of $X_m$ and $Y_n$, respectively, and $(m + n)\hat{H}(t) = m\hat{F}_X(t) + n\hat{G}_Y(t)$. Another approach is based on the energy statistic, which is an estimate of the squared energy distance (Székely and Rizzo, 2013):

$$E^2 = 2 \mathbb{E}_{X,Y}[|X - Y|] - \mathbb{E}_{X,Y'}[|X - X'|] - \mathbb{E}_{Y,Y'}[|Y - Y'|].$$

The energy distance is well-defined assuming a finite first moment and it can be written in a form that is similar to the Cramér’s distance, namely,

$$E^2 = 2 \int_{-\infty}^{\infty} (F_X(t) - G_Y(t))^2 dt.$$
The CvM statistic has several advantages over the energy statistic for univariate two-sample testing. For instance, the CvM statistic is distribution-free under $H_0$ (Anderson, 1962) and its population quantity is well-defined without any moment assumptions. It also has an intuitive probabilistic interpretation as described in Baringhaus and Henze (2017). Nevertheless, the CvM statistic has rarely been studied for multivariate testing. A primary reason is that the CvM statistic is essentially rank-based, which leads to a challenge to generalize it in a multivariate space. In contrast, the energy statistic can be easily applied in arbitrary dimensions as in Baringhaus and Franz (2004); Székely and Rizzo (2004). Specifically, they defined the squared multivariate energy distance by

$$E_2^2 = 2\mathbb{E}_{X,Y}[||X - Y||] - \mathbb{E}_{X,X'}[||X - X'||] - \mathbb{E}_{Y,Y'}[||Y - Y'||],$$

where $||\cdot||$ is the Euclidean norm in $\mathbb{R}^d$. The multivariate energy distance maintains the characteristic property that it is always non-negative and equal to zero if and only if $F_X = G_Y$. It can also be viewed as the average of univariate Cramér’s distances of projected random variables (Baringhaus and Franz, 2004):

$$E_2^2 = \frac{\sqrt{\pi}(d-1)\Gamma\left(\frac{d-1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} \int_{S^{d-1}} \int_{\mathbb{R}} (F_{\beta^\top X}(t) - G_{\beta^\top Y}(t))^2 dt d\lambda(\beta),$$

where $\lambda$ represents the uniform probability measure on the $d$-dimensional unit sphere $S^{d-1} = \{x \in \mathbb{R}^d : ||x|| = 1\}$ and for each direction $\beta \in S^{d-1}$, $F_{\beta^\top X}$ and $G_{\beta^\top Y}$ denote the distribution functions of $\beta^\top X$ and $\beta^\top Y$, respectively.

Although the multivariate energy distance can be simply estimated in any dimension, it still requires the finite moment assumption as in the univariate case; thereby, the resulting test might suffer from low power when the underlying distributions are heavy-tailed or contain outliers. Given that outlying observations arise frequently in practice with high-dimensional data, there is a need to develop a robust counterpart of the energy distance. The primary goal of this work is to introduce a robust, tuning parameter free, two-sample testing procedure that is easily applicable in arbitrary dimensions and consistent against all alternatives. Specifically, we modify the univariate CvM statistic to generalize it to an arbitrary dimension by averaging over all one-dimensional projections. In detail, the proposed test statistic is an unbiased estimate of the squared multivariate CvM distance defined as follows:

$$W_2^2 = \int_{S^{d-1}} \int_{\mathbb{R}} (F_{\beta^\top X}(t) - G_{\beta^\top Y}(t))^2 \frac{d(F_{\beta^\top X}(t)/2 + G_{\beta^\top Y}(t)/2)}{d\lambda(\beta)}.$$

Throughout this paper, we refer to the process of averaging over all projections as projection pursuit.

### 1.1 Summary of our results

The proposed multivariate CvM distance shares some attractive properties of the energy distance while being robust to heavy-tailed distributions or outliers. For example, $W_d$ satisfies the characteristic property (Lemma 2.1) and is invariant to orthogonal transformations. More importantly, it is straightforward to estimate $W_d$ without using any tuning parameters (Theorem 2.1). For the first part of this paper, we investigate properties of the multivariate CvM distance as well as the resulting testing procedure. Our main results regarding the first topic are summarized as follows:
• **Closed form expression** (Section 2) We show that the test statistic has a simple closed-form expression.

• **Asymptotic power** (Section 2) We prove that the permutation test based on the proposed statistic has the same asymptotic power as the oracle test.

• **Robustness** (Section 3) We show that the permutation test based on the proposed statistic maintains good power in the Huber contamination model, while the energy test becomes completely powerless in this setting.

• **Minimax optimality** (Section 4) We analyze the finite-sample power of the proposed permutation test and prove its minimax rate optimality.

• **Adaptivity in high-dimensions** (Section 5) We consider a high-dimensional regime where the dimension tends to infinity. Under this regime, we show that the limiting power of the proposed test, the energy test and the high-dimensional mean tests introduced by Chen and Qin (2010) and Chakraborty and Chaudhuri (2017) are the same against certain location models.

• **Angular distance** (Section 6) We introduce the angular distance between two vectors and use this to show that the multivariate CvM distance is a member of the generalized energy distance (Sejdinovic et al., 2013) as well as the maximum mean discrepancy (MMD) (Gretton et al., 2012) associated with the angular distance.

Beyond the CvM statistic, the projection pursuit technique can be widely applicable to other nonparametric statistics. For the second part of this study, we revisit some famous univariate sign- or rank-based statistics and propose their multivariate counterparts via projection pursuit. Although there has been much effort to extend univariate sign- or rank-based statistics in a multivariate space (see, e.g., Liu, 2006; Oja, 2010, for a survey), they are either computationally expensive to implement or less intuitive to understand. Our projection pursuit approach addresses these issues by providing a tractable calculation form of statistics and by having a direct interpretation in terms of projections. In Section 7, we demonstrate the generality of the projection pursuit approach by presenting multivariate extensions of several existing univariate statistics.

### 1.2 Literature review

There are a number of multivariate two-sample testing procedures available in the literature. To list a few: Anderson et al. (1994) proposed the two-sample statistic based on the integrated square distance between two kernel density estimates. The energy statistic is introduced by Baringhaus and Franz (2004) and Székely and Rizzo (2004) independently. Gretton et al. (2012) introduced a class of distances between two probability distributions, called the maximum mean discrepancy (MMD). They applied the MMD to multivariate two-sample testing with the Gaussian kernel. Sejdinovic et al. (2013) showed that the energy distance is a special case of the MMD associated with the kernel induced by the Euclidean distance. Another line of work is based on graph constructions including the k-nearest neighbor (NN) (Schilling, 1986; Henze, 1988; Mondal et al., 2015) and the minimum spanning tree (Friedman and Rafsky, 1979; Chen and Friedman, 2017). However, there are certain drawbacks associated with the use of these methods. For example, one needs to choose tuning parameters such as the bandwidths for the Gaussian MMD and the number of the nearest neighbors in the k-NN. In addition, certain types of graph-based tests require an algorithm with high computational complexity. On the other hand, the proposed approach is completely free of tuning parameters and can be computed in $O(m + n)$ if the linear-type statistic is used (see Section 4) and
Table 1: The comparisons between different two-sample testing methods: 1) Cramér-von Mises distance (CvM), 2) Energy distance (Energy), 3) Gaussian MMD (MMD) and 4) k-nearest neighbor (NN).

| Characteristic          | CvM | Energy | MMD | NN |
|-------------------------|-----|--------|-----|----|
| 1. Robustness           | ✓   | ×      | ×   | ×  |
| 2. No tuning parameters | ✓   | ✓      | ×   | ×  |
| 3. Interpretability     | ✓   | ×      | ×   | ✓  |
| 4. Computational cost   | ✓   | ✓      | ✓   | ✓  |
| 5. Minimax optimality   | ✓   | ?      | ?   | ?  |

in $O((m + n)^3)$ if the U-statistic is used.

The projection pursuit approach to CvM-type statistics can be found in other statistical problems. For example, Zhu et al. (1997) and Cui (2002) considered the CvM statistic using projection pursuit to investigate one-sample goodness-of-fit tests for multivariate distributions. Escanciano (2006) proposed the CvM-based goodness-of-fit test for parametric regression models. To the best of our knowledge, however, this is the first study that investigates the CvM statistic for the multivariate two-sample problem using projection pursuit.

The advantages of our approach are summarized in Table 1 where we compare the proposed test to some popular existing tests.

Outline. The rest of this paper is organized as follows. In Section 2, we introduce our test statistic and study its limiting behavior. In Section 3, we compare the power of the CvM test with that of the energy test and we highlight the robustness of the CvM test. Section 4 establishes the minimax rate optimality of the proposed test. In Section 5, we study the limiting power of the CvM test in high-dimensions and prove that it becomes equivalent to that of the existing high-dimensional mean tests against certain location models. We introduce the angular distance between two vectors in Section 6 to show that the CvM distance is a special case of the generalized energy distance associated with the introduced distance. In Section 7, the projection pursuit technique is applied to other sign- or rank-based statistics and this allows us to provide new multivariate extensions. Simulation results are presented in Section 8 to demonstrate the competitive power performance of the proposed approach with finite sample size. All proofs not contained in the main text are in the supplementary material.

Notation. The following notation is used throughout this paper. For $U_1, U_2 \in \mathbb{R}^d$, we denote the angle between $U_1$ and $U_2$ by $\Psi(U_1, U_2) = \arccos(U_1 / \|U_1\|, U_2 / \|U_2\|)$ where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^d$. Let $S^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$ and let $\lambda(\cdot)$ be the uniform uniform probability measure on $S^{d-1}$. For given $\beta \in S^{d-1}$, $F_{\beta^\top X}(t)$, $G_{\beta^\top Y}(t)$ denote the distribution functions of $\beta^\top X$ and $\beta^\top Y$, respectively. For $1 \leq q < p$, we let $(p)_q = p(p - 1) \cdots (p - q + 1)$. Let $\mathbb{P}_0$ and $\mathbb{P}_1$ be the probability measures under $H_0$ and $H_1$, respectively. Similarly, $\mathbb{E}_0$ and $\mathbb{E}_1$ stand for the expectations under $H_0$ and $H_1$. We denote the distribution function of $N(0, 1)$ by $\Phi$ and let $z_\alpha$ be the upper $\alpha$ quantile of $N(0, 1)$. We use $a_n \asymp b_n$ if there exist constants $C, C' > 0$ such that $C < |a_n / b_n| < C'$. Similarly, $a_n \gtrsim b_n$ if there exists $C > 0$ such that $Cb_n \geq a_n$. 

4
2 Projection Pursuit-Type Cramér-von Mises Statistics

In this section, we present the basic properties of the CvM distance and introduce our test statistic. To start, we establish the characteristic property of the CvM distance, meaning that \( W_d \) is nonnegative and equal to zero if and only if \( F_X = G_Y \). As a consequence, the resulting tests are consistent against all fixed alternatives.

Lemma 2.1. \( W_d \) is nonnegative and has the characteristic property:

\[
W_d(F_X, G_Y) = 0 \quad \text{if and only if} \quad F_X = G_Y.
\]

Note that \( W_d \) involves integration over the unit sphere. One way to approximate this integral is to consider a subset of \( S^{d-1} \), namely \( \{\beta_1, \ldots, \beta_k\} \), and then taking the sample mean over \( k \) different univariate CvM statistics (see e.g. Zhu et al., 1997). However, this approach has a clear trade-off between accuracy and computational time depending on the choice of \( k \). Our approach does not suffer from this issue by explicitly calculating the integral over \( S^{d-1} \). The explicit form of the integration is mainly due to Escanciano (2006) who provided the following lemma:

Lemma 2.2. (Escanciano, 2006) For any two vectors \( U_1, U_2 \in \mathbb{R}^d \),

\[
\int_{S^{d-1}} I(\beta^T U_1 \leq 0) I(\beta^T U_2 \leq 0) d\lambda(\beta) = \frac{1}{2} - \frac{1}{2\pi} \Psi(U_1, U_2),
\]

where

\[
\Psi(U_1, U_2) := \arccos \left( \frac{U_1^T U_2}{||U_1|| ||U_2||} \right).
\]

Based on Lemma 2.2, we present another representation for \( W_d^2 \) in terms of the expected angle involving three independent random vectors.

Theorem 2.1. Let \( X, X' \overset{i.i.d.}{\sim} F_X \) and \( Y, Y' \overset{i.i.d.}{\sim} G_Y \) where \( \beta^T X \) and \( \beta^T Y \) have continuous distribution for all \( \beta \in S^{d-1} \) almost surely. Then the squared multivariate CvM distance can be written as

\[
W_d^2 = \frac{1}{3} - \frac{1}{2\pi} \mathbb{E} \left[ \Psi(X - Y, X' - Y) \right] - \frac{1}{2\pi} \mathbb{E} \left[ \Psi(Y - X, Y' - X) \right].
\]

2.1 Test Statistic

Theorem 2.1 leads to a straightforward empirical estimate of \( W_d^2 \) based on a \( U \)-statistic. Consider the kernel of order two:

\[
h(x_1, x_2; y_1, y_2) = \frac{1}{3} - \frac{1}{2\pi} \Psi(x_1 - y_1, x_2 - y_1) - \frac{1}{2\pi} \Psi(y_1 - x_1, y_2 - x_1).
\]

Then we define our test statistic as follows:

\[
U_{\text{CvM}} = \frac{1}{(m)_2(n)_2} \sum_{i_1 \neq i_2} \sum_{j_1 \neq j_2} h(X_{i_1}, X_{i_2}; Y_{j_1}, Y_{j_2}).
\]

Based on the theory of \( U \)-statistics, it is clear that \( U_{\text{CvM}} \) is an unbiased estimator of \( W_d^2 \). Next we explore the asymptotic distribution of the proposed test statistic under the null hypothesis.
Theorem 2.2 (Asymptotic null distribution of $U_{CvM}$). Let $m/N \to \pi_1 \in (0, 1)$ and $n/N \to \pi_2 \in (0, 1)$ with $N = m + n$ as $N \to \infty$. Let $\lambda_k$ be the eigenvalue with the corresponding eigenfunction $\phi_k$ satisfying the integral equation

$$
E_{X_2} \left\{ E_{Y_1,Y_2} \left[ \tilde{h}(x_1, X_2; Y_1, Y_2 | X_2) \phi_k(X_2) \right] \right\} = \lambda_k \phi_k(x_1) \quad \text{for } k = 1, 2, \ldots,
$$

where $\tilde{h}(x_1, x_2; y_1, y_2) = 1/2h(x_1, x_2; y_1, y_2) + 1/2h(x_2, x_1; y_2, y_1)$. Then $U_{CvM}$ has the limiting null distribution as

$$
\frac{mn}{m+n} U_{CvM} \xrightarrow{d} \sum_{k=1}^{\infty} \lambda_k (Z_k^2 - 1),
$$

where $Z_k \overset{i.i.d.}{\sim} N(0, 1)$.

Remark 2.1. When $F_X \neq G_Y$, $U_{CvM}$ is a non-degenerate $U$-statistic. Followed by the asymptotic theory of non-degenerate two-sample $U$-statistics (see e.g. Chapter 3.2.3 of Kowalski and Tu, 2008), $\sqrt{N} U_{CvM}$ converges to a normal distribution under the alternative hypothesis.

2.2 Critical Value and Permutation Test

Unfortunately, when $d \geq 2$, the null distribution of $U_{CvM}$ depends on the underlying distributions $F_X$ and $G_Y$; therefore, the resulting test is not distribution-free. Suppose that the mixture distribution $\pi_1 F_X + \pi_2 G_Y$ is known where $m/N \to \pi_1$ and $n/N \to \pi_2$ as $N \to \infty$. Then the critical value of the oracle test can be decided as follows:

• Oracle Test

1. Consider i.i.d. samples $\{Z_1, \ldots, Z_N\}$ from the mixture $\pi_1 F_X + \pi_2 G_Y$.

2. Let $T_{m,n}(Z)$ be the test statistic of interest calculated based on $X_m = \{Z_1, \ldots, Z_m\}$ and $Y_n = \{Z_{m+1}, \ldots, Z_N\}$.

3. Return the critical value $c_{\alpha,m,n}$ defined by

$$
c_{\alpha,m,n} := \inf \left\{ t \in \mathbb{R} : 1 - \alpha \leq P \left( T_{m,n}(Z) \leq t \right) \right\}.
$$

(7)

In practice, $\pi_1 F_X + \pi_2 G_Y$ is unknown and thus $c_{\alpha,m,n}$ is not available. Instead, we consider the permutation procedure described below.

• Permutation Test

1. Let $(Z_1, \ldots, Z_N) = (X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ be the pooled samples and $Z(\pi) = (Z_{\pi(1)}, \ldots, Z_{\pi(N)})$ where $\pi = (\pi(1), \ldots, \pi(N))$ is a permutation of $\{1, \ldots, N\}$.

2. Let $T_{m,n}(Z_\pi)$ be the test statistic of interest calculated based on $X_m = \{Z_{\pi(1)}, \ldots, Z_{\pi(m)}\}$ and $Y_n = \{Z_{\pi(m+1)}, \ldots, Z_{\pi(N)}\}$. 
3. Return the critical value $c^*_{\alpha,m,n}$ defined by

$$c^*_{\alpha,m,n} := \inf \left\{ t \in \mathbb{R} : 1 - \alpha \leq \frac{1}{N!} \sum_{\pi \in G_N} I \left( T_{m,n}(Z_\pi) \leq t \right) \right\},$$

where $G_N$ is the set of all permutations of $\{1,\ldots,N\}$.

In Theorem 2.3, we show that the difference between $c_{\alpha,m,n}$ and $c^*_{\alpha,m,n}$ of the proposed statistic is asymptotically negligible. This implies that the permutation test has the same asymptotic power as the oracle test while being exact for any finite sample size. We formally state this in Corollary 2.1.

**Theorem 2.3.** Suppose that $m/N \to \pi_1 \in (0, 1)$ and $n/N \to \pi_2 \in (0, 1)$ as $N \to \infty$. Let $c_{\alpha,m,n}$ and $c^*_{\alpha,m,n}$ be the critical values of the oracle test and the permutation test based on $\frac{mn}{m+n} U_{\text{CVM}}$ respectively defined in (7) and (8). Then under the null and the alternative hypothesis

$$c_{\alpha,m,n} - c^*_{\alpha,m,n} \xrightarrow{p} 0.$$  

**Corollary 2.1.** Let us denote the oracle and permutation test functions by

$$\phi_{\text{oracle}} := I \left( \frac{mn}{m+n} U_{\text{CVM}} \geq c_{\alpha,m,n} \right) \quad \text{and} \quad \phi_{\text{CVM}} := I \left( \frac{mn}{m+n} U_{\text{CVM}} \geq c^*_{\alpha,m,n} \right).$$

Then under the null hypothesis, $\mathbb{E}_0[\phi_{\text{oracle}}] \leq \alpha$ and $\mathbb{E}_0[\phi_{\text{CVM}}] \leq \alpha$ for any sample size. Whereas under the alternative hypothesis,

$$\mathbb{E}_1[\phi_{\text{oracle}}] - \mathbb{E}_1[\phi_{\text{CVM}}] \to 0 \text{ as } N \to \infty.$$  

**Proof.** The control of type I error is obvious from the definition of $c_{\alpha,m,n}$ and the well-known property of permutation tests (e.g. Chapter 15 of Lehmann and Romano, 2006). The power statement is a consequence of Theorem 2.3. \qed

### 3 Robustness

Recall that the energy distance and the CvM distance can be represented by integrals of the $L_2$-type difference between two distribution functions. In view of this, the main difference between the energy distance and the CvM distance is in their weight function. More precisely, the energy distance is defined with $dt$, which gives the uniform weight on the whole real line. On the other hand, the CvM distance is defined with $d(F_{\beta^\top X}(t)/2 + G_{\beta^\top Y}(t)/2)$, which gives the most weight on high-density regions. As a result, the test based on the CvM distance is more robust to extreme observations than the one based on the energy distance. It is also important to note that the CvM distance does not require any moment conditions. On the other hand, the energy distance is only well-defined assuming a finite first moment. When the moment condition is violated or there exist extreme observations, the test based on the energy distance may suffer from low power. In this section, we demonstrate this point by using contaminated distribution models.

Suppose we observe samples from an $\epsilon$-contamination model:

$$X \sim P_m := (1 - \epsilon)F_X + \epsilon H_m \quad \text{and} \quad Y \sim Q_n := (1 - \epsilon)G_Y + \epsilon H'_n,$$  

(9)
where $F_X$ and $G_Y$ are fixed but $H_m$ and $H'_n$ can change arbitrarily with $m, n$. With this model assumption, we would like to assess whether a test can maintain adequate power even when $H_m, H'_n$ have an adverse impact on power performance. We mainly focus on statistical power to study robustness because one can always employ the permutation procedure to control the type I error under $H_0 : P_m = Q_n$.

To state the main result, consider the energy statistic based on a $U$-statistic:

$$E_{m,n} = \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} ||X_i - Y_j|| - \frac{1}{m^2} \sum_{i_1 \neq i_2} ||X_{i_1} - X_{i_2}|| - \frac{1}{n^2} \sum_{j_1 \neq j_2} ||Y_{j_1} - Y_{j_2}||. \quad (10)$$

Under the given setting, the following result shows that the CvM test is uniformly powerful over different types of contaminations as the sample size tends to infinity. However, the energy test becomes asymptotically powerless against certain contaminations.

**Theorem 3.1.** Suppose we observe samples $X_m$ and $Y_n$ from the contaminated model in (9) with a fixed contamination ratio $0 < \epsilon < 1$. Assume that there exist $\delta_1, \delta_2 > 0$ such that

$$E_d(P_m, Q_n) \geq \delta_1 \quad \text{and} \quad W_d(P_m, Q_n) \geq \delta_2, \quad \text{for every } m, n. \quad (11)$$

Consider the tests based on $U_{CvM}$ and $E_{m,n}$, defined by

$$\phi_{CvM} := I(U_{CvM} \geq c_{\alpha,m,n}^*) \quad \text{and} \quad \phi_{\text{Energy}} := I(E_{m,n} \geq c'_{\alpha,m,n}),$$

where $c_{\alpha,m,n}^*$ and $c'_{\alpha,m,n}$ are $\alpha$ level critical values from their permutation distributions. Then the power of the CvM test is asymptotically powerful uniformly over any choice of $(H_m, H'_n)$, whereas the power of the energy test becomes asymptotically powerless against certain choices of $(H_m, H'_n)$, that is

$$\lim_{m,n \to \infty} \inf_{H_m, H'_n} E_1[\phi_{CvM}] = 1 \quad \text{and} \quad \lim_{m,n \to \infty} \inf_{H_m, H'_n} E_1[\phi_{\text{Energy}}] \leq \alpha. \quad (12)$$

To illustrate Theorem 3.1 with finite sample size, we carried out simulation studies based on the normal contamination models. We consider two alternatives with 1) location differences and 2) scale differences. For both examples, we let $H_m, H'_n$ have the same multivariate normal distribution given by

$$H_m = H'_n := N((0, \ldots, 0)^\top, \sigma^2 I_d),$$

where $\sigma$ controls the degree of heavy-tailedness. Note that the condition in (11) is easily met whenever $H_m = H'_n$, but $F_X \neq G_Y$ (with the extra first moment condition for $E_d$), which can be seen from the representations for $E_d^2$ and $W_d^2$ in (3) and (4).

**Example 3.1** (Location difference). For the location alternative, we compare two multivariate normal distributions, where the means are different but the covariance matrices are identical. Specifically, we set

$$F_X = N((-0.5, \ldots, -0.5)^\top, I_d), \quad \text{and} \quad G_Y = N((0.5, \ldots, 0.5)^\top, I_d),$$

with $\epsilon = 0.05$. We then change $\sigma = 1, 40, 80, 120, 160, 200$ and $240$ to investigate the robustness of the tests against heavy-tail contaminations.
Figure 1: Empirical power of $\Phi_{\text{CvM}}$ (solid blue) and $\Phi_{\text{Energy}}$ (dashed red) under the contamination models with $\epsilon = 0.05$. See Example 3.1 and 3.2 for details.

**Example 3.2** (Scale difference). Similar to the location alternative, we again choose multivariate normal distributions which differ in their scale but not in their location parameters. In detail, we have

$$F_X = N((0, \ldots, 0)^\top, 0.1^2 \times I_d), \quad \text{and} \quad G_Y = N((0, \ldots, 0)^\top, I_d),$$

with $\epsilon = 0.05$. Again, we change $\sigma = 1, 40, 80, 120, 160, 200$ and 240 to assess the effect of heavy-tail contaminations.

Experiments were run with $m = n = 40$ and $d = 10$ at significance level $\alpha = 0.05$ and the critical values were decided by the permutation procedure. As can be seen from the result in Figure 1, the power of the CvM test is consistently robust to the change of $\sigma$, while the power of the energy test drops down significantly as $\sigma$ increases. Roughly speaking, when $\sigma$ is large, it is likely that the energy statistic is dominated by extreme observations from $H_m$ and $H_n$. Furthermore, when $H_m$ and $H_n$ are identical (i.e. they are in favor of the null hypothesis), the null and alternative distributions of the energy statistic tend to be close as $\sigma$ increases. As a consequence, the resulting test based on the energy statistic performs poorly. In contrast, the CvM statistic puts less weight on extreme observations, which leads to robust performance of the CvM test under contaminations. A mathematical justification can be found in the proof of Theorem 3.1.

### 4 Minimax Optimality

In this section, we investigate the minimax optimality of the permutation test based on $U_{\text{CvM}}$. To formulate the minimax problem, we define the set of two multivariate distributions which are at least $\epsilon$ far apart in terms of the Cramér-von Mises distance, i.e.

$$\mathcal{F}(\epsilon) := \{(F_X, G_Y) : W_d(F_X, G_Y) \geq \epsilon\}.$$
For a given significance level $\alpha \in (0, 1)$, let $\Phi_{m,n}(\alpha)$ be the set of measurable functions $\phi : \{X_m, Y_n\} \mapsto \{0, 1\}$ such that

$$\Phi_{m,n}(\alpha) = \{\phi : \mathbb{P}_0(\phi = 0) \leq \alpha\}.$$  

We then define the minimax type II error as follows:

$$1 - \beta_{m,n}(\epsilon) = \inf_{\phi \in \Phi_{m,n}(\alpha)} \sup_{F_X, G_Y \in \mathcal{F}(\epsilon)} \mathbb{P}^{m,n}_1(\phi = 0).$$  

Our primary interest is in finding the minimum separation rate $\epsilon_{m,n}$ satisfying

$$\epsilon_{m,n} = \inf \left\{ \epsilon : 1 - \beta_{m,n}(\epsilon) \leq \xi \right\},$$

for some $0 < \xi < 1 - \alpha$.

We start with presenting a lower bound of the multivariate CvM distance.

**Lemma 4.1.** The multivariate CvM distance is lower bounded by

$$W_d(F_X, G_Y) \geq \int_{\mathbb{S}^{d-1}} \left| \frac{1}{2} - \mathbb{P}_{X,Y} \left( \beta^T X \leq \beta^T Y \right) \right| d\lambda(\beta).$$  

Consider two random vectors $X^*$ and $Y^*$ where their first coordinates have normal distributions $N(\mu_X, 1)$ and $N(\mu_Y, 1)$ respectively and the other coordinates have the degenerate distribution at zero, i.e.

$$X^* := (Z_1, 0, \ldots, 0)^\top \quad \text{and} \quad Y^* := (Z_2, 0, \ldots, 0)^\top$$

where $Z_1 \sim N(\mu_X, 1)$ and $Z_2 \sim N(\mu_Y, 1)$. Given $\beta = (\beta_1, \ldots, \beta_d) \in \mathbb{S}^{d-1}$, we have $\beta^T X^* \sim N(\beta_1 \mu_X, \beta_1^2)$ and $\beta^T Y^* \sim N(\beta_1 \mu_Y, \beta_1^2)$; therefore $\beta^T X^*$ and $\beta^T Y^*$ are continuous for all $\beta \in \mathbb{S}^{d-1}$ almost surely. Under this setting, the multivariate CvM distance is lower bounded as follows:

**Lemma 4.2.** Consider random vectors $X^*$ and $Y^*$ described above with $\mu_{X^*} = cm^{-1/2}$ and $\mu_{Y^*} = -c'n^{-1/2}$ for some $c > 0$. Let us denote the corresponding distributions by $F_{X^*}$ and $G_{Y^*}$. Then there exists a constant $C > 0$ independent of the dimension satisfying

$$W_d(F_{X^*}, G_{Y^*}) \geq C \left( \frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}} \right).$$

Furthermore, the lower bound is tight up to constant factors.

**Proof.** From Lemma 4.1, it is enough to show

$$\int_{\mathbb{S}^{d-1}} \left| \frac{1}{2} - \mathbb{P}_{X^*, Y^*} \left( \beta^T X^* \leq \beta^T Y^* \right) \right| d\lambda(\beta) \geq C \left( \frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}} \right).$$

For any fixed $\beta \in \mathbb{S}^{d-1}$, we have $\beta^T (X^* - Y^*) \sim N(\beta_1 (\mu_{X^*} - \mu_{Y^*}), 2\beta_1^2)$ and thus

$$\left| \frac{1}{2} - \mathbb{P}_{X^*, Y^*} \left( \beta^T X^* \leq \beta^T Y^* \right) \right| = \left| \frac{1}{2} - \Phi \left( -\text{sign}(\beta_1) \cdot \frac{c}{\sqrt{2}} \left( \frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}} \right) \right) \right|
\geq \frac{c}{\sqrt{2}} \left( \frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}} \right) \cdot \phi \left( \frac{c}{2\sqrt{2}} \left( \frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}} \right) \right)
\geq \frac{c}{\sqrt{2}} \left( \frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}} \right) \cdot \phi \left( \frac{c}{2\sqrt{2}} \right),$$
where $\Phi(\cdot)$ and $\phi(\cdot)$ are the cumulative distribution function and density function of the standard normal random distribution respectively. This lower bound holds almost surely for all $\beta \in S^{d-1}$ and thus the result follows. To have an upper bound, notice that

$$W^2_d(F_{X^*}, G_{Y^*}) \leq \int_{S^{d-1}} \sup_{t \in \mathbb{R}} (F_{\beta^\top X}(t) - G_{\beta^\top Y}(t))^2 d\lambda(\beta)$$

$$\leq \frac{1}{2} \int_{S^{d-1}} \text{KL}(N(\beta_1 \mu_{X^*}, \Sigma_1), N(\beta_1 \mu_{Y^*}, \Sigma_1)) d\lambda(\beta)$$

$$= \frac{c^2}{2} \left( \frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}} \right)^2,$$

where we used the Pinsker’s inequality for $(i)$. This shows the tightness of the lower bound.

The previous result together with Neyman-Pearson lemma establishes a lower bound for the minimum separation rate in the next theorem.

**Theorem 4.1 (Lower Bound).** For $0 < \xi < 1 - \alpha$, there exists some constant $b = b(\alpha, \xi)$ independent of the dimension such that $\epsilon_{m,n} = b(m^{-1/2} + n^{-1/2})$ and the minimax type II error is lower bounded by $\xi$, i.e.

$$1 - \beta_{m,n}(\epsilon_{m,n}) \geq \xi.$$

Theorem 4.1 shows that no test can have considerable power against all alternatives when $\epsilon_{m,n}$ is of order $m^{-1/2} + n^{-1/2}$; thereby, it provides a lower bound of the minimum separation rate. Now, we show that this lower bound is tight by establishing a matching upper bound. In particular, the upper bound is obtained by the permutation test based on $U_{CV^*}$.

**Theorem 4.2 (Upper Bound).** Consider the test based on $U_{CV^*}$ given by

$$\phi_{CV^*} := I(U_{CV^*} \geq \epsilon^*_\alpha_{m,n})$$

where $\epsilon^*_\alpha_{m,n}$ is the $\alpha$ level critical value from the permutation distribution. For a sufficiently large $c > 0$, let $\epsilon^*_{m,n}$ be the radius of interest defined by

$$\epsilon^*_{m,n} := c \left( \frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}} \right). \quad (15)$$

Then there exists $\xi \in (0, 1 - \alpha)$ such that the type II error of $\phi_{CV^*}$ is uniformly bounded by $\xi$, i.e.

$$\sup_{F_X, G_Y \in \mathcal{F}(\epsilon^*_{m,n})} \mathbb{P}_1(\phi_{CV^*} = 0) < \xi.$$

**Proof.** Note that $\epsilon^*_{\alpha_{m,n}}$ is a random quantity depending on $\mathcal{X}_m \cup \mathcal{Y}_n$. To control the randomness from $\epsilon^*_{\alpha_{m,n}}$, we use a similar idea in Fromont et al. (2013) (see also Albert, 2015) where they considered the quantile of $\epsilon^*_{\alpha_{m,n}}$. Specifically, let $c^*_{\xi/2}$ be the upper $\xi/2$ quantile of the distribution of $\epsilon^*_{\alpha_{m,n}}$. Then it suffices to show that

$$\mathbb{E}_1[U_{CV^*}] \geq c^*_{\xi/2} + \sqrt{\frac{2}{\xi} \text{Var}_1(U_{CV^*})} \quad (16)$$
uniformly over $F_X, G_Y \in \mathcal{F}(\epsilon_{m,n})$ by choosing a sufficiently large $c$. Here $E_1$ and $\text{Var}_1$ are the expectation and the variance under the alternative. In detail, we have

$$P_1(U_{\text{CVM}} < c_{a,m,n}^*) = P_1(U_{\text{CVM}} < c_{a,m,n}^*, c_{a,m,n}^* > c_{\xi/2}^*) + P_1(U_{\text{CVM}} < c_{a,m,n}^*, c_{a,m,n}^* \leq c_{\xi/2}^*)$$

$$\leq P_1(c_{a,m,n}^* > c_{\xi/2}^*) + P_1(U_{\text{CVM}} \leq c_{\xi/2}^*)$$

$$\leq \frac{\xi}{2} + P_1(U_{\text{CVM}} \leq c_{\xi/2}^*),$$

where the second inequality is by the definition of $c_{\xi/2}^*$. To control the second term, we apply

the Chebyshev’s inequality

$$P_1(U_{\text{CVM}} \leq c_{\xi/2}^*) = P_1(U_{\text{CVM}} - E_1[U_{\text{CVM}}] \sqrt{\text{Var}_1(U_{\text{CVM}})} \leq c_{\xi/2}^* - E_1[U_{\text{CVM}}] \sqrt{\text{Var}_1(U_{\text{CVM}})})$$

$$\leq \frac{\text{Var}_1(U_{\text{CVM}})}{(E_1[U_{\text{CVM}}] - c_{\xi/2}^*)^2}$$

$$\leq \frac{\xi}{2},$$

where the last inequality uses (16). Indeed, (16) is true and the details are given in the appendix. Hence, the result follows.

It is worth noting that there are computationally more efficient ways of estimating $W_d^2$. For example, one can use the linear-type statistic defined as

$$L_{\text{CVM}} = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{2} [h(X_{2i-1}, X_{2i}; Y_{2i-1}, Y_{2i}) + h(X_{2i}, X_{2i-1}; Y_{2i}, Y_{2i-1})], \quad (17)$$

where $M = [n/2]$ and $m = n$ for simplicity. While $L_{\text{CVM}}$ is also an unbiased estimator of $W_d^2$ and can be computed in linear time, the test based on $L_{\text{CVM}}$ is notably sub-optimal in terms of minimax power. In particular, the following lemma shows that the oracle test based on $L_{\text{CVM}}$ can have full power only against alternatives shrinking slower than $N^{-1/4}$ rate, whereas the minimax optimal rate is $N^{-1/2}$ when $m = n$. This illustrates a certain trade-off between computational cost and statistical power.

**Lemma 4.3.** Let $\eta_{a,m,n}$ be the $\alpha$ level critical value of the oracle test (see Section 2.2) based on $L_{\text{CVM}}$ in (17) and define the corresponding test function by

$$\phi_{L_{\text{CVM}}} := I(L_{\text{CVM}} \geq \eta_{a,m,n}).$$

Consider a sequence of alternatives such that

$$W_d(F_X, G_Y) \asymp N^{-\varepsilon} \quad \text{where} \quad \varepsilon > 1/4.$$\n
Then for $0 < \alpha < 1/2$,

$$\lim_{m,n \to \infty} P_1(\phi_{L_{\text{CVM}}} = 0) > 1/2.$$
Figure 2: Empirical power of the multivariate CvM tests based on the U-statistic (solid blue), the linear-type statistic (dashed red) and Hotelling’s test (dotted gray) where $X \sim N((-\mu, \ldots, -\mu), I_d)$ and $Y \sim N((\mu, \ldots, \mu), I_d)$.

In addition to the rate optimality, the proposed tests have comparable power to existing optimal tests against certain directions. To illustrate this point, we consider an example of multivariate normal distributions where $X \sim N(\mu_X, I_d)$ and $Y \sim N(\mu_Y, I_d)$. Under this particular setting, it is well-known that Hotelling’s test is asymptotically the most powerful test when the dimension is fixed. On the other hand, when the dimension is close to the sample size, Hotelling’s test performs poorly due to a high variance from the sample precision matrix (Bai and Saranadasa, 1996). We carried out simulation studies with $m = n = 20$ to compare the power of the multivariate CvM test with Hotelling’s $T^2$. For our simulation, we set $\mu_X = (-\mu, \ldots, -\mu)^\top$ and $\mu_Y = (\mu, \ldots, \mu)^\top$ at significance level $\alpha = 0.05$ and the critical values were decided by the permutation procedure. According to the results presented in Figure 2, when the dimension is small ($d = 2$), the CvM test based on $U_{\text{CVM}}$ (denoted by $\phi_{\text{CVM}}$) shows comparable (in fact, almost identical) power to Hotelling’s test. However, it is apparent to see that $\phi_{\text{CVM}}$ outperforms Hotelling’s test when the dimension ($d = 30$) is close to the sample size. Indeed, under the given setting, $\phi_{\text{CVM}}$ has the same limiting power as the high-dimensional mean tests introduced by Chen and Qin (2010) and Chakraborty and Chaudhuri (2017). This is our next topic in Section 5. In general, the test based on the linear statistic shows substantially lower power than $\phi_{\text{CVM}}$ as we expected.

5 Adaptable Power in High-Dimensions

In this section, we study the power of the multivariate CvM test in high-dimensions against a simple location shift model:

$$X = \mu_X + V \quad \text{and} \quad Y = \mu_Y + V',$$

where $V = (V_1, \ldots, V_d)^\top \in \mathbb{R}^d$ and $V_i$’s are i.i.d. having a continuous distribution with $\text{Var}(V_1) = \sigma^2$ and $\mathbb{E}[V_1^8] < \infty$. $V'$ is independent and identically distributed with $V$. In
addition, we consider $\boldsymbol{\mu}_X = (\mu_X, \ldots, \mu_X)^\top \in \mathbb{R}^d$ and $\boldsymbol{\mu}_Y = (\mu_Y, \ldots, \mu_Y)^\top \in \mathbb{R}^d$ such that
\[ d^{1/4-\epsilon}(\boldsymbol{\mu}_X - \boldsymbol{\mu}_Y) \to 0 \quad \text{for any } \epsilon > 0 \text{ as } d \to \infty. \] (19)

There are many types of high-dimensional inference procedures for testing
\[ H_0 : \boldsymbol{\mu}_X = \boldsymbol{\mu}_Y \quad \text{versus} \quad H_1 : \boldsymbol{\mu}_X \neq \boldsymbol{\mu}_Y. \] (20)

For example, Chen and Qin (2010) suggested the test statistic based on the sample means. Specifically, their statistic is based on the $U$-statistic defined by
\[ T_{CQ} = \frac{1}{(m/2)(n/2)} \sum_{i_1, i_2=1}^m \sum_{j_1, j_2=1}^n (X_{i_1} - Y_{j_1})(X_{i_2} - Y_{j_2}). \]

Recently, Chakraborty and Chaudhuri (2017) defined their test statistic based on spatial ranks by
\[ T_{WMW} = \frac{1}{(m/2)(n/2)} \sum_{i_1, i_2=1}^m \sum_{j_1, j_2=1}^n \frac{(X_{i_1} - Y_{j_1})^\top}{||X_{i_1} - Y_{j_1}||} \frac{(X_{i_2} - Y_{j_2})}{||X_{i_2} - Y_{j_2}||}. \]

Somewhat surprisingly, under this given scenario, the limiting power of the tests based on $U_{CQ}$, $E_{m,n}$, $T_{CQ}$ and $T_{WMW}$ tend to be the same when $d \to \infty$ and then $N \to \infty$. (In fact, the latter condition is unnecessary from our proof). This finding is striking because the tests based on $T_{CQ}$ and $T_{WMW}$ ($CQ$ test and WMW test for short, respectively) are specifically designed for high-dimensional mean inference, while the proposed test as well as energy test can be sensitive to much broader alternatives. In other words, the CvM test and the energy test are adaptive to high-dimensional location alternatives. To begin, we introduce the power function of each statistic:
\[ \beta_{CvM} = \mathbb{P} \left( 2\pi \sqrt{3d} \gamma_{m,n}^{-1} \cdot U_{CvM} > -z_\alpha \right), \quad \beta_{Energy} = \mathbb{P} \left( \sqrt{2} \sigma^{-1} \gamma_{m,n}^{-1} \cdot E_{m,n} > -z_\alpha \right), \]
\[ \beta_{CQ} = \mathbb{P} \left( d^{-1/2} \sigma^{-2} \gamma_{m,n}^{-1} \cdot T_{CQ} > -z_\alpha \right), \quad \beta_{WMW} = \mathbb{P} \left( 2\sqrt{d} \gamma_{m,n}^{-1} \cdot T_{WMW} > -z_\alpha \right), \] (21)

where $\gamma_{m,n} = \sqrt{2(m/2)^{-1} + 2(n/2)^{-1} + 4(mn)^{-1}}$. Then we present the following result.

**Theorem 5.1.** Suppose that $X$ and $Y$ satisfy the model assumption in (18) with
\[ \mu_X - \mu_Y = \frac{\psi}{d^{1/4} \sqrt{m+n}} \quad \text{and} \quad \text{Var} [X] = \text{Var} [Y] = \sigma^2 I_d. \]

In addition, we assume that $m/N \to \pi_1 \in (0, 1)$ and $n/N \to \pi_2 \in (0, 1)$ as $N \to \infty$. Then the power of the tests based on $U_{CvM}$, $E_{m,n}$, $T_{CQ}$ and $T_{WMW}$ in (21) are asymptotically the same if we let $d \to \infty$ and then $N \to \infty$. In other words,
\[ \lim_{N \to \infty} \lim_{d \to \infty} \beta_{CvM} = \lim_{N \to \infty} \lim_{d \to \infty} \beta_{Energy} = \lim_{N \to \infty} \lim_{d \to \infty} \beta_{CQ} = \lim_{N \to \infty} \lim_{d \to \infty} \beta_{WMW}, \]
and they converge to
\[ \Phi \left( -z_\alpha + \frac{\pi_1 \pi_2}{\sqrt{2} \sigma^2 \psi^2} \right). \]
We would like to point out that the restricted model assumption in (18) is mainly due to technical convenience. We believe that the main result in Theorem 5.1 is still valid under much weaker assumptions supported by our simulations in Section 8 and the supplementary material. Nevertheless, when the moment assumption is violated, the power of these tests can be entirely different. For instance, our simulations in Section 8 further demonstrate that the CQ and (especially) energy tests perform poorly when $X$ and $Y$ have Cauchy distributions with different location parameters. In contrast, the CvM and WMW tests maintain robust power against the same Cauchy alternative.

6 Connection to the Generalized Energy Distance and MMD

Recall that the energy distance is defined with the Euclidean distance under the finite first moment condition. By considering a semimetric space $(\mathcal{Z}, \rho)$ of negative type, Sejdinovic et al. (2013) generalized the energy distance by

$$E_\rho^2 = 2E_{X,Y} [\rho(X,Y)] - E_{X,X'} [\rho(X,X')] - E_{Y,Y'} [\rho(Y,Y')].$$

They further established the equivalence between the generalized energy distance and the MMD with a kernel induced by $\rho(\cdot, \cdot)$. Given a distance-induced kernel $k(\cdot, \cdot)$, the squared MMD is given by

$$\text{MMD}_k^2 = E_{X,X'} [k(X,X')] + E_{Y,Y'} [k(Y,Y')] - 2E_{X,Y} [k(X,Y)].$$

In this section, we will show that the multivariate CvM distance is a member of the generalized energy distance by the use of the angular distance and thus also a member of the MMD. To start, let $S_X$ and $S_Y$ be the support of $X$ and $Y$, respectively and let $S = S_X \cup S_Y \subseteq \mathbb{R}^d$. Then we define the angular distance as follows:

**Definition 6.1.** Let $W_*$ be a random vector having mixture distribution $(1/2)F_X + (1/2)G_Y$. For $z, z' \in S$, denote the scaled angle between $z - W_*$ and $z' - W_*$ by

$$\rho_{\text{Angle}}(z, z'; W_*) = \frac{1}{\pi} \Psi (z - W_*, z' - W_*).$$

The angular distance is defined as the expected value of the scaled angle:

$$\rho_{\text{Angle}}(z, z') = E_{W_*} [\rho_{\text{Angle}}(z, z'; W_*)]. \quad (22)$$

The next lemma shows that $\rho_{\text{Angle}}$ is a metric of negative type defined on $S$.

**Lemma 6.1.** For $\forall z, z', z'' \in S$ and $\rho_{\text{Angle}}: S \times S \mapsto [0, \infty)$, the following conditions are satisfied:

1. $\rho_{\text{Angle}}(z, z') \geq 0$ and $\rho_{\text{Angle}}(z, z') = 0$ if and only if $z = z'$.
2. $\rho_{\text{Angle}}(z, z') = \rho_{\text{Angle}}(z', z)$.
3. $\rho_{\text{Angle}}(z, z') \leq \rho_{\text{Angle}}(z, z'') + \rho_{\text{Angle}}(z', z'')$.

In addition, for $\forall n \geq 2$, $z_1, \ldots, z_n \in S$, and $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, with $\sum_{i=1}^n \alpha_i = 0$,

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \rho_{\text{Angle}}(z_i, z_j) \leq 0.$$
By the use of the angular distance, we establish the identity between the generalized energy distance and the CvM distance in the next proposition. As a result, we conclude that the multivariate CvM distance is a special case of the generalized energy distance based on the angular distance.

**Proposition 6.1.** Let us consider the angular distance defined in (22). Then the following identity holds:

$$2W_d^2 = 2\mathbb{E}_{X,Y}[\rho_{\text{Angle}}(X,Y)] - \mathbb{E}_{X,X'}[\rho_{\text{Angle}}(X,X')] - \mathbb{E}_{Y,Y'}[\rho_{\text{Angle}}(Y,Y')] .$$

**Remark 6.1.** The angular distance can be generalized by taking the expectation with respect to a different measure. For instance, when the expectation is taken with respect to the Lebesgue measure $\nu$, the generalized angular distance is proportional to the Euclidean distance, i.e.

$$\mathbb{E}_{\nu}[\rho_{\text{Angle}}(z,z';\nu)] = \gamma_d||z - z'||,$$

where $\gamma_d$ solely depends on the dimension (see the proof of Lemma 6.1 for more details). The main difference between the Euclidean distance and the proposed angular distance is that the latter takes into account information from the underlying distribution and is less sensitive to outliers. In this aspect, the introduced angular distance can be viewed as a robust alternative for the Euclidean distance.

### 7 Other Multivariate Extensions via Projection Pursuit

The projection pursuit approach used for the multivariate CvM statistic can be applied to other univariate robust statistics. In this section, we illustrate the utility of the projection pursuit approach by considering several examples including Kendall’s tau, the coefficient by Blum et al. (1961) and the sign covariance (Bergsma and Dassios, 2014). We begin by considering one-sample and two-sample robust statistics. Given a pair of random variables $(X,Y)$, define $Z = X - Y$. The univariate sign test statistic is an estimate of $\theta_{\text{sign}} := \mathbb{P}(Z > 0) - 1/2$ and it is used to test whether

$$H_0 : \mathbb{P}(Z > 0) = 1/2 \quad \text{versus} \quad H_1 : \mathbb{P}(Z > 0) \neq 1/2 .$$

The projection pursuit technique extends $\theta_{\text{sign}}$ to a multivariate case as follows:

**Proposition 7.1 (One-sample sign test statistic).** For i.i.d. random vectors $Z_1, Z_2$ from a multivariate distribution $F_Z$ where $Z \in \mathbb{R}^d$, the projection pursuit approach generalizes $\theta_{\text{sign}}$ as

$$\int_{d-1} \left( \mathbb{P}(\beta^T Z_1 > 0) - \frac{1}{2} \right)^2 d\lambda(\beta) = \frac{1}{4} - \frac{1}{2\pi} \mathbb{E}[\Psi(Z_1, Z_2)] .$$

**Proof.** Given $\beta \in \mathbb{R}^{d-1}$, note that

$$\left( \mathbb{P}(\beta^T Z_1 > 0) - \frac{1}{2} \right)^2 = \frac{1}{4} - \mathbb{E}[I(\beta^T Z_1 > 0)] + \mathbb{E}[I(\beta^T Z_1 > 0, \beta^T Z_2 > 0)] .$$

16
Apply Lemma 2.2 with Fubini’s theorem to have
\[
\mathbb{E} \left[ \int_{S^{d-1}} I(\beta^T Z_1 > 0) d\lambda(\beta) \right] = \frac{1}{2}
\]
\[
\mathbb{E} \left[ \int_{S^{d-1}} I(\beta^T Z_1 > 0, \beta^T Z_2 > 0) d\lambda(\beta) \right] = \frac{1}{2} - \frac{1}{2\pi} \mathbb{E} [\Psi (Z_1, Z_2)].
\]
This completes the proof.

Given univariate two samples \(X_m = \{X_1, \ldots, X_m\}\) and \(Y_n = \{Y_1, \ldots, Y_n\}\), the Wilcoxon-Mann-Whitney test aims at distinguishing
\[H_0: \mathbb{P}(X > Y) = \frac{1}{2} \text{ versus } H_1: \mathbb{P}(X > Y) \neq \frac{1}{2},\]
and its test statistic is based on an estimate of \(\theta_{WMW} := \mathbb{P}(X > Y) - 1/2\). The next proposition extends \(\theta_{WMW}\) to a multivariate case via projection pursuit.

**Proposition 7.2** (Two-sample Wilcoxon-Mann-Whitney test statistic). Let \(X_1, X_2 \overset{i.i.d.}{\sim} F_X\) and \(Y_1, Y_2 \overset{i.i.d.}{\sim} G_Y\) where \(X, Y \in \mathbb{R}^d\). The projection pursuit approach generalizes \(\theta_{WMW}\) as
\[
\int_{S^{d-1}} \left( \mathbb{P}(\beta^T X > \beta^T Y) - \frac{1}{2} \right)^2 d\lambda(\beta) = \frac{1}{4} - \frac{1}{2\pi} \mathbb{E} [\Psi (X_1 - Y_1, X_2 - Y_2)].
\]

**Proof.** The result follows by replacing \(Z_1, Z_2\) with \(X_1 - Y_1, X_2 - Y_2\) in Proposition 7.1.

**Remark 7.1.** The first order Taylor approximation of the arccosine function shows that the representations given in the right-side of (23) and (24) are related to the spatial sign-statistics introduced by Wang et al. (2015) and Chakraborty and Chaudhuri (2017), respectively. We believe, however, that our projection pursuit-type statistics — which can be viewed as the average of univariate statistics based on projected random variables — is more intuitive to understand.

The same technique can be further applied to some robust statistics for independence testing. To test for independence between two random variables, Kendall’s tau statistic is defined as an estimate of \(\tau = 4 \mathbb{P}(X_1 < X_2, Y_1 < Y_2) - 1\). We present a multivariate extension of \(\tau\) as follows:

**Theorem 7.1** (Kendall’s tau). For \(i.i.d.\) pairs of random vectors \((X_1, Y_1), \ldots, (X_4, Y_4)\) from a joint distribution \(F_{XY}\) where \(X \in \mathbb{R}^p\) and \(Y \in \mathbb{R}^q\), the multivariate extension of \(\tau\) via projection pursuit is given by
\[
\int_{S^{p-1}} \int_{S^{q-1}} \left[ 4 \mathbb{P} \left( \alpha^T (X_1 - X_2) < 0, \beta^T (Y_1 - Y_2) < 0 \right) - 1 \right]^2 d\lambda(\alpha) d\lambda(\beta)
\]
\[
= \mathbb{E} \left[ \left( 2 - \frac{2}{\pi} \Psi (X_1 - X_2, X_3 - X_4) \right) \left( 2 - \frac{2}{\pi} \Psi (Y_1 - Y_2, Y_3 - Y_4) \right) \right] - 1.
\]

Kendall’s tau has been frequently used in practice due to its robustness, simplicity and interpretability. Nonetheless, the main limitation of Kendall’s tau is that it can be zero even when there exists a certain association between random variables. There have been alternative
approaches to resolve this issue in the literature. For a multivariate case, Zhu et al. (2017) extended Hoeffding’s coefficient (Hoeffding, 1948) via projection pursuit. Specifically, they defined the projection correlation between $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ as

$$
\int_{S^{p-1}} \int_{S^{q-1}} \int_{\mathbb{R}^2} \left[ F_{\alpha'X,\beta'Y}(u,v) - F_{\alpha'X}(u)F_{\beta'Y}(v) \right]^2 d\omega_1(u,v,\alpha,\beta),
$$

(25)

where $d\omega_1(u,v,\alpha,\beta) = dF_{\alpha'X,\beta'Y}(u,v)d\lambda(\alpha)d\lambda(\beta)$. Although the projection correlation is more broadly sensitive than Kendall’s tau is in detecting dependence among random variables, it can still be zero even when $X$ and $Y$ are dependent. A counterexample for the univariate case can be found in Hoeffding (1948).

On the other hand, the coefficient introduced by Blum et al. (1961) overcomes this issue by replacing $dF_{XY}$ with $dF_XdF_Y$. The univariate Blum-Kiefer-Rosenblatt (BKR) coefficient (Blum et al., 1961) is defined by

$$
\int_{\mathbb{R}^2} \left[ F_{XY}(u,v) - F_X(u)F_Y(v) \right]^2 dF_X(u)dF_Y(v).
$$

Next, we generalize the univariate BKR coefficient to a multivariate space via projection pursuit.

**Theorem 7.2** (Blum–Kiefer–Rosenblatt (BKR) coefficient). Let us consider weight function $d\omega_2(u,v,\alpha,\beta) = dF_{\alpha'X}(u)dF_{\beta'Y}(v)d\lambda(\alpha)d\lambda(\beta)$. For i.i.d. random vectors $(X_1,Y_1),\ldots,(X_6,Y_6)$ from a joint distribution $F_{XY}$ where $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, the univariate BKR coefficient can be extended to a multivariate case by

$$
\int_{S^{p-1}} \int_{S^{q-1}} \int_{\mathbb{R}^2} \left[ F_{\alpha'X,\beta'Y}(u,v) - F_{\alpha'X}(u)F_{\beta'Y}(v) \right]^2 d\omega_2(u,v,\alpha,\beta)
$$

$$
= E \left[ \left( \frac{1}{2} - \frac{1}{2\pi} \Psi (X_1 - X_3, X_2 - X_3) \right) \cdot \left( \frac{1}{2} - \frac{1}{2\pi} \Psi (Y_1 - Y_4, Y_2 - Y_4) \right) \right]
$$

$$
+ E \left[ \left( \frac{1}{2} - \frac{1}{2\pi} \Psi (X_1 - X_5, X_2 - X_5) \right) \cdot \left( \frac{1}{2} - \frac{1}{2\pi} \Psi (Y_3 - Y_6, Y_4 - Y_6) \right) \right]
$$

$$
- 2E \left[ \left( \frac{1}{2} - \frac{1}{2\pi} \Psi (X_1 - X_4, X_2 - X_4) \right) \cdot \left( \frac{1}{2} - \frac{1}{2\pi} \Psi (Y_1 - Y_5, Y_3 - Y_5) \right) \right].
$$

Recently, Bergsma and Dassios (2014) introduce a modification of Kendall’s tau, which is zero if and only if random variables are independent. Let us denote the univariate Bergsma-Dassios sign covariation by

$$
\tau^* = E \left[ a_{\text{sign}}(X_1, X_2, X_3, X_4) \cdot a_{\text{sign}}(Y_1, Y_2, Y_3, Y_4) \right],
$$

(26)

with $a_{\text{sign}}(z_1, z_2, z_3, z_4) = \text{sign}(|z_1 - z_2| + |z_3 - z_4| - |z_1 - z_3| - |z_2 - z_4|)$. Motivated by the projection pursuit approach, we propose the multivariate $\tau^*$ as follows:

**Definition 7.1** (Multivariate $\tau^*$). Suppose $(X_1,Y_1),\ldots,(X_4,Y_4)$ are i.i.d. random vectors from a joint distribution $F_{XY}$ where $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$. We define the multivariate $\tau^*$ by

$$
\tau^*_{p,q} = \int_{S^{p-1}} \int_{S^{q-1}} E \left[ a_{\text{sign}}(\alpha'X_1, \alpha'X_2, \alpha'X_3, \alpha'X_4) \right. \times a_{\text{sign}}(\beta'Y_1, \beta'Y_2, \beta'Y_3, \beta'Y_4) \left. \right] d\lambda(\alpha)d\lambda(\beta).
$$
Since the kernel of $\tau^*$ is sign-invariant, i.e.

$$a_{\text{sign}}(z_1, z_2, z_3, z_4) = a_{\text{sign}}(-z_1, -z_2, -z_3, -z_4),$$

it is easy to see that $\tau^*_{p,q}$ becomes the univariate $\tau^*$ when $p = q = 1$. Also, note that since $X$ and $Y$ are independent if and only if $\alpha^T X$ and $\beta^T Y$ are independent for all $\alpha \in \mathbb{S}^{p-1}$ and $\beta^T \in \mathbb{S}^{q-1}$, the characteristic property of $\tau^*_{p,q}$ follows by that of the univariate $\tau^*$.

To have an expression for $\tau^*_{p,q}$ without involving integrations over the unit sphere, we first generalize Lemma 2.2 with three indicator functions presented in Lemma 7.1. Then based on this result, we provide an alternative expression for $\tau^*_{p,q}$ in Theorem 7.3.

**Lemma 7.1.** For arbitrary vectors $U_1, U_2, U_3 \in \mathbb{R}^d$, the following identity holds

$$\int_{\mathbb{S}^{d-1}} I(\beta^T U_1 \leq 0) I(\beta^T U_2 \leq 0) I(\beta^T U_3 \leq 0) d\lambda(\beta)$$

$$= \frac{1}{2} - \frac{1}{4\pi} \left[ \Psi(U_1, U_2) + \Psi(U_1, U_3) + \Psi(U_2, U_3) \right].$$

For $U_1, U_2, U_3 \in \mathbb{R}^d$, define $g_d(U_1, U_2, U_3)$ and $h_d(Z_1, Z_2, Z_3, Z_4)$ by

$$g_d(U_1, U_2, U_3) = \frac{1}{2} - \frac{1}{4\pi} \left[ \Psi(U_1, U_2) + \Psi(U_1, U_3) + \Psi(U_2, U_3) \right]$$

and

$$h_d(Z_1, Z_2, Z_3, Z_4)$$

$$= g_d(Z_1 - Z_2, Z_2 - Z_3, Z_3 - Z_4) + g_d(Z_2 - Z_1, Z_1 - Z_3, Z_3 - Z_4)$$

$$+ g_d(Z_1 - Z_2, Z_2 - Z_4, Z_4 - Z_3) + g_d(Z_2 - Z_1, Z_1 - Z_4, Z_4 - Z_3).$$

Based on the kernel $h_d$, we present an alternative expression for $\tau^*_{p,q}$ as follows:

**Theorem 7.3.** For i.i.d. random vectors $(X_1, Y_1), \ldots, (X_4, Y_4)$ from a joint distribution $F_{XY}$ where $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, $\tau^*_{p,q}$ can be written as

$$\tau^*_{p,q} = \mathbb{E} [h_p(X_1, X_2, X_3, X_4) \cdot h_q(Y_1, Y_2, Y_3, Y_4)]$$

$$+ \mathbb{E} [h_p(X_1, X_2, X_3, X_4) \cdot h_q(Y_3, Y_4, Y_1, Y_2)]$$

$$- 2\mathbb{E} [h_p(X_1, X_2, X_3, X_4) \cdot h_q(Y_1, Y_3, Y_2, Y_4)].$$

Theorem 7.3 leads to a straightforward empirical estimate of $\tau^*_{p,q}$ based on a $U$-statistic. This is also true for the other multivariate generalizations introduced in this section. Using these estimates, some theoretical and empirical properties of the proposed measures can be further investigated. These topics are reserved for future work.

**Remark 7.2.** Dhar et al. (2016) studied the robustness of the univariate $\tau^*$ and contrasted it to the non-robustness of the distance covariance (Székely et al., 2007). One can similarly show that the multivariate $\tau^*_{p,q}$ also retains the robustness in terms of the maximum bias functional considered in Dhar et al. (2016).
8 Simulations

In this section, we present some simulation examples to support the argument in Section 5 as well as to illustrate the finite sample performance of the proposed test. Throughout our experiments, the significance level was set at 0.05 and the permutation procedure was used to set up the critical value. Since the computational costs of $U$-statistics together with permutation tests are expensive for large $n$ and $d$, we set the sample size and the dimension relatively small by $m = n = 20$ and $d = 200$, but which are sufficient to emphasize our claim.

First, we consider several examples where the power of the considered tests in Section 5 are approximately equivalent to each other. Here we use multivariate normal distributions with different means $\mu^{(0)} = (0, \ldots, 0)^T$, $\mu^{(1)} = (0.1, \ldots, 0.1)^T$ and $\mu^{(2)} = \sqrt{2}(0.1, \ldots, 0.1, 0, \ldots, 0)^T$, $d/2$ elements $d/2$ elements and covariance matrices:

1. Identity matrix (denoted by $I$) where $\sigma_{i,i} = 1$ and $\sigma_{i,j} = 0$ for $i \neq j$.
2. Banded matrix (denoted by $\Sigma_{Band}$) where $\sigma_{i,i} = 1$, $\sigma_{i,j} = 0.6$ for $|i - j| = 1$, $\sigma_{i,j} = 0.3$ for $|i - j| = 2$ and $\sigma_{i,j} = 0$ otherwise.
3. Autocorrelation matrix (denoted by $\Sigma_{Auto}$) where $\sigma_{i,i} = 1$ and $\sigma_{i,j} = 0.2^{|i-j|}$ when $i \neq j$.
4. Block diagonal matrix (denoted by $\Sigma_{Block}$) where the $5 \times 5$ main diagonal blocks $A$ are defined by $a_{i,i} = 1$ and $a_{i,j} = 0.2$ when $i \neq j$, and the off-diagonal blocks are zeros.

Then we generate random samples from $X \sim N(\mu^{(0)}, \Sigma)$ and either $Y \sim N(\mu^{(1)}, \Sigma)$ or $Y \sim N(\mu^{(2)}, \Sigma)$. The results are summarized in Table 2. As we expected, the empirical power of the considered tests are close when the moment assumptions are met and the covariates are independent or weakly dependent. More empirical evidence can be found in the supplementary material.

Table 2: Empirical power of the considered tests against the normal location models at $\alpha = 0.05$.

| methods | $I$ | $\Sigma_{Band}$ | $\Sigma_{Auto}$ | $\Sigma_{Block}$ |
|---------|-----|-----------------|-----------------|-----------------|
| $\mu^{(1)}$ | 0.251 | 0.271 | 0.266 | 0.257 |
| $\mu^{(2)}$ | 0.266 | 0.191 | 0.246 | 0.267 |
| CvM      | 0.246 | 0.261 | 0.261 | 0.267 |
| Energy   | 0.256 | 0.251 | 0.251 | 0.256 |
| CQ       | 0.256 | 0.266 | 0.261 | 0.267 |
| WMW      | 0.241 | 0.266 | 0.261 | 0.267 |

In our second experiment, we consider several examples where the moment conditions are not satisfied. Simply, we focus on random samples generated from multivariate Cauchy distributions. Let $\text{Cauchy}(\gamma, s)$ denote the univariate Cauchy distribution where $\gamma, s$ are the location parameter and the scale parameter, respectively. Let $X = (X^{(1)}, \ldots, X^{(d)})$ and $Y = (Y^{(1)}, \ldots, Y^{(d)})$ be random vectors where $X^{(i)} \overset{i.i.d.}{\sim} \text{Cauchy}(0, 1)$ and $Y^{(i)} \overset{i.i.d.}{\sim} \text{Cauchy}(\gamma, s)$ for $i = 1, \ldots, d$. We first consider location differences where $\gamma$ is not zero but the scale parameters are identical, i.e. $s = 1$. Similarly, we consider scale differences where the scale parameter $s$ changes, but the location parameters are identical, i.e. $\gamma = 0$. We also report
Table 3: Empirical power of the considered tests against multivariate Cauchy distributions at $\alpha = 0.05$ where $\gamma, s$ represent the location and scale parameter, respectively.

| $\gamma$ | 0 | 1 | 2 | 3 | 4 | 5 |
|-----------|---|---|---|---|---|---|
| Location  |   |   |   |   |   |   |
| $s$       |   |   |   |   |   |   |
| CvM       | 0.146 | 0.231 | 0.593 | 0.874 | 0.628 | 0.940 | 0.990 | 1.000 |
| Energy    | 0.075 | 0.065 | 0.136 | 0.131 | 0.342 | 0.678 | 0.844 | 0.859 |
| CQ        | 0.136 | 0.246 | 0.322 | 0.442 | 0.070 | 0.060 | 0.055 | 0.090 |
| WMW       | 0.296 | 0.643 | 0.915 | 0.985 | 0.090 | 0.060 | 0.055 | 0.106 |
| MMD       | 0.101 | 0.101 | 0.161 | 0.171 | 0.492 | 0.794 | 0.945 | 0.970 |
| NN        | 0.281 | 0.683 | 0.869 | 0.970 | 0.221 | 0.211 | 0.196 | 0.241 |
| Scale     |   |   |   |   |   |   |
| $s$       |   |   |   |   |   |   |
| CvM       | 0.940 | 0.990 | 1.000 |
| MMD       | 0.678 | 0.844 |
| Energy    | 0.060 | 0.055 |
| NN        | 0.794 | 0.945 |

the power results based on the Gaussian MMD with the median heuristic (Gretton et al., 2012) and $k$-NN with $k = 3$ (Schilling, 1986) as reference points. These tests are denoted by MMD and NN, respectively. From the results presented in Table 3, we can see that, unlike the multivariate normal cases, there are significant differences between power performance among CvM, Energy, CQ and WMW tests. In particular, the tests based on the Energy and CQ statistics have relatively low power against the heavy-tail location alternatives, whereas the tests based on the CvM and WMW statistics show better performance than the others. Turning to the scale problems, it can be seen that the CQ and WMW tests have no power to detect scale differences, which makes sense because they are specifically designed for location problems. On the other hand, the CvM and Energy tests are reasonably sensitive to these alternatives. Among the omnibus tests, the MMD test shows a good performance against the scale differences, but not against the location differences. The NN test outperforms the other omnibus tests against the location alternatives; however, it has low power against the scale alternatives. In summary, the overall ranks among the omnibus tests against the Cauchy alternatives are

1. Cauchy location: NN $\succ$ CvM $\succ$ MMD $\succ$ Energy.
2. Cauchy scale: CvM $\succ$ MMD $\succ$ Energy $\succ$ NN.

9 Concluding Remarks

In this work, we extended the univariate Cramér-von Mises statistic for two-sample testing to the multivariate case using projection pursuit. The proposed statistic has a straightforward calculation formula in arbitrary dimensions and the resulting test has good properties. Throughout this paper, we demonstrated its robustness, minimax optimality and high-dimensional power properties. In addition, we applied the same projection technique to other robust statistics and presented their multivariate extensions.

Beyond nonparametric testing problems, we believe that our approach can be used for other problems. For example, our work can be viewed as an application of the angular distance to the two-sample problem. The angular distance is closely connected to the Euclidean distance (Remark 6.1) but is more robust to outliers by incorporating information from the underlying distribution. Given that the use of distances is of fundamental importance in many statistical applications (including clustering, classification and regression), we expect that the
angular distance can be applied to other statistical problems as a robust alternative for the Euclidean distance.

References

Albert, M. (2015). *Tests of independence by bootstrap and permutation: an asymptotic and non-asymptotic study. Application to neurosciences*. PhD thesis, Université Nice Sophia Antipolis.

Anderson, N. H., Hall, P., and Titterington, D. M. (1994). Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 50(1):41–54.

Anderson, T. W. (1962). On the distribution of the two-sample cramér-von mises criterion. *The Annals of Mathematical Statistics*, pages 1148–1159.

Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, pages 311–329.

Baraud, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5):577–606.

Baringhaus, L. and Franz, C. (2004). On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88(1):190–206.

Baringhaus, L. and Henze, N. (2017). Cramér–von mises distance: probabilistic interpretation, confidence intervals, and neighbourhood-of-model validation. *Journal of Nonparametric Statistics*, 29(2):167–188.

Bergsma, W. and Dassios, A. (2014). A consistent test of independence based on a sign covariance related to kendall’s tau. *Bernoulli*, 20(2):1006–1028.

Bhat, B. V. (1995). *Theory of U-statistics and its applications*. PhD thesis, Karnataka University.

Blum, J. R., Kiefer, J., and Rosenblatt, M. (1961). Distribution free tests of independence based on the sample distribution function. *The Annals of Mathematical Statistics*, pages 485–498.

Bogomolny, E., Bohigas, O., and Schmit, C. (2007). Distance matrices and isometric embeddings. *arXiv preprint arXiv:0710.2063*.

Chakraborty, A. and Chaudhuri, P. (2017). Tests for high-dimensional data based on means, spatial signs and spatial ranks. *The Annals of Statistics*, 45(2):771–799.

Chen, H. and Friedman, J. H. (2017). A new graph-based two-sample test for multivariate and object data. *Journal of the American Statistical Association*, 112(517):397–409.

Chen, S. X. and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*, pages 808–835.

Chung, E. and Romano, J. P. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics*, pages 484–507.
Chung, E. and Romano, J. P. (2016). Asymptotically valid and exact permutation tests based on two-sample u-statistics. *Journal of Statistical Planning and Inference*, 168:97–105.

Cui, H. (2002). Average projection type weighted cramér-von mises statistics for testing some distributions. *Science in China Series A: Mathematics*, 45(5):562–577.

Dhar, S. S., Dassios, A., and Bergsma, W. (2016). A study of the power and robustness of a new test for independence against contiguous alternatives. *Electronic Journal of Statistics*, 10(1):330–351.

Escanciano, J. C. (2006). A consistent diagnostic test for regression models using projections. *Econometric Theory*, 22(6):1030–1051.

Friedman, J. H. and Rafsky, L. C. (1979). Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of Statistics*, pages 697–717.

Fromont, M., Laurent, B., and Reynaud-Bouret, P. (2013). The two-sample problem for poisson processes: Adaptive tests with a nonasymptotic wild bootstrap approach. *The Annals of Statistics*, pages 1431–1461.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773.

Henze, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *The Annals of Statistics*, pages 772–783.

Hoeffding, W. (1948). A non-parametric test of independence. *The Annals of Mathematical Statistics*, pages 546–557.

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics*, pages 169–192.

Kowalski, J. and Tu, X. M. (2008). *Modern applied U-statistics*, volume 714. John Wiley & Sons.

Lee, J. (1990). U-statistics: Theory and practice.

Lehmann, E. L. (1951). Consistency and unbiasedness of certain nonparametric tests. *The Annals of Mathematical Statistics*, pages 165–179.

Lehmann, E. L. and Romano, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

Liu, R. Y. (2006). *Data depth: robust multivariate analysis, computational geometry, and applications*, volume 72. American Mathematical Soc.

Mondal, P. K., Biswas, M., and Ghosh, A. K. (2015). On high dimensional two-sample tests based on nearest neighbors. *Journal of Multivariate Analysis*, 141:168–178.

Oja, H. (2010). *Multivariate nonparametric methods with R: an approach based on spatial signs and ranks*. Springer Science & Business Media.

Ramdas, A., Reddi, S. J., Poczos, B., Singh, A., and Wasserman, L. (2015). Adaptivity and computation-statistics tradeoffs for kernel and distance based high dimensional two sample testing. *arXiv preprint arXiv:1508.00655*. 

23
Schilling, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, 81(395):799–806.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, pages 2263–2291.

Székely, G. J. and Rizzo, M. L. (2004). Testing for equal distributions in high dimension. *InterStat*, 5(16.10).

Székely, G. J. and Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794.

Thas, O. (2010). *Comparing distributions*. Springer.

Wang, L., Peng, B., and Li, R. (2015). A high-dimensional nonparametric multivariate test for mean vector. *Journal of the American Statistical Association*, 110(512):1658–1669.

Zhu, L., Xu, K., Li, R., and Zhong, W. (2017). Projection correlation between two random vectors. *Biometrika*.

Zhu, L.-X., Fang, K.-T., and Bhatti, M. I. (1997). On estimated projection pursuit-type cramér-von mises statistics. *Journal of Multivariate Analysis*, 63(1):1–14.
A Permutation Tests

Permutation tests are among the most widely used conditional procedures of nonparametric inference. An essential feature of the permutation tests is that they are guaranteed to obtain an exact rejection probability whenever the exchangeability condition is satisfied under $H_0$. However, except for the finite sample exactness, there has been little theory for permutation tests. The goal here is to enhance our understanding of permutation tests for the two-sample problem. Specifically, we would like to establish fairly general conditions under which the permutation distribution is asymptotically equivalent to the corresponding unconditional null distribution based on a $U$-statistic. We demonstrate our results using the proposed CvM statistic.

A.1 Asymptotic null behavior of permutation $U$-statistics

In this section, we study the limiting null behavior of the permutation distribution based on a degenerating $U$-statistic. The limiting behavior under the alternative will be studied in the next section based on the coupling argument.

We start with introducing some notations. Consider a symmetric kernel $g(X_1,\ldots,X_r;Y_1,\ldots,Y_r)$ of degree $(r,r)$ such that

$$E\left[g(X_1,\ldots,X_r;Y_1,\ldots,Y_r)\right] = \theta,$$

$$E\left[\left\{g(X_1,\ldots,X_r;Y_1,\ldots,Y_r)\right\}^2\right] < \infty.$$  

(27)

Let $U_{m,n}$ be the corresponding $U$-statistic:

$$U_{m,n} = \frac{1}{(m)_r(n)_r} \sum_{\alpha \in A_m} \sum_{\beta \in B_n} g(X_{\alpha_1},\ldots,X_{\alpha_r};Y_{\beta_1},\ldots,Y_{\beta_r}),$$

where $A_m$ and $B_n$ denote the set of all subsets that consist of $r$ unique elements from $\{1,\ldots,m\}$ and $\{1,\ldots,n\}$, respectively. Let $g_{c,d}(x_1,\ldots,x_c;y_1,\ldots,y_d)$ be the conditional expectation given by

$$g_{c,d}(x_1,\ldots,x_c;y_1,\ldots,y_d) := E\left[g(x_1,\ldots,x_c,X_{c+1},\ldots,X_r;y_1,\ldots,y_d,Y_{d+1},\ldots,Y_r)\right].$$  

(28)

Further define

$$g_{c,d}^*(x_1,\ldots,x_c;y_1,\ldots,y_d) := g_{c,d}(x_1,\ldots,x_c;y_1,\ldots,y_d) - \theta$$

and

$$\sigma_{c,d}^2 := \text{Var}\left[g_{c,d}(X_1,\ldots,X_c;Y_1,\ldots,Y_d)\right] = E\left[\left\{g_{c,d}^*(X_1,\ldots,X_c;Y_1,\ldots,Y_d)\right\}^2\right].$$  

(29)

We say the kernel $g$ is non-degenerate if both $\sigma_{0,1}$ and $\sigma_{1,0}$ are positive, and degenerate if $\sigma_{0,1} = \sigma_{1,0} = 0$.

It is well-known that a $U$-statistic has different limiting behaviors depending on the existence of degeneracy (see e.g. Lee, 1990). When the kernel is non-degenerate, Chung and Romano (2016) provide a sufficient condition under which the permutation distribution approximates the unconditional distribution of $U_{m,n}$. In this work, we would like to develop a similar result for a degenerate $U$-statistic.

To begin, consider the centered $U$-statistic scaled by $N = m + n$:

$$U_{m,n}^*(X_1,\ldots,X_m,Y_1,\ldots,Y_n) := N(U_{m,n} - \theta),$$

25
and let \( \{Z_1, \ldots, Z_{m+n}\} = \{X_1, \ldots, X_m, Y_1, \ldots, Y_n\} \) be the pooled samples. Using these notations, the permutation distribution of \( U^*_{m,n} \) is given by

\[
\hat{R}_{m,n}(t) = \frac{1}{N!} \sum_{\pi \in G_N} I\{U^*_{m,n}(Z_{\pi(1)}, \ldots, Z_{\pi(N)}) \leq t\},
\]

where \( G_N \) is the set of all permutations of \{1, \ldots, N\}. Also, let \( R(t) \) be the unconditional limiting distribution of \( U^*_{m,n} \). Then we present the following Theorem.

**Theorem A.1.** Suppose the kernel \( g \) is degenerate with \( \mathbb{E}[g^2] < \infty \) and satisfies

Condition 1. \( g_{0,2}^*(z_1, z_2) = g_{2,0}^*(z_1, z_2) \) and \( g_{1,1}^*(z_1, z_2) = \frac{1-r}{r} g_{0,2}^*(z_1, z_2) \)

Condition 2. \( \sigma_{0,1}^2 = \sigma_{1,0}^2 = 0 \) and \( \sigma_{0,2}^2, \sigma_{2,0}^2, \sigma_{1,1}^2 > 0 \).

Condition 3. \( m/N \to \pi_1 \in (0, 1) \) and \( n/N \to \pi_2 \in (0, 1) \) as \( N \to \infty \).

Then under the null hypothesis,

\[
\sup_{t \in \mathbb{R}} \left| \hat{R}_{m,n}(t) - R(t) \right| \overset{p}{\to} 0. \tag{30}
\]

**Proof.** The proof can be found in Section C.17. \( \square \)

### A.2 The coupling argument

The proof of Theorem A.1 relies on the fact that \( Z_{\pi(1)}, \ldots, Z_{\pi(N)} \) are i.i.d. samples under the null hypothesis for any permutations. The main difficulty of generalizing this result to the alternative hypothesis is that the given samples are not identically distributed. We instead have \( m \) samples \( \{X_1, \ldots, X_m\} \) from \( F_X \) and \( n \) samples \( \{Y_1, \ldots, Y_n\} \) from \( G_Y \). To overcome the difficulty from non i.i.d. samples, we employ the coupling argument considered in Chung and Romano (2013), which can be summarized as follows:

**Algorithm 1:** Coupling

**Data:** \( \{Z_1, \ldots, Z_N\} := \{X_1, \ldots, X_m, Y_1, \ldots, Y_n\} \) where \( \{X_1, \ldots, X_m\} \overset{i.i.d.}{\sim} F_X \) and \( \{Y_1, \ldots, Y_n\} \overset{i.i.d.}{\sim} G_Y \), a random permutation \( \pi_0 \) of \{1, \ldots, N\}.

**Result:** \( \{\bar{Z}_{\pi_0(1)}, \ldots, \bar{Z}_{\pi_0(N)}\} \).

begin

\[ B \sim \text{Binomial}(N, m/N); \]

if \( B \geq m \) then

Generate \( \{X_{m+1}, \ldots, X_B\} \) i.i.d. samples from \( F_X \);

return \( \{\bar{Z}_{\pi_0(1)}, \ldots, \bar{Z}_{\pi_0(N)}\} := \{X_1, \ldots, X_m, Y_1, \ldots, Y_{N-B}, X_{m+1}, \ldots, X_B\} \);

end

else

Generate \( \{Y_{n+1}, \ldots, Y_{N-B}\} \) i.i.d. samples from \( G_X \);

return \( \{\bar{Z}_{\pi_0(1)}, \ldots, \bar{Z}_{\pi_0(N)}\} := \{X_1, \ldots, X_B, Y_{n+1}, \ldots, Y_{N-B}, Y_1, \ldots, Y_n\} \);

end

end
Note that the output of Algorithm 1 consists of \( i.i.d. \) samples from \( \frac{m}{N} F_X + \frac{n}{N} G_Y \). Also note that there are \( D = |m - n| \) different observations between the original samples \( \{Z_1, \ldots, Z_N\} \) and the coupled samples \( \{\tilde{Z}_{\pi_0(1)}, \ldots, \tilde{Z}_{\pi_0(N)}\} \) and \( D = O_P(N^{-1/2}). \) The main strategy of studying the permutation distribution under the alternative is to show that

\[
U_{m,n}^*(Z_{\pi(1)}, \ldots, Z_{\pi(N)}) = U_{m,n}^*(\tilde{Z}_{\pi_0(1)}, \ldots, \tilde{Z}_{\pi_0(N)}) + o_P(1).
\]

If this is the case, then both statistics have the same limiting behavior and we can work with \( i.i.d. \) samples to apply Theorem A.1. We demonstrate this procedure by using the proposed CvM statistic and prove Theorem 2.3 in the main text. The details can be found in the proof of Theorem 2.3.

**Remark A.1.** Note that the coupling argument in Chung and Romano (2013) requires the following assumption:

\[
\frac{m}{N} - \pi_1 = O\left(\frac{1}{\sqrt{N}}\right), \tag{31}
\]

which turns out to be unnecessary in our application; we only need the assumption that \( m/N \to \pi_1 \) and \( b/N \to \pi_2 \) as \( N \to \infty \) without a specific rate. To remove the assumption in (31), we first show that the test statistic based on permuted samples is close to that based on \( i.i.d. \) samples from \( \frac{m}{N} F_X + \frac{n}{N} G_Y \). Then we will show that the two test statistic — one is based on \( i.i.d. \) samples from \( \frac{m}{N} F_X + \frac{n}{N} G_Y \) and the other one is based on \( i.i.d. \) samples from \( \pi_1 F_X + \pi_2 G_Y \) — have the same asymptotic behavior.

## B Auxiliary Lemmas

In this section, we collect some auxiliary lemmas used in our main proofs. To begin, we present another expression for the CvM distance (Lemma B.1) and the proposed statistic (Lemma B.2).

**Lemma B.1** (Another expression for the CvM distance). Let \( X, X', X'' \overset{i.i.d.}{\sim} F_X \) and \( Y, Y', Y'' \overset{i.i.d.}{\sim} G_Y \). The squared multivariate CvM distance can be written as

\[
W_d^2 = \frac{1}{2\pi} \mathbb{E}\left[ \Psi(X - X', Y - Y') \right] + \frac{1}{2\pi} \mathbb{E}\left[ \Psi(X - Y', Y - Y') \right] - \frac{1}{4\pi} \mathbb{E}\left[ \Psi(X - X'', X' - X'') \right] \\
- \frac{1}{4\pi} \mathbb{E}\left[ \Psi(X - Y, X' - Y) \right] - \frac{1}{4\pi} \mathbb{E}\left[ \Psi(Y - Y'', Y' - Y'') \right] - \frac{1}{4\pi} \mathbb{E}\left[ \Psi(Y - X, Y' - X) \right].
\]

**Proof.** Recall that

\[
W_d^2 = \int_{S^{d-1}} \int_{\mathbb{R}} \left( F_{\beta^\top X}(t) - G_{\beta^\top Y}(t) \right)^2 d(F_{\beta^\top X}(t)/2 + G_{\beta^\top Y}(t)/2)d\lambda(\beta)
\]

\[
= \mathbb{E}_{\beta, W_*} \left[ F_{\beta^\top X}(\beta^\top W_* )^2 \right] + \mathbb{E}_{\beta, W_*} \left[ (G_{\beta^\top Y}(\beta^\top W_* )^2 \right] \\
- 2 \mathbb{E}_{\beta, W_*} \left[ F_{\beta^\top X}(\beta^\top W_* )G_{\beta^\top Y}(\beta^\top W_* ) \right],
\]

where

- **I**
  \[
  \mathbb{E}_{\beta, W_*} \left[ F_{\beta^\top X}(\beta^\top W_* ) \right] = \frac{1}{n} \sum_{i=1}^{n} \left( F_{\beta^\top X}(\beta^\top W_* ) \right)
  \]

- **II**
  \[
  \mathbb{E}_{\beta, W_*} \left[ (G_{\beta^\top Y}(\beta^\top W_* )^2 \right] = \frac{1}{n} \sum_{i=1}^{n} \left( G_{\beta^\top Y}(\beta^\top W_* )^2 \right)
  \]

- **III**
  \[
  \mathbb{E}_{\beta, W_*} \left[ F_{\beta^\top X}(\beta^\top W_* )G_{\beta^\top Y}(\beta^\top W_* ) \right] = \frac{1}{n} \sum_{i=1}^{n} \left( F_{\beta^\top X}(\beta^\top W_* )G_{\beta^\top Y}(\beta^\top W_* ) \right)
  \]
Similarly, define the corresponding term $(II)$ has the identity
\begin{align*}
(II) &= \mathbb{E}_{\beta, W_*, Y, Y'} \left[ I(\beta^T Y \leq \beta^T W_*, \beta^T Y' \leq \beta^T W_*) \right] \\
&= \frac{1}{2} \mathbb{E}_{\beta, X, X', Y} \left[ I(\beta^T X \leq \beta^T Y, \beta^T X' \leq \beta^T Y) \right] \\
&+ \frac{1}{2} \mathbb{E}_{\beta, X', Y} \left[ I(\beta^T Y \leq \beta^T X, \beta^T Y' \leq \beta^T X) \right].
\end{align*}
and
\begin{align*}
(III) &= \mathbb{E}_{\beta, X, Y, W_*} \left[ I(\beta^T X \leq \beta^T W_*, \beta^T Y \leq \beta^T W_*) \right] \\
&= \frac{1}{2} \mathbb{E}_{\beta, X, X', Y} \left[ I(\beta^T X \leq \beta^T Y, \beta^T Y' \leq \beta^T Y) \right] \\
&+ \frac{1}{2} \mathbb{E}_{\beta, X, Y, Y'} \left[ I(\beta^T Y \leq \beta^T X, \beta^T Y' \leq \beta^T X) \right].
\end{align*}
We then apply Lemma 2.2 to obtain the desired result.

Lemma B.2 (Another expression for the CvM statistic). Consider the kernel of order three:
\begin{align*}
h^*(x_1, x_2, x_3; y_1, y_2, y_3) &= \frac{1}{2} \mathbb{E}_\beta \left\{ I(\beta^T x_1 \leq \beta^T x_3) - I(\beta^T y_1 \leq \beta^T x_3) \right\} \cdot \left\{ I(\beta^T x_2 \leq \beta^T x_3) - I(\beta^T y_2 \leq \beta^T x_3) \right\} \\
&+ \frac{1}{2} \mathbb{E}_\beta \left\{ I(\beta^T x_1 \leq \beta^T y_3) - I(\beta^T y_1 \leq \beta^T y_3) \right\} \cdot \left\{ I(\beta^T x_2 \leq \beta^T y_3) - I(\beta^T y_2 \leq \beta^T y_3) \right\}
\end{align*}
and define the corresponding $U$-statistic by
\begin{align*}
U^*_{\text{CVM}} := \frac{1}{(m)_3(n)_3} \sum_{i_1, i_2, i_3} \sum_{j_1, j_2, j_3} h^* (X_{i_1}, X_{i_2}, X_{i_3}; Y_{j_1}, Y_{j_2}, Y_{j_3}).
\end{align*}
Then $U^*_{\text{CVM}}$ is an unbiased estimator of $W^2_d$. Furthermore when $\beta^T X$ and $\beta^T Y$ are continuous for all $\beta \in \mathbb{S}^{d-1}$ almost surely, it becomes
\begin{align*}
U^*_{\text{CVM}} = U_{\text{CVM}} &= \frac{1}{(m)_2(n)_2} \sum_{i_1, i_2=1}^m \sum_{j_1, j_2=1}^n \sum_{i_1 \neq i_2} \sum_{j_1 \neq j_2} h(X_{i_1}, X_{i_2}; Y_{j_1}, Y_{j_2}) \text{ almost surely.}
\end{align*}
Proof. The unbiasedness property is trivial. We will show that (33) holds when $\beta^TX$ and $\beta^TY$ are continuous for all $\beta \in \mathbb{S}^{d-1}$ almost surely. Since there is no tie with probability one, we have

$$
\frac{1}{(m)_3} \sum_{i_1, i_2, i_3 \text{ all distinct}}^m \beta[I(\beta^TX_{i_1} \leq \beta^TX_{i_3})I(\beta^TX_{i_2} \leq \beta^TX_{i_3})] = \frac{1}{3} \quad \text{and}
$$

$$
\frac{1}{(n)_3} \sum_{j_1, j_2, j_3 \text{ all distinct}}^n \beta[I(\beta^TY_{j_1} \leq \beta^TY_{j_3})I(\beta^TY_{j_2} \leq \beta^TY_{j_3})] = \frac{1}{3}.
$$

Also the following identities are true:

$$
\frac{2}{(m)_2 \cdot n} \sum_{i_1, i_2=1}^m \sum_{j=1}^n \beta[I(\beta^TX_{i_1} \leq \beta^TX_{i_2})I(\beta^TY_{j} \leq \beta^TX_{i_2})]
$$

$$
= 1 - \frac{1}{(m)_2 \cdot n} \sum_{i_1, i_2=1}^m \sum_{j=1}^n \beta[I(\beta^TX_{i_1} \leq \beta^TY_{j})I(\beta^TY_{j} \leq \beta^TX_{i_2})]
$$

and

$$
\frac{2}{m \cdot (n)_2} \sum_{i=1}^m \sum_{j_1, j_2=1, j_1 \neq j_2}^n \beta[I(\beta^TY_{j_1} \leq \beta^TY_{j_2})I(\beta^TX_{i} \leq \beta^TY_{j_2})]
$$

$$
= 1 - \frac{1}{m \cdot (n)_2} \sum_{i=1}^m \sum_{j_1, j_2=1, j_1 \neq j_2}^n \beta[I(\beta^TY_{j_1} \leq \beta^TX_{i})I(\beta^TY_{j_2} \leq \beta^TX_{i})].
$$

After expanding the terms in $h^*$ and replacing the above identities, we obtain

$$
U_{\text{CVM}}^* = \frac{1}{(m)_2 \cdot n} \sum_{i_1, i_2=1}^m \sum_{j=1}^n \beta[I(\beta^TX_{i_1} \leq \beta^TY_{j})I(\beta^TX_{i_2} \leq \beta^TY_{j})]
$$

$$
+ \frac{1}{m \cdot (n)_2} \sum_{i=1}^m \sum_{j_1, j_2=1, j_1 \neq j_2}^n \beta[I(\beta^TY_{j_1} \leq \beta^TX_{i})I(\beta^TY_{j_2} \leq \beta^TX_{i})] - \frac{2}{3},
$$

$$
= \frac{1}{(m)_2(n)_2} \sum_{i_1, i_2=1, j_1, j_2=1, j_1 \neq j_2}^m \sum_{j_1, j_2=1, j_1 \neq j_2}^n h(X_{i_1}, X_{i_2}, Y_{j_1}, Y_{j_2}).
$$

Hence the result follows. \(\square\)

Next, we present the exact variance of the two-sample $U$-statistic, which will be used to bound the variance of the proposed statistic.
Lemma B.3 (Theorem 2 of Lee (1990) in Chapter 2). Let $U_{m,n}$ be a two-sample $U$-statistic based on a kernel having degrees $k_1$ and $k_2$. Then

$$\text{Var} (U_{m,n}) = \sum_{c=0}^{k_1} \sum_{d=0}^{k_2} \binom{k_1}{c} \binom{k_2}{d} \frac{(m-k_1-c)!}{k_1!} \frac{(n_2-k_2-d)!}{k_2!} \sigma_{c,d}^2,$$

where $\sigma_{c,d}^2$ is defined in (29).

Hoeffding (1952) established a sufficient condition (indeed the necessary condition proved by Chung and Romano, 2013) under which the permutation distribution approximates the corresponding unconditional distribution. The condition is stated as follows:

Lemma B.4 (Theorem 5.1 of Chung and Romano (2013)). Suppose $X_n$ has distribution $P_n$ in $X_n$, and $G_n$ is a finite group of transformations from $X_n$ to $X_n$. Let $G_n$ be a random variable that is uniform on $G_n$. Also, let $G'_n$ have the same distribution as $G_n$, with $X_n$, $G_n$ and $G'_n$ mutually independent. Suppose, under $P_n$,

$$(T_n(G_nX_n), T_n(G'_nX_n)) \overset{d}{\rightarrow} (T, T'), \quad (34)$$

where $T$ and $T'$ are independent, each with common c.d.f. $R(\cdot)$. Then, under $P_n$,

$$\hat{R}_n(t) \overset{p}{\rightarrow} R(t), \quad (35)$$

for every $t$ which is a continuity point of $R(\cdot)$. Conversely, if (35) holds for some limiting c.d.f. $R(\cdot)$ whenever $t$ is a continuity point, then (34) holds.

C Proofs

C.1 Proof of Lemma 2.1

From the definition of $W^2_d$, it is clear to see that $W^2_d \geq 0$ and it becomes zero if $F_X = G_Y$. For the other direction, we will show that if $W^2_d = 0$, then $X$ and $Y$ have the same characteristic function:

$$\mathbb{E}_X \left[ e^{it^\top \beta} X \right] = \mathbb{E}_Y \left[ e^{it^\top \beta} Y \right] \quad \text{for all } (\beta, t) \in \mathbb{S}^{d-1} \times \mathbb{R},$$

which implies $F_X = G_Y$.

1. Univariate case

We begin with the univariate case. Suppose that $X$ and $Y$ are univariate random variables. We claim that the univariate Cvm distance has the characteristic property:

$$W^2 = \int_{\mathbb{R}} (F_X(t) - G_Y(t))^2 \frac{d(F_X(t)/2 + G_Y(t)/2)}{dF_X} = 0,$$

if and only if $F_X(t) = G_Y(t)$ for all $t \in \mathbb{R}$. Define $E$ to be the smallest set $E \subseteq \mathbb{R}$ such that

$$\int_E dF_X = \int_E dG_Y = 1.$$
Since $\mathbb{P}(X \in E^c) = \mathbb{P}(Y \in E^c) = 0$, we only need to show that $W^2 = 0$ implies $F_X(t) = G_Y(t)$ for all $t \in E$. Suppose there exists $t_0 \in E$ such that $F_X(t_0) \neq G_Y(t_0)$. Suppose further that either $F_X$ or $G_Y$ is discontinuous at $t_0$. Then, there is a nonzero probability mass at $t_0$ and thus $W^2 > 0$. This contradicts the assumption. Next, suppose that both $F_X$ and $G_Y$ are continuous at $t_0$. Then $W^2 > 0$ follows similarly from Lemma 4.1 of Lehmann (1951). Therefore, we conclude that $W^2 = 0$ implies that $X$ and $Y$ have the same distribution.

2. Multivariate case
Recall that $\lambda(\cdot)$ is the uniform measure on $S^{d-1}$. From the characteristic property of the univariate CvM distance, $W^2 = 0$ implies that $\beta^\top X$ and $\beta^\top Y$ are identically distributed for $\lambda$-almost all $\beta \in S^{d-1}$. Now, by the continuity of the characteristic function, we conclude that

$$
\mathbb{E}_X \left[ e^{it\beta^\top X} \right] = \mathbb{E}_Y \left[ e^{it\beta^\top Y} \right] \quad \text{for all } (\beta, t) \in S^{d-1} \times \mathbb{R}.
$$

C.2 Proof of Theorem 2.1
The key ingredient of the proof is Lemma 2.2 originally provided in Escanciano (2006). In the original paper, the integration was with respect to $d\beta$ rather than $d\lambda(\beta)$:

$$
\int_{S^{d-1}} I(\beta^\top U_1 \leq 0)I(\beta^\top U_2 \leq 0)d\beta = c(d)\left\{ \pi - \arccos \left( \frac{U_1^\top U_2}{||U_1|| ||U_2||} \right) \right\},
$$

where $c(d) = \pi^{d/2-1}/\Gamma(d/2)$ as corrected in Zhu et al. (2017). Our expression in Lemma 2.2 follows by the fact that the area of the surface of the $d$-dimensional unit ball is $2\pi^{d/2}/\Gamma(d/2)$.

We assume that $\beta^\top X$ and $\beta^\top Y$ are continuous for all $\beta \in S^{d-1}$ almost surely. This implies that there will be no tie between $\beta^\top X, \beta^\top X'$ and $\beta^\top X''$ where $X, X', X'' \overset{i.i.d.}{\sim} F_X$ for $\lambda$-almost all $\beta \in S^{d-1}$ with probability one. This is also true for $Y, Y', Y'' \overset{i.i.d.}{\sim} G_Y$. Therefore, the following identities hold for all $\beta \in S^{d-1}$ almost surely:

\begin{align*}
\int (F_{\beta^\top X}(t))^2 dF_{\beta^\top X}(t) &= \mathbb{P}\left( \max\{\beta^\top X, \beta^\top X'\} \leq \beta^\top X'' \right) = \frac{1}{3}, \\
\int (G_{\beta^\top Y}(t))^2 dG_{\beta^\top Y}(t) &= \mathbb{P}\left( \max\{\beta^\top Y, \beta^\top Y'\} \leq \beta^\top Y'' \right) = \frac{1}{3}, \\
\int (F_{\beta^\top X}(t))^2 dG_{\beta^\top Y}(t) &= \mathbb{P}\left( \max\{\beta^\top X, \beta^\top X'\} \leq \beta^\top Y \right), \\
\int (G_{\beta^\top Y}(t))^2 dF_{\beta^\top X}(t) &= \mathbb{P}\left( \max\{\beta^\top Y, \beta^\top Y'\} \leq \beta^\top X \right). 
\end{align*}

Also note that

$$
\mathbb{P}\left( \max\{\beta^\top X, \beta^\top X'\} \leq \beta^\top Y \right) + \mathbb{P}\left( \max\{\beta^\top X, \beta^\top Y\} \leq \beta^\top X' \right) + \mathbb{P}\left( \max\{\beta^\top X', \beta^\top Y\} \leq \beta^\top X \right) = 1
$$

and

$$
\mathbb{P}\left( \max\{\beta^\top X, \beta^\top Y\} \leq \beta^\top X' \right) = \mathbb{P}\left( \max\{\beta^\top X', \beta^\top Y\} \leq \beta^\top X \right).
$$
These two identities give
\[ \int F_{\beta^\top X(t)}G_{\beta^\top Y(t)}(t)\,dF_{\beta^\top X}(t) = \mathbb{P}\left( \max\{\beta^\top X, \beta^\top Y\} \leq \beta^\top X' \right) \]
\[ = \frac{1}{2} - \frac{1}{2}\mathbb{P}\left( \max\{\beta^\top X, \beta^\top X'\} \leq \beta^\top Y' \right). \tag{37} \]

Similarly,
\[ \int F_{\beta^\top X(t)}G_{\beta^\top Y(t)}(t)\,dG_{\beta^\top Y}(t) = \mathbb{P}\left( \max\{\beta^\top Y, \beta^\top X\} \leq \beta^\top Y' \right) \]
\[ = \frac{1}{2} - \frac{1}{2}\mathbb{P}\left( \max\{\beta^\top Y, \beta^\top Y'\} \leq \beta^\top X \right). \tag{38} \]

Now, combine (36), (37) and (38) to obtain
\[ \int_{S^d} \int_{\mathbb{R}} (F_{\beta^\top X}(t) - G_{\beta^\top Y}(t))^2 \,d(F_{\beta^\top X}(t)/2 + G_{\beta^\top Y}(t)/2)\,d\lambda(\beta) \]
\[ = \int_{S^d} \mathbb{P}\left( \max\{\beta^\top X, \beta^\top X'\} \leq \beta^\top Y \right)\,d\lambda(\beta) + \int_{S^d} \mathbb{P}\left( \max\{\beta^\top Y, \beta^\top Y'\} \leq \beta^\top X \right)\,d\lambda(\beta) - \frac{2}{3}. \]

Hence,
\[ W_d^2 = \mathbb{E}_{\beta, X, X', Y} \left[ I(\beta^\top X \leq \beta^\top Y, \beta^\top X' \leq \beta^\top Y) \right] \]
\[ + \mathbb{E}_{\beta, Y, Y', X} \left[ I(\beta^\top Y \leq \beta^\top X, \beta^\top Y' \leq \beta^\top X) \right] - \frac{2}{3}. \]

Then apply Lemma 2.2 to obtain the result.

### C.3 Proof of Theorem 2.2

We first show that \( h \) is degenerate under \( H_0 \). Then apply the limit theorem for \( U \)-statistics.

1. **Degeneracy**

   Recall the definition of the kernel \( h \):
   \[ h(x_1, x_2; y_1, y_2) = \frac{1}{3} - \frac{1}{2\pi} \Psi(x_1 - y_1, x_2 - y_1) - \frac{1}{2\pi} \Psi(y_1 - x_1, y_2 - x_1), \]
   and denote the symmetrized kernel \( h \) by \( \tilde{h} \), i.e.
   \[ \tilde{h}(x_1, x_2; y_1, y_2) = \frac{1}{2} h(x_1, x_2; y_1, y_2) + \frac{1}{2} h(x_2, x_1; y_2, y_1). \]
   
   To begin, let us focus on univariate cases where \( x_1, x_2, y_1, y_2 \in \mathbb{R} \). Let \( h^{(1)} \) denote the symmetrized \( h \) for the univariate case, that is
   \[ h^{(1)}(x_1, x_2; y_1, y_2) := \frac{1}{2} \left( I(\max\{x_1, x_2\} \leq y_1) + I(\max\{x_1, x_2\} \leq y_2) \right. \]
   \[ + I(\max\{y_1, y_2\} \leq x_1) + I(\max\{y_1, y_2\} \leq x_2) \left. \right) - \frac{2}{3}. \]
From the identity
\[
I(\max\{x_1, x_2\} \leq y_1) + I(\max\{x_1, x_2\} \leq y_2) + I(\max\{y_1, y_2\} \leq x_1) + I(\max\{y_1, y_2\} \leq x_2) - 1
\]
\[= I(\max\{x_1, x_2\} \leq \min\{y_1, y_2\}) + I(\max\{y_1, y_2\} \leq \min\{x_1, x_2\}),
\]
the symmetric kernel \( h^{(1)} \) becomes
\[
2h^{(1)}(x_1, x_2; y_1, y_2) = I(\max\{x_1, x_2\} \leq \min\{y_1, y_2\})
\]
\[+ I(\max\{y_1, y_2\} \leq \min\{x_1, x_2\}) - \frac{1}{3},
\]
which is equivalent to the kernel for Lehmann’s two-sample statistic (Lehmann, 1951). In addition, when \( F_X = G_Y \), we have
\[
h_{1,0}^{(1)}(x_1) := \mathbb{E} \left[ h^{(1)}(x_1, X_2; Y_1, Y_2) \right] = 0 \quad \text{and}
\]
\[
h_{0,1}^{(1)}(y_1) := \mathbb{E} \left[ h^{(1)}(X_1, X_2; y_1, Y_2) \right] = 0. \tag{39}
\]
See Chapter 4 of Bhat (1995) for details.

Now, come back to multivariate cases where \( x_1, x_2, y_1, y_2 \in \mathbb{R}^d \). By the definition of \( \tilde{h} \),
\[
\tilde{h}(x_1, x_2, y_1, y_2) = \int_{\mathbb{S}^{d-1}} h^{(1)}(\beta^T x_1, \beta^T x_2; \beta^T y_1, \beta^T x_2) d\lambda(\beta).
\]
Due to (39) for any \( \beta \in \mathbb{S}^{d-1} \) almost surely, the Fubini’s theorem presents
\[
\mathbb{E} \left[ h^{(1)}(\beta^T x_1, \beta^T x_2; \beta^T y_1, \beta^T y_2) \right] = \mathbb{E} \left[ h^{(1)}(X_1, X_2; y_1, \beta^T y_2) \right] = 0,
\]
which results in
\[
\tilde{h}_{1,0}(x_1) := \mathbb{E} \left[ \tilde{h}(x_1, X_2; Y_1, Y_2) \right]
\]
\[= \int_{\mathbb{S}^{d-1}} \mathbb{E} \left[ h^{(1)}(\beta^T x_1, \beta^T x_2; \beta^T y_1, \beta^T y_2) \right] d\lambda(\beta) = 0,
\]
\[
\tilde{h}_{0,1}(y_1) := \mathbb{E} \left[ \tilde{h}(X_1, X_2; y_1, Y_2) \right]
\]
\[= \int_{\mathbb{S}^{d-1}} \mathbb{E} \left[ h^{(1)}(\beta^T X_1, \beta^T X_2; \beta^T y_1, \beta^T y_2) \right] d\lambda(\beta) = 0.
\]
On the other hand,
\[
\tilde{h}_{2,0}(x_1, x_2) := \mathbb{E} \left[ \tilde{h}(x_1, x_2; Y_1, Y_2) \right]
\]
\[= \frac{1}{2} \int_{\mathbb{S}^{d-1}} \left( 1 - F_{\beta^T X}(\max\{\beta^T x_1, \beta^T x_2\}) \right)^2 d\lambda(\beta)
\]
\[+ \frac{1}{2} \int_{\mathbb{S}^{d-1}} F_{\beta^T X}^2(\min\{\beta^T x_1, \beta^T x_2\}) d\lambda(\beta) - \frac{1}{6}.
\]
\[ \tilde{h}_{0,2}(y_1, y_2) := \mathbb{E} \left[ \tilde{h}(X_1, X_2; y_1, y_2) \right], \]
\[ = \frac{1}{2} \int_{S^{d-1}} \left( 1 - G_{\beta^T y} \left( \max \{ \beta^T y_1, \beta^T y_2 \} \right) \right)^2 d\lambda(\beta) \]
\[ + \frac{1}{2} \int_{S^{d-1}} G_{\beta^T y} \left( \min \{ \beta^T y_1, \beta^T y_2 \} \right) d\lambda(\beta) - \frac{1}{6}, \]
\[ \tilde{h}_{1,1}(x_1, y_1) := \mathbb{E} \left[ \tilde{h}(x_1, X_2; y_1, Y_2) \right] \]
\[ = -\frac{1}{2} \tilde{h}_{2,0}(x_1, y_1). \]

Note that \( \tilde{h}_{2,0}(x_1, x_2) \neq 0 \) for some \((x_1, x_2)\). For example, when \( x_1 = x_2 \), it is seen that
\[ \frac{1}{2} \left( 1 - F_{\beta^T X}(\beta^T x_1) \right)^2 + \frac{1}{2} F_{\beta^T X}^2(\beta^T x_1) - \frac{1}{6} \geq \frac{1}{12} \]
for all \( \beta \in S^{d-1} \), which implies \( \tilde{h}_{2,0}(x_1, x_1) \geq 1/12 \). By the continuity of \( \tilde{h}_{2,0} \) at \((x_1, x_1)\), there exist a set \( E \) with nonzero measure such that \( \tilde{h}_{2,0}(z_1, z_2) > 0 \) for \((z_1, z_2) \in E\). Therefore, we conclude that \( \tilde{h} \) (and \( h \)) has a degeneracy of order one under \( H_0 \).

### 2. Limiting distribution of the U-statistic

To obtain the limiting null distribution of \( U_{\text{CVM}} \), we apply the result given in Chapter 3 of Bhat (1995) to have
\[ NU_{\text{CVM}} \overset{d}{\to} \frac{1}{\pi_1} \sum_{k=1}^{\infty} \lambda_k (Z_k^2 - 1) + \frac{1}{\pi_2} \sum_{k=1}^{\infty} \lambda_k (Z_k'^2 - 1) - \frac{2}{\sqrt{\pi_1 \pi_2}} \sum_{k=1}^{\infty} \lambda_k Z_k Z_k', \]
where \( Z_k, Z_k' \overset{i.i.d.}{\sim} N(0, 1) \). Next, by Slutsky’s theorem with \( m/N \to \pi_1 \) and \( n/N \to \pi_2 \),
\[ \frac{mn}{m+n} U_{\text{CVM}} \overset{d}{\to} \sum_{k=1}^{\infty} \lambda_k \left[ \left( \sqrt{\pi_2} Z_k - \sqrt{\pi_1} Z_k' \right)^2 - 1 \right] \]
and
\[ \sqrt{\pi_2} Z_k - \sqrt{\pi_1} Z_k' \sim N(0, 1), \]
the result follows.

### C.4 Proof of Theorem 2.3

Under the null hypothesis, we need to verify the conditions given in Theorem A.1. Indeed, these conditions are satisfied with \( r = 2 \) as we showed in the proof of Theorem 2.2. Hence, the result follows under \( H_0 \).

Next, we focus on the alternative hypothesis. We use the coupling argument (Algorithm 1) to show that the difference between the two \( U \)-statistics — one is based on the randomly permuted original samples and the other is based on the corresponding coupled \( i.i.d. \) samples — is asymptotically negligible. Formally, we state the result in the following:
Lemma C.1 (Coupling for the CvM statistic). Consider the two sets of samples \( \{Z_1, \ldots, Z_N\} \) and \( \{\tilde{Z}_{\pi_0(1)}, \ldots, \tilde{Z}_{\pi_0(N)}\} \) from Algorithm 1 and their random permutations \( \{Z_{\pi(1)}, \ldots, Z_{\pi(N)}\} \) and \( \{\tilde{Z}_{\pi(\pi_0(1))}, \ldots, \tilde{Z}_{\pi(\pi_0(N))}\} \). Then we have

\[
NU_{\text{CVM}}(Z_{\pi(1)}, \ldots, Z_{\pi(N)}) - NU_{\text{CVM}}(\tilde{Z}_{\pi(\pi_0(1))}, \ldots, \tilde{Z}_{\pi(\pi_0(N))}) = o_P(1). \tag{40}
\]

Proof. Using the result in Lemma B.2, we work with the three order kernel \( h^*(x_1, x_2, x_3; y_1, y_2, y_3) \). First notice that the expectation of both \( UC_{\text{CVM}}(Z_{\pi(1)}, \ldots, Z_{\pi(N)}) \) and \( UC_{\text{CVM}}(\tilde{Z}_{\pi(\pi_0(1))}, \ldots, \tilde{Z}_{\pi(\pi_0(N))}) \) are zero. For example, the law of total expectation gives

\[
\mathbb{E} \left[ \{I(\beta^T Z_{\pi(1)} \leq \beta^T Z_{\pi(3)}) - I(\beta^T Z_{\pi(m+1)} \leq \beta^T Z_{\pi(3)})\} \times \{I(\beta^T Z_{\pi(2)} \leq \beta^T Z_{\pi(3)}) - I(\beta^T Z_{\pi(m+2)} \leq \beta^T Z_{\pi(3)})\} \right]
\]

\[
= \mathbb{E} \left[ \mathbb{P}(\pi(1), \pi(m+1)) \left\{ \{I(\beta^T Z_{\pi(1)} \leq \beta^T Z_{\pi(3)}) - I(\beta^T Z_{\pi(m+1)} \leq \beta^T Z_{\pi(3)})\} \mid \beta, Z, \pi(2), \pi(3), \pi(m+2) \right\} \right. \\

\times \left. \left\{ I(\beta^T Z_{\pi(2)} \leq \beta^T Z_{\pi(3)}) - I(\beta^T Z_{\pi(m+2)} \leq \beta^T Z_{\pi(3)}) \right\} \right] .
\]

Note that the conditional expectation \((*)\) is zero and thus the entire expectation is zero. By applying the same logic to the other terms, we see that the expectations of \( UC_{\text{CVM}}(Z_{\pi(1)}, \ldots, Z_{\pi(N)}) \) and \( UC_{\text{CVM}}(\tilde{Z}_{\pi(\pi_0(1))}, \ldots, \tilde{Z}_{\pi(\pi_0(N))}) \) are zero.

Now it suffices to show that

\[
\mathbb{E} \left[ (NU_{\text{CVM}}(Z_{\pi(1)}, \ldots, Z_{\pi(N)}) - NU_{\text{CVM}}(\tilde{Z}_{\pi(\pi_0(1))}, \ldots, \tilde{Z}_{\pi(\pi_0(N))}))^2 \right] = o(1) \tag{41}
\]

to establish (40). Denote

\[
d_\pi(i_1, i_2, i_3; j_1, j_2, j_3) = h^*(Z_{\pi(i_1)}, Z_{\pi(i_2)}, Z_{\pi(i_3)}; Z_{\pi(j_1+m)}, Z_{\pi(j_2+m)}, Z_{\pi(j_3+m)})
\]

\[
- h^*(\tilde{Z}_{\pi(\pi_0(i_1))}, \tilde{Z}_{\pi(\pi_0(i_2))}, \tilde{Z}_{\pi(\pi_0(i_3))}; \tilde{Z}_{\pi(\pi_0(j_1+m))}, \tilde{Z}_{\pi(\pi_0(j_2+m))}, \tilde{Z}_{\pi(\pi_0(j_3+m))}).
\]

Then the square of \( NU_{\text{CVM}}(Z_{\pi(1)}, \ldots, Z_{\pi(N)}) - NU_{\text{CVM}}(\tilde{Z}_{\pi(\pi_0(1))}, \ldots, \tilde{Z}_{\pi(\pi_0(N))}) \) can be written as

\[
D_{m,n} = \frac{N^2}{m^2(m-1)^2n^2(n-1)^2} \times
\]

\[
\sum_{i_1 \neq i_2 \neq i_3} \sum_{j_1 \neq j_2 \neq j_3} \sum_{j_1' \neq j_2' \neq j_3'} \sum_{i_1', i_2', i_3', j_1', j_2', j_3' = 1}^{\infty} d_\pi(i_1, i_2, i_3; j_1, j_2, j_3) d_\pi(i_1', i_2', i_3'; j_1', j_2', j_3').
\]

\( (**\) \)

By the law of total expectation, we can see that

\[
\mathbb{E}_{\pi} \left[ d_\pi(i_1, i_2, i_3; j_1, j_2, j_3) d_\pi(i_1', i_2', i_3'; j_1', j_2', j_3') \mid \beta, Z, \tilde{Z} \right] = 0
\]

when

- Case 1(a). \( \{i_1, i_2, i_3\} \cap \{i_1', i_2', i_3'\} = 0 \) and \( \{j_1, j_2, j_3\} \cap \{j_1', j_2', j_3'\} = 0 \) or
- Case 1(b). \( \{i_1, i_2, i_3\} \cap \{i_1', i_2', i_3'\} = 1 \) and \( \{j_1, j_2, j_3\} \cap \{j_1', j_2', j_3'\} = 0 \) or
- Case 1(c). \( \{i_1, i_2, i_3\} \cap \{i_1', i_2', i_3'\} = 0 \) and \( \{j_1, j_2, j_3\} \cap \{j_1', j_2', j_3'\} = 1 \).
Thus the unconditional expectation is also zero in these cases. Next consider the cases where

- Case 2(a). \(|\{i_1, i_2, i_3\} \cap \{i'_1, i'_2, i'_3\}| = 2\) and \(|\{j_1, j_2, j_3\} \cap \{j'_1, j'_2, j'_3\}| = 0\).
- Case 2(b). \(|\{i_1, i_2, i_3\} \cap \{i'_1, i'_2, i'_3\}| = 0\) and \(|\{j_1, j_2, j_3\} \cap \{j'_1, j'_2, j'_3\}| = 2\).
- Case 2(c). \(|\{i_1, i_2, i_3\} \cap \{i'_1, i'_2, i'_3\}| = 1\) and \(|\{j_1, j_2, j_3\} \cap \{j'_1, j'_2, j'_3\}| = 1\).

Suppose there are \(B_1\) number of different observations between \(\{Z_{\pi(1)}, \ldots, Z_{\pi(m)}\}\) and \(\{Z_{\pi(\pi_0(1))}, \ldots, Z_{\pi(\pi_0(m))}\}\) and \(B_2\) number of different observations between \(\{Z_{\pi(m+1)}, \ldots, Z_{\pi(m+n)}\}\) and \(\{Z_{\pi(\pi_0(m+1))), \ldots, Z_{\pi(\pi_0(m+n))}\}\). Hence, \(D = B_1 + B_2\). Then for Case 2(a), there are at least \(3C_2 \times 3! \times (m - B_1)(m - B_1 - 1)(m - B_1 - 2)(m - B_1 - 3)(m - B_1 - 4)\) numbers of zero product terms in (**) out of total number of product terms.

Similarly for Case 2(b), there are at least \(3C_2 \times 3! \times (m - B_1)(m - B_1 - 1)(m - B_1 - 2)(m - B_1 - 3)(m - B_1 - 4)\) numbers of zero product terms in (**) out of total number of product terms.

For Case 2(c), there are at least \(3C_2 \times 3! \times (m - B_1)(m - B_1 - 1)(m - B_1 - 2)(m - B_1 - 3)(m - B_1 - 4)\) numbers of zero product terms in (**) out of total number of product terms.

To summarize it, the number of non-zero product terms for each case is bounded by

\[
\text{Case 2(a)} \lesssim B_1 m^3 n^6 + B_2 m^4 n^5,
\]
\[
\text{Case 2(b)} \lesssim B_1 m^5 n^4 + B_2 m^6 n^3,
\]
\[
\text{Case 2(c)} \lesssim B_1 m^4 n^5 + B_2 m^5 n^4.
\]

Note that the number of the rest cases except for Case 1(a), 1(b), 1(c), 2(a), 2(b) and 2(c) are order of \(O(N^9)\). Since \(E[B_1] = O(\sqrt{N})\), \(E[B_2] = O(\sqrt{N})\) and the kernel \(d_{\pi}\) is bounded, we can conclude that

\[
E[D_{m,n}] = O\left(\frac{1}{\sqrt{N}}\right) = o(1).
\]

Hence, the result in (40) follows, which completes the proof. \(\square\)

From Lemma C.1, we have established that \(NU_{cvm}(Z_{\pi(1)}, \ldots, Z_{\pi(N)})\) and \(NU_{cvm}(\tilde{Z}_{\pi(\pi_0(1))}, \ldots, \tilde{Z}_{\pi(\pi_0(N))})\) have the same limiting distribution. Note that \(\tilde{Z}_{\pi(\pi_0(1))}, \ldots, \tilde{Z}_{\pi(\pi_0(N))}\) are sampled from \(\frac{n}{N} F_X + \frac{n}{N} G_Y\). Next, we will further show that the limiting distribution of \(NU_{cvm}\) based on samples from \(\frac{n}{N} F_X + \frac{n}{N} G_Y\) and that based on samples from \(\pi_1 F_X + \pi_2 G_Y\) are equivalent when \(\frac{n}{N} \rightarrow \pi_1\) and \(\frac{n}{N} \rightarrow \pi_2\) as \(N \rightarrow \infty\) where \(0 < \pi_1, \pi_2 < 1\). Since the limiting distribution of \(NU_{cvm}\) is the weighted sum of independent chi-square statistics, the limiting distribution is decided by the weights, which are eigenvalues of the integral equation associated with the kernel. Using the symmetrized kernel of \(h\) denoted by \(\tilde{h}\), write

\[
\tilde{h}^{(m,n)}_{2,0}(x_1, x_2) = \int \tilde{h}(x_1, x_2; y_1, y_2) dH_{m,n}(y_1) dH_{m,n}(y_2)
\]
where \( H_{m,n} = \frac{m}{N} F_X + \frac{n}{N} G_Y \). Similarly, write
\[
\tilde{h}_{2,0}(x_1, x_2) = \int \tilde{h}(x_1, x_2; y_1, y_2) dH(y_1) dH(y_2)
\]
where \( H = \pi_1 F_X + \pi_1 G_Y \). We can see that

\[
|\tilde{h}^{(m,n)}_{2,0}(x_1, x_2) - \tilde{h}_{2,0}(x_1, x_2)| \leq \sum_{i=0, j=0}^{4} \left| \left( \frac{m}{N} \right)^i \left( \frac{n}{N} \right)^j - \pi_1^i \pi_2^j \right|
\]  

(42)

by the boundness of the kernel, i.e. \(|\tilde{h}| \leq 1\). Let \( \{\lambda_i^{(m,n)}\}_{i=1}^{\infty} \) and \( \{\phi_i^{(m,n)}(\cdot)\}_{i=1}^{\infty} \) be eigenvalues and square integrable eigenfunctions of the following integral equation:

\[
\int \tilde{h}^{(m,n)}_{2,0}(x_1, x_2) \phi_i^{(m,n)}(x_2) dH_{m,n}(x_2) = \lambda_i^{(m,n)} \phi_i^{(m,n)}(x_1).
\]  

(43)

Let us denote the limits by \( \lambda_i^* = \lim_{N \to \infty} \lambda_i^{(m,n)}\) and \( \phi_i^*(z) = \lim_{N \to \infty} \phi_i^{(m,n)}(z) \). We will show that \( \lambda_i^* \) and \( \phi_i^* \) satisfy the following integral equation

\[
\int \tilde{h}_{2,0}(x_1, x_2) \phi_i^*(x_2) dH(x_2) = \lambda_i^* \phi_i^*(x_1)
\]  

(44)

for all \( x_1 \). Therefore, the limits are the eigenvalues and the eigenfunctions of (44). We formally state this in the following lemma.

**Lemma C.2.** Let us denote the eigenvalues and the eigenfunctions of the integral equation in (43) by \( \{\lambda_i^{(m,n)}\}_{i=1}^{\infty} \) and \( \{\phi_i^{(m,n)}(\cdot)\}_{i=1}^{\infty} \), respectively. Further denote their limits by \( \lambda_i^* = \lim_{N \to \infty} \lambda_i^{(m,n)}\) and \( \phi_i^*(z) = \lim_{N \to \infty} \phi_i^{(m,n)}(z) \). Then \( \{\lambda_i^*\}_{i=1}^{\infty} \) and \( \{\phi_i^*(\cdot)\}_{i=1}^{\infty} \) are the eigenvalues and the eigenfunctions of the integral equation in (44). In addition, we have

\[
\sum_{i=1}^{\infty} \left( \lambda_i^{(m,n)} \right)^2 \to \sum_{i=1}^{\infty} \lambda_i^2 \quad \text{as } N \to \infty.
\]

**Proof.** Note that

\[
\left| \int \tilde{h}^{(m,n)}_{2,0}(x_1, x_2) \phi_i^{(m,n)}(x_2) dH_{m,n}(x_2) - \int \tilde{h}_{2,0}(x_1, x_2) \phi_i^{(m,n)}(x_2) dH(x_2) \right|
\]

\[
\leq \left| \int \tilde{h}^{(m,n)}_{2,0}(x_1, x_2) \phi_i^{(m,n)}(x_2) dH_{m,n}(x_2) - \int \tilde{h}^{(m,n)}_{2,0}(x_1, x_2) \phi_i^{(m,n)}(x_2) dH(x_2) \right|
\]

\[
+ \left| \int \tilde{h}^{(m,n)}_{2,0}(x_1, x_2) \phi_i^{(m,n)}(x_2) dH(x_2) - \int \tilde{h}_{2,0}(x_1, x_2) \phi_i^{(m,n)}(x_2) dH(x_2) \right|
\]

\[
= (I) + (II).
\]

For \( I \), we have

\[
(I) = \left( \frac{m}{N} - \pi_1 \right) \int \tilde{h}^{(m,n)}_{2,0}(x_1, x_2) \phi_i^{(m,n)}(x_2) dF_X(x_2)
\]

37
\[\begin{align*}
+ \left( \frac{n}{N} - \pi_2 \right) & \int \tilde{h}_{2,0}^{(m,n)}(x_1, x_2) \phi_i^{(m,n)}(x_2) dG_Y(x_2) \\
\leq & \left| \frac{m}{N} - \pi_1 \right| \int \tilde{h}_{2,0}^{(m,n)}(x_1, x_2) \phi_i^{(m,n)}(x_2) dF_X(x_2) \\
+ & \left| \frac{n}{N} - \pi_2 \right| \int \tilde{h}_{2,0}^{(m,n)}(x_1, x_2) \phi_i^{(m,n)}(x_2) dG_Y(x_2) \\
\leq & \left| \frac{m}{N} - \pi_1 \right| \sqrt{\int \left( \phi_i^{(m,n)}(x_2) \right)^2 dF_X(x_2)} + \left| \frac{n}{N} - \pi_2 \right| \sqrt{\int \left( \phi_i^{(m,n)}(x_2) \right)^2 dG_Y(x_2)}
\end{align*}\]

where the last inequality is due to Cauchy-Schwartz inequality and the boundedness of the kernel. Since \(\phi_i^{(m,n)}\) is an orthonormal eigenfunction, i.e.

\[\int \left( \phi_i^{(m,n)}(x_2) \right)^2 dH_{m,n}(x_2) = \frac{m}{N} \int \left( \phi_i^{(m,n)}(x_2) \right)^2 dF_X(x_2) + \frac{n}{N} \int \left( \phi_i^{(m,n)}(x_2) \right)^2 dG_Y(x_2) = 1,\]

we obtain the following upper bound:

\[\int \left( \phi_i^{(m,n)}(x_2) \right)^2 dF_X(x_2) + \int \left( \phi_i^{(m,n)}(x_2) \right)^2 dG_Y(x_2) \leq \frac{N}{\min\{m, n\}}. \tag{45}\]

Using this, (I) is further bounded by

\[\begin{align*}
(I) & \leq \sqrt{\frac{N}{\min\{m, n\}}} \left( \left| \frac{m}{N} - \pi_1 \right| + \left| \frac{n}{N} - \pi_2 \right| \right). \\
\end{align*}\]

Next, focusing on (II), we have

\[\begin{align*}
(II) & \leq \int \left| \tilde{h}_{2,0}^{(m,n)}(x_1, x_2) - \tilde{h}_{2,0}(x_1, x_2) \right| \phi_i^{(m,n)}(x_2) dH(x_2) \\
& \leq \sum_{i=0}^{4} \sum_{j=0}^{4} \left| \left( \frac{m}{N} \right)^i \left( \frac{n}{N} \right)^j - \pi_1^i \pi_2^j \right| \sqrt{\max(\pi_1, \pi_2)} \frac{N}{\min\{m, n\}}.
\end{align*}\]

Since the upper bounds are uniform over \(x_1\) and \(m/N \to \pi_1, n/N \to \pi_2\) as \(N \to \infty\),

\[\lim_{N \to \infty} \sup_{x_1 \in \mathbb{R}^d} \left| \int \tilde{h}_{2,0}^{(m,n)}(x_1, x_2) \phi_i^{(m,n)}(x_2) dH_{m,n}(x_2) - \int \tilde{h}_{2,0}(x_1, x_2) \phi_i^{(m,n)}(x_2) dH(x_2) \right| = 0.\]

In addition,

\[\lim_{N \to \infty} \sup_{x_1 \in \mathbb{R}^d} \left| \int \tilde{h}_{2,0}^{(m,n)}(x_1, x_2) \phi_i^{(m,n)}(x_2) dH_{m,n}(x_2) - \int \tilde{h}_{2,0}(x_1, x_2) \phi_i^{(m,n)}(x_2) dH(x_2) \right|,\]

\[\geq \sup_{x_1 \in \mathbb{R}^d} \lim_{N \to \infty} \left| \int \tilde{h}_{2,0}^{(m,n)}(x_1, x_2) \phi_i^{(m,n)}(x_2) dH_{m,n}(x_2) - \int \tilde{h}_{2,0}(x_1, x_2) \phi_i^{(m,n)}(x_2) dH(x_2) \right|,\]

38
and thus conclude that $NU^{(1)}_{\text{CVM}}$ is the CvM statistic based on i.i.d. samples from $\frac{m}{N}F_X + \frac{n}{N}G_Y$. Similarly, let $NU^{(2)}_{\text{CVM}}$ be the CvM statistic based on i.i.d. samples from $\pi_1 F_X + \pi_2 G_Y$. Then $NU^{(1)}_{\text{CVM}}$ and $NU^{(2)}_{\text{CVM}}$ have the same limiting distribution.
Proof. The proof can be complete by following the similar steps in Section C.17. Based on i.i.d. samples \( \{Z_1, \ldots, Z_{m+n}\} \sim \frac{m}{N} F_X + \frac{n}{N} G_Y \), we can arrive at

\[
N \hat{U}^{(1)}_{m,n,K} = \sum_{k=1}^{K} \lambda_k^{(m,n)} \left( \sqrt{\frac{N}{m}} \sum_{i=1}^{m} \phi_k^{(m,n)}(Z_i) - \sqrt{\frac{N}{n}} \sum_{i=m+1}^{m+n} \phi_k^{(m,n)}(Z_i) \right)^2 - \frac{1}{\pi_1\pi_2} \sum_{k=1}^{K} \lambda_k + o_P(1).
\]

By the multivariate central limit theorem and Slutsky’s theorem with \( \lambda_k^{(m,n)} \rightarrow \lambda_i \), \( i = 1, \ldots, K \) and \( m/N \rightarrow \pi_1, n/N \rightarrow \pi_2 \), it is seen that

\[
N \hat{U}^{(1)}_{m,n,K} \overset{d}{\rightarrow} \frac{1}{\pi_1\pi_2} \sum_{k=1}^{K} \lambda_k (W_k^2 - 1),
\]

where \( W_k^2 \) are independent chi-square random variables with one degree of freedom. The remainder terms \( k = K + 1, \ldots \) can be similarly controlled by noting that

\[
\lim_{N \to \infty} \sum_{k=K+1}^{\infty} \left( \lambda_k^{(m,n)} \right)^2 = \sum_{k=K+1}^{\infty} \lambda_k^2
\]

from Lemma C.2. This concludes that \( NU_{CvM}^{(1)} \) has the same limiting distribution as \( NU_{CvM}^{(2)} \).

\[ \square \]

C.5 Proof of Theorem 3.1

The boundedness property of the multivariate CvM statistic guarantees that the variance of the \( U \)-statistic is bounded regardless of the presence of contamination. On the other hand, the energy statistic can be adversely affected by outliers, which lead to an extremely high variance of the test statistic. In fact, as the same size tends to infinity, the energy statistic can be completely dominated by these outliers and the resulting test becomes powerless in the end. We will make this statement more rigorous below.

1. Multivariate CvM statistic

We follow the similar steps used in the proof of Theorem 4.2. Recall that \( W_d(P_m, Q_n) \geq \delta_1 \) and thus \( E[U_{CvM}] \geq \delta_1^2 \). We first upper bound the type II error as

\[
\mathbb{P} \left( U_{CvM} < c_{\alpha,m,n}^* \right) = \mathbb{P} \left( U_{CvM} < c_{\alpha,m,n}^*, c_{\alpha,m,n}^* > \delta_1^2 / 2 \right) + \mathbb{P} \left( U_{CvM} < c_{\alpha,m,n}^*, c_{\alpha,m,n}^* \leq \delta_1^2 / 2 \right) \leq \mathbb{P} \left( c_{\alpha,m,n}^* > \delta_1^2 / 2 \right) + \mathbb{P} \left( U_{CvM} \leq \delta_1^2 / 2 \right).
\]

For the first part, Lemma C.7 and Chebyshev’s inequality present

\[
\mathbb{P} \left( U_{CvM} \geq t \right) \leq \frac{\text{Var}_{\pi}(U_{CvM})}{t^2} \leq \frac{C_0}{t^2} \cdot \left( \frac{1}{m} + \frac{1}{n} \right)^2
\]

where \( C_0 \) is a universal constant. This conclude that the critical value of the permutation test is uniformly bounded by

\[
c_{\alpha,m,n}^* \leq \sqrt{\frac{C_0}{\alpha} \left( \frac{1}{m} + \frac{1}{n} \right)}.
\]
Hence, we bound \((I)\) by
\[
(I) = P_1(c_{\alpha,m,n} > \delta_1^2/2) \leq \frac{4}{\delta_1^2} \mathbb{E}[c_{\alpha,m,n}^2] \leq \frac{4C_0}{\alpha \delta_1^4} \left( \frac{1}{m} + \frac{1}{n} \right)^2.
\]
Next,
\[
(II) = P_1(U_{\text{CvM}} \leq \delta_1^2/2) = P_1\left( \frac{U_{\text{CvM}} - \mathbb{E}[U_{\text{CvM}}]}{\sqrt{\text{Var}(U_{\text{CvM}})}} \leq \frac{\delta_1^2/2 - \mathbb{E}[U_{\text{CvM}}]}{\sqrt{\text{Var}(U_{\text{CvM}})}} \right)
\]
\[
\leq P_1\left( \frac{U_{\text{CvM}} - \mathbb{E}[U_{\text{CvM}}]}{\sqrt{\text{Var}(U_{\text{CvM}})}} \leq -\frac{\delta_1^2/2}{\sqrt{\text{Var}(U_{\text{CvM}})}} \right)
\]
\[
= P_1\left( \frac{-U_{\text{CvM}} + \mathbb{E}[U_{\text{CvM}}]}{\sqrt{\text{Var}(U_{\text{CvM}})}} \geq \frac{\delta_1^2/2}{\sqrt{\text{Var}(U_{\text{CvM}})}} \right)
\]
\[
\leq \frac{2\text{Var}(U_{\text{CvM}})}{\delta_1^2}
\]
\[
\leq \frac{C_1}{\delta_1^2} \left( \frac{1}{m} + \frac{1}{n} \right)
\]
where \((i)\) uses \(\mathbb{E}[U_{\text{CvM}}] \geq \delta_1^2\), \((ii)\) is by Chebyshev’s inequality and \((iii)\) uses Lemma C.6 with a universal constant \(C_1\). In the end, we have
\[
\lim_{m,n \to \infty} \inf_{H_m,H_n'} \mathbb{E}_1[\phi_{\text{CvM}}] \geq 1 - \lim_{m,n \to \infty} \inf_{H_m,H_n'} \left\{ \frac{4C_0}{\alpha \delta_1^4} \left( \frac{1}{m} + \frac{1}{n} \right)^2 + \frac{C_1}{\delta_1^2} \left( \frac{1}{m} + \frac{1}{n} \right) \right\}
\]
\[
= 1.
\]

2. Energy statistic
For the energy statistic, we will pick \(F_X, G_Y, H_m, H_n'\) properly to show that the energy statistic is dominated by samples from the contaminations \(H_m\) and \(H_n'\). When \(H_m\) and \(H_n'\) behave in favor of \(H_0\) (i.e. \(H_m = H_n'\)), the resulting test based on the energy statistic has no power even if \(F_X \neq G_Y\) for all \(m,n\). For this part of proof, we assume \(m/N \to \pi_1 \in (0,1)\) as \(N \to \infty\) and the dimension \(d\) is fixed for convenience.

To begin, we choose
\[
F_X = N((\mu, \ldots, \mu)^\top, I_d), \quad G_Y = N((-\mu, \ldots, -\mu)^\top, I_d) \text{ and}
\]
\[
H_m = H_n' = N((0, \ldots, 0)^\top, \sigma_N^2 I_d),
\]
where \(\sigma_N^2 \propto (m + n)^p\) with \(p > 2\). It is worth noting that there is nothing special about the choice of normal distributions. The rest of the proof can be similarly applied to other distributions with some moment conditions. Under this setting, it is clear to see that
\[
E_d^2 = (1 - \epsilon)^2 \int_{S^{d-1}} \int_{\mathbb{R}} \left( F_{\beta^\top X}(t) - G_{\beta^\top Y}(t) \right)^2 dt d\lambda(\beta) = \delta_1^2 > 0.
\]
Next, we define the truncated random vectors $\tilde{X}$ and $\tilde{Y}$ coupled with $X$ and $Y$ as follows:

$$
\tilde{X} = \begin{cases} 
(0, \ldots, 0)^\top, & \text{if } X \sim F_X, \\
X/\sigma_N, & \text{if } X \sim H_m,
\end{cases}
$$

and

$$
\tilde{Y} = \begin{cases} 
(0, \ldots, 0)^\top, & \text{if } Y \sim G_Y, \\
Y/\sigma_N, & \text{if } Y \sim H'_m.
\end{cases}
$$

By the construction, $\tilde{X}$ and $\tilde{Y}$ have the same mixture distribution as:

$$
\tilde{X}, \tilde{Y} \sim (1 - \epsilon)\theta_0 + \epsilon N(0, I_d),
$$

where $\theta_0$ has the degenerating distribution at $(0, \ldots, 0)^\top$. At a high-level, when the energy statistic of the original samples is asymptotically equivalent to that of the truncated samples from the identical distributions, then it implies that the resulting test based on the energy statistic is asymptotically powerless. We start with the following lemma:

**Lemma C.4.** Suppose $\sigma_N^2 \gg N^p$ for some $p > 2$ and $m/N \to \pi_1 \in (0, 1)$ as $N \to \infty$. Let $\tilde{E}_{m,n}$ be the energy statistic based on $\{\tilde{X}_1, \ldots, \tilde{X}_m, \tilde{Y}_1, \ldots, \tilde{Y}_n\}$ coupled with the original samples $\{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$ and $E_{m,n}$ be the energy statistic based on the original samples. Then we have

$$
N \left( \frac{1}{\sigma_N} E_{m,n} - \tilde{E}_{m,n} \right) = o_P(1).
$$

**Proof.** Define the difference between the Euclidean distances based on the original samples $(X_1, X_2)$ and the truncated samples $(\tilde{X}_1, \tilde{X}_2)$ by

$$
\Delta_{m,n}(X_1, X_2) = \frac{1}{\sigma_N} ||X_1 - X_2|| - ||\tilde{X}_1 - \tilde{X}_2||.
$$

Note that there are four possible cases of $\Delta_{m,n}(X_1, X_2)$:

$$
\Delta_{m,n}(X_1, X_2) = \begin{cases} 
(a) \frac{1}{\sigma_N}||X_1 - X_2||, & \text{if } X_1, X_2 \sim F_X, \\
(b) \frac{1}{\sigma_N}||X_1 - X_2|| - \frac{1}{\sigma_N}||X_2||, & \text{if } X_1 \sim F_X, X_2 \sim H_m, \\
(c) \frac{1}{\sigma_N}||X_1 - X_2|| - \frac{1}{\sigma_N}||X_1||, & \text{if } X_1 \sim H_m, X_2 \sim F_X, \\
(d) 0, & \text{if } X_1, X_2 \sim H_m,
\end{cases}
$$

and for any case we have

$$
\mathbb{E} \left[ \Delta_{m,n}^2(X_1, X_2) \right] \lesssim \sigma_N^{-2}. \quad (47)
$$

For instance, the square of the case $(b)$ can be upper bounded by $\sigma_N^{-2}||X_1||^2$ based on the triangle inequality. Since $||X_1||^2$ has the noncentral chi-square distribution with $d$ degrees of freedom and $d\mu^2$ non-central parameter, the expectation of $||X_1||^2$ becomes $d(1 + \mu^2)$. Hence (47) holds. Similarly, $\mathbb{E} \left[ \Delta_{m,n}^2(X_i, X_j) \right] \lesssim \sigma_N^{-2}$, $\mathbb{E} \left[ \Delta_{m,n}^2(Y_i, Y_j) \right] \lesssim \sigma_N^{-2}$ and $\mathbb{E} \left[ \Delta_{m,n}^2(X_i, Y_j) \right] \lesssim \sigma_N^{-2}$ for any $i, j$.

Recall that the symmetrized kernel of the energy statistic is given by

$$
\chi_{\text{Energy}}^*(X_1, X_2; Y_1, Y_2)
$$
\[ \frac{1}{2}||X_1 - Y_1|| + \frac{1}{2}||X_1 - Y_1|| + \frac{1}{2}||X_2 - Y_1|| + \frac{1}{2}||X_2 - Y_2|| - ||X_1 - X_2|| - ||Y_1 - Y_2||. \]

Now define the energy statistic based on the truncated random samples:

\[ \tilde{E}_{m,n} = \frac{1}{(m)_2(n)_2} \sum_{i_1, i_2=1}^{m} \sum_{j_1, j_2=1}^{n} \chi_{\text{Energy}}(\tilde{X}_{i_1}, \tilde{X}_{i_2}, \tilde{Y}_{j_1}, \tilde{Y}_{j_2}). \]

Then the goal is to show

\[ N \left( \frac{1}{\sigma_N} E_{m,n} - \tilde{E}_{m,n} \right) \]

\[ = \frac{N}{(m)_2(n)_2} \sum_{i_1, i_2=1}^{m} \sum_{j_1, j_2=1}^{n} \left\{ \frac{1}{\sigma_N} \chi_{\text{Energy}}(X_{i_1}, X_{i_2}, Y_{j_1}, Y_{j_2}) - \chi_{\text{Energy}}(\tilde{X}_{i_1}, \tilde{X}_{i_2}, \tilde{Y}_{j_1}, \tilde{Y}_{j_2}) \right\} \]

\[ \overset{\text{let}}{=} \frac{N}{(m)_2(n)_2} \sum_{i_1, i_2=1}^{m} \sum_{j_1, j_2=1}^{n} D((X_{i_1}, \tilde{X}_{i_2}),(X_{i_2}, \tilde{X}_{i_1}); (Y_{j_1}, \tilde{Y}_{j_2}), (Y_{j_2}, \tilde{Y}_{j_1})) \]

\[ = o_P(1). \]

For this purpose, we use the Cauchy-Schwartz inequality to bound

\[ \mathbb{E} \left[ D(i_1, i_2; j_1, j_2) D(i_1', i_2'; j_1', j_2') \right] \leq \sqrt{\mathbb{E} \left[ D^2(i_1, i_2; j_1, j_2) \right]} \sqrt{\mathbb{E} \left[ D^2(i_1', i_2'; j_1', j_2') \right]}, \]

\[ \lesssim \sigma_N^{-2}, \]

for all \( i_1 \neq i_2, j_1 \neq j_2, i_1' \neq i_2', j_1' \neq j_2' \). For the second inequality, we used

\[ \mathbb{E} \left[ D^2(i_1, i_2; j_1, j_2) \right] \lesssim \mathbb{E}[\Delta_{m,n}^2(X_{i_1}, X_{i_2})] + \mathbb{E}[\Delta_{m,n}^2(X_{i_1}, Y_{j_1})] + \mathbb{E}[\Delta_{m,n}^2(X_{i_1}, Y_{j_2})] \]

\[ + \mathbb{E}[\Delta_{m,n}^2(X_{i_2}, Y_{j_1})] + \mathbb{E}[\Delta_{m,n}^2(X_{i_2}, Y_{j_2})] + \mathbb{E}[\Delta_{m,n}^2(Y_{j_1}, Y_{j_2})], \]

\[ \lesssim \sigma_N^{-2}. \]

As a consequence,

\[ \mathbb{E} \left[ N^2 \left( \sigma_N^{-1} E_{m,n} - \tilde{E}_{m,n} \right)^2 \right] \lesssim \frac{N^2}{\sigma_N^2}. \]

Under the given assumptions that \( \sigma_N^2 \asymp (m + n)^p \) with \( p > 2 \) and \( m/N \to \pi_1 \in (0, 1) \), we obtain \( N \left( \sigma_N^{-1} E_{m,n} - \tilde{E}_{m,n} \right) = o_P(1) \) as desired.

Since \( \tilde{E}_{m,n} \) has degeneracy of order one with the finite variance of the kernel, \( N \tilde{E}_{m,n} \) converges to an infinite weighted sum of chi-square random variables:

\[ N \tilde{E}_{m,n} \overset{d}{\to} \sum_{k=1}^{\infty} \lambda_k (Z_k^2 - 1), \]

43
for some \( \{\lambda_k\}_{k=1}^\infty \). Lemma C.4 implies that \( NE_{m,n}/\sigma_N \) converges to the same distribution:
\[
\frac{N}{\sigma_N} E_{m,n} \overset{d}{\to} \sum_{k=1}^\infty \lambda_k (Z_k^2 - 1).
\]
Furthermore, the permutation distribution of \( N\sigma_N^{-1} E_{m,n} \) is asymptotically equivalent to the limiting distribution of \( N\tilde{E}_{m,n} \) as shown in the next lemma.

**Lemma C.5.** Consider the settings in Lemma C.4. Let \( D \) where the kernel \( \pi \)
\[
\text{Proof.} \text{ Let } \{Z_1, \ldots, Z_{m+n}\} \text{ be the pooled samples of } \{X_1, \ldots, X_m, Y_1, \ldots, Y_n\} \text{ and similarly } \{\tilde{Z}_1, \ldots, \tilde{Z}_{m+n}\} \text{ be the pooled samples of } \{\tilde{X}_1, \ldots, \tilde{X}_m, \tilde{Y}_1, \ldots, \tilde{Y}_n\}. \text{ For any random permutation } \pi = (\pi(1), \ldots, \pi(N)) \text{ of } \{1, \ldots, N\}, \text{ we will show that }
\[
N\sigma_N^{-1} E_{m,n}(Z_\pi) - N\tilde{E}_{m,n}(\tilde{Z}_\pi) \overset{p}{\to} 0,
\]
where \( Z_\pi = (Z_{\pi(1)}, \ldots, Z_{\pi(N)}) \) and \( \tilde{Z}_\pi = (\tilde{Z}_{\pi(1)}, \ldots, \tilde{Z}_{\pi(N)}) \). If this is the case, then for independent \( \pi \) and \( \pi' \), the following limiting behavior
\[
(N\tilde{E}_{m,n}(\tilde{Z}_\pi), N\tilde{E}_{m,n}(\tilde{Z}_{\pi'})) \overset{d}{\to} (T, T')
\]
implies
\[
(N\sigma_N^{-1} E_{m,n}(\tilde{Z}_\pi), N\sigma_N^{-1} E_{m,n}(\tilde{Z}_{\pi'})) \overset{d}{\to} (T, T'),
\]
where \( T \) and \( T' \) are independent having the same distribution of \( R(t) \). Then Hoeffding’s condition in (B.4) presents (49). Indeed, (51) holds from Theorem A.1; hence it is enough to show (50) to complete the proof. Note that
\[
N\sigma_N^{-1} E_{m,n}(Z_\pi) - N\tilde{E}_{m,n}(\tilde{Z}_\pi) =
\]
\[
\frac{N}{(m2(n))^2} \sum_{i_1,i_2=1}^m \sum_{j_1,j_2=1}^n D((Z_{\pi(i_1)}, \tilde{Z}_{\pi(i_1)}), (Z_{\pi(i_2)}, \tilde{Z}_{\pi(i_2)}); (Z_{\pi(j_1+m)}, \tilde{Z}_{\pi(j_1+m)}), (Z_{\pi(j_2+m)}, \tilde{Z}_{\pi(j_2+m)}))
\]
where the kernel \( D \) is given in (48). Note further that
\[
E[D^2((Z_{\pi(i_1)}, \tilde{Z}_{\pi(i_1)}), (Z_{\pi(i_2)}, \tilde{Z}_{\pi(i_2)}); (Z_{\pi(j_1+m)}, \tilde{Z}_{\pi(j_1+m)}), (Z_{\pi(j_2+m)}, \tilde{Z}_{\pi(j_2+m)}))]
\]
\[
\lesssim E[\Delta^2_{m,n}(Z_{\pi(i_1)}, Z_{\pi(i_2)})] + E[\Delta^2_{m,n}(Z_{\pi(i_1)}, Z_{\pi(j_1+m)})] + E[\Delta^2_{m,n}(Z_{\pi(j_1+m)}, Z_{\pi(j_2+m)})]
\]
\[
+ E[\Delta^2_{m,n}(Z_{\pi(i_2)}, Z_{\pi(j_1+m)})] + E[\Delta^2_{m,n}(Z_{\pi(i_2)}, Z_{\pi(j_2+m)})] + E[\Delta^2_{m,n}(Z_{\pi(j_1+m)}, Z_{\pi(j_2+m)})]
\]
\[
\lesssim \sigma_N^2
\]
by (47) and similarly for the other cases. Then it is easy to see that
\[
E[(N\sigma_N^{-1} E_{m,n}(Z_\pi) - N\tilde{E}_{m,n}(\tilde{Z}_\pi))^2] \lesssim \frac{N^2}{\sigma_N^2} = o(1)
\]
when \( \sigma_N^2 \asymp Np \) for some \( p > 2 \). This implies (50), which completes the proof.
Combining the previous results, we see that
\[ N\sigma_N^{-1} E_{m,n} - \nu'_{\alpha,m,n} = N \bar{E}_{m,n} - \bar{\nu}'_{\alpha,m,n} + o_P(1), \]
where \( \nu'_{\alpha,m,n} \) and \( \bar{\nu}'_{\alpha,m,n} \) are the upper \( \alpha \) quantile of the permutation distributions of \( N\sigma_N^{-1} E_{m,n} \) and \( N \bar{E}_{m,n} \). This implies that
\[
\lim_{N \to \infty} P\left( E_{m,n} \geq c'_{\alpha,m,n} \right) = \lim_{N \to \infty} P\left( N\sigma_N^{-1} E_{m,n} \geq \nu'_{\alpha,m,n} \right) = \lim_{N \to \infty} P\left( N \bar{E}_{m,n} \geq \bar{\nu}'_{\alpha,m,n} \right) \leq \alpha.
\]
Hence the result follows.

C.6 Proof of Lemma 4.1
Let \( \beta^T Z \sim \frac{1}{2} F_{\beta^T X} + \frac{1}{2} G_{\beta^T Y} \). From the definition of the multivariate CvM distance,
\[
W_d^2 = \mathbb{E}_{\beta,Z} \left[ \left( F_{\beta^T X} - G_{\beta^T Y} \right)^2 \right] \geq \left\{ \mathbb{E}_{\beta,Z} \left| F_{\beta^T X} - G_{\beta^T Y} \right| \right\}^2,
\]
where we used Jensen’s inequality. From the definition of \( \beta^T Z \),
\[
\mathbb{E}_{\beta,Z} \left| F_{\beta^T X} - G_{\beta^T Y} \right| = \frac{1}{2} \mathbb{E}_{\beta,X} \left| F_{\beta^T X'} - G_{\beta^T Y'} \right| + \frac{1}{2} \mathbb{E}_{\beta,Y} \left| F_{\beta^T Y'} - G_{\beta^T Y'} \right| \\
\geq \frac{1}{2} \mathbb{E}_{\beta} \left| \mathbb{E}_{X,X,Y} \left( I(\beta^T X \leq \beta^T X') - I(\beta^T Y \leq \beta^T X') \right) \right| \\
+ \frac{1}{2} \mathbb{E}_{\beta} \left| \mathbb{E}_{X,Y,Y'} \left( I(\beta^T X \leq \beta^T Y') - I(\beta^T Y \leq \beta^T Y') \right) \right|,
\]
where we used Jensen’s inequality once again. The last expression can be simplified based on \( P(\beta^T X \leq \beta^T X') = P(\beta^T Y \leq \beta^T Y') = 1/2 \), which becomes
\[
\mathbb{E}_{\beta} \left| \frac{1}{2} - P_{X,Y} \left( \beta^T X \leq \beta^T Y \right) \right|.
\]
Therefore,
\[
W_d^2 \geq \left\{ \int_{S^{d-1}} \frac{1}{2} - P_{X,Y} \left( \beta^T X \leq \beta^T Y \right) d\lambda(\beta) \right\}^2,
\]
which completes the proof.

C.7 Proof of Theorem 4.1
The minimax lower bound is based on a standard application of Neyman-Pearson lemma (see, e.g. Baraud, 2002):
\[
\inf_{\phi \in \Phi_{m,n}} \sup_{F_X,G_Y \in \mathcal{F}(\epsilon_{m,n})} \mathbb{P}^n \left( \phi = 0 \right) \geq 1 - \alpha - \sup_{A \in \mathcal{A}} \left| P_0^n(A) - F_1^n(A) \right|
\]
\[ \geq 1 - \alpha - \sqrt{\frac{1}{2} KL(P_{1m,n}^{m,n}, P_{0m,n}^{m,n})}, \]  
\[ (52) \]

where the second inequality is by Pinsker’s inequality.

Recall the example considered in Lemma 4.2:

\[ X^* := (Z_1, 0, \ldots, 0)^T \quad \text{and} \quad Y^* := (Z_2, 0, \ldots, 0)^T, \]

where \( Z_1 \sim N(\mu_{X^*}, 1) \) and \( Z_2 \sim N(\mu_{Y^*}, 1) \). Under the null, we let \( \mu_{X^*} = \mu_{Y^*} = 0 \) and under the alternative, we let

\[ \mu_{X^*} = \frac{\sqrt{2}(1 - \alpha - \xi)}{\sqrt{m}} \quad \text{and} \quad \mu_{Y^*} = -\frac{\sqrt{2}(1 - \alpha - \xi)}{\sqrt{n}}. \]

Lemma 4.2, we have \( F_{X^*}, G_{Y^*} \in \mathcal{F}(\epsilon_{m,n}^*) \) for all \( d \). In this case, the Kullback-Leibler divergence is calculated by

\[ KL(P_{1m,n}^{m,n}, P_{0m,n}^{m,n}) = \frac{m}{2} \mu_{X^*}^2 + \frac{n}{2} \mu_{Y^*}^2 = 2(1 - \alpha - \xi)^2. \]

By plugging this into (52), we conclude that

\[ \inf_{\phi \in \Phi_{m,n,\alpha}} \sup_{F_{X^*}, G_{Y^*} \in \mathcal{F}(\epsilon_{m,n}^*)} P_{1m,n} (\phi = 0) \geq \xi. \]

Hence the result follows.

C.8 Proof of Theorem 4.2

To finish the proof, we need to verify the condition in (16). To start, we present two lemmas: in Lemma C.6, we bound the variance of \( U_{\text{CVM}} \) and in Lemma C.7, we consider the two moments of \( U_{\text{CVM}} \) under permutations.

**Lemma C.6 (Variance of \( U_{\text{CVM}} \)).** Consider the U-statistic in (6). Then there exist universal constants \( C_1, C_2, C_3, C_4 > 0 \) such that

\[ \text{Var} [U_{m,n}] \leq C_1 \mathbb{E}[U_{m,n}] \cdot \left( \frac{1}{m} + \frac{1}{n} \right) + \frac{C_2}{m^2} + \frac{C_3}{n^2} + \frac{C_4}{mn}. \]

**Proof.** Let \( \psi(x_1, x_2, x_3; y_1, y_2, y_3) \) be the symmetrized kernel of \( h^*(x_1, x_2, x_3; y_1, y_2, y_3) \) in (32) and define \( \psi_{c,d} \) as in (28). Further denote the variance of \( \psi_{c,d} \) by \( \sigma_{c,d}^2 \) as in (29). Then the variance of \( U_{\text{CVM}} \) can be written as (Lemma B.3)

\[ \text{Var} (U_{\text{CVM}}) = \sum_{c=0}^{3} \sum_{d=0}^{3} \binom{3}{c} \binom{3}{d} \binom{m-3}{3-c} \binom{n-3}{3-d} \sigma_{c,d}^2. \]

(53)

First we bound \( \sigma_{1,0}^2 \). After applying the law of total expectation repeatedly, we obtain that

\[ \psi_{1,0}(x_1) - \mathbb{E}[\psi_{1,0}(x_1)] = \mathbb{E}_{\beta,X} \left[ (I(\beta^T x_1 \leq \beta^T X) - F_{\beta^T X}(\beta^T X)) \cdot \left( G_{\beta^T Y}(\beta^T X) - F_{\beta^T X}(\beta^T X) \right) \right] \]
Using $(f_1(x_1) + f_2(x_1) + f_3(x_1))^2 \leq 3f_1^2(x_1) + 3f_2^2(x_1) + 3f_3^2(x_1)$, we obtain that
\[
\sigma_{1,0}^2 = \mathbb{E}_X \left[ (\psi_{1,0}(X) - \mathbb{E}_X[\psi_{1,0}(X)])^2 \right] \leq 3 \mathbb{E}_X \left[ f_1^2(X) \right] + 3 \mathbb{E}_X \left[ f_2^2(X) \right] + 3 \mathbb{E}_X \left[ f_3^2(X) \right].
\]
By Cauchy-Schwartz inequality, the first two terms are bounded by
\[
\mathbb{E}_X \left[ f_1^2(X) \right] \leq \mathbb{E}_{\beta,X} \left[ (f_{\beta,X}(\beta^T X) - G_{\beta,Y}(\beta^T X))^2 \right],
\]
\[
\mathbb{E}_X \left[ f_2^2(X) \right] \leq \mathbb{E}_{\beta,Y} \left[ (f_{\beta,Y}(\beta^T Y) - G_{\beta,Y}(\beta^T Y))^2 \right].
\]
Since $0 \leq \mathbb{E}_\beta \left[ (f_{\beta,X}(\beta^T x_1) - G_{\beta,Y}(\beta^T x_1))^2 \right] \leq 1$ for all $x_1 \in \mathbb{R}^d$, the third term is also bounded by
\[
\mathbb{E}_X \left[ f_3^2(X) \right] \leq \frac{1}{4} \mathbb{E}_X \left[ (\mathbb{E}_\beta \left[ (f_{\beta,X}(\beta^T X) - G_{\beta,Y}(\beta^T X))^2 \right])^2 \right]
\leq \frac{1}{4} \mathbb{E}_{\beta,X} \left[ (f_{\beta,X}(\beta^T X) - G_{\beta,Y}(\beta^T X))^2 \right].
\]
Therefore, the following fact
\[
\mathbb{E}[U_{\text{CVM}}] = \frac{1}{2} \mathbb{E}_{\beta,X} \left[ (f_{\beta,X}(\beta^T X) - G_{\beta,Y}(\beta^T X))^2 \right] + \frac{1}{2} \mathbb{E}_{\beta,Y} \left[ (f_{\beta,Y}(\beta^T Y) - G_{\beta,Y}(\beta^T Y))^2 \right],
\]
implies $\sigma_{1,0}^2 \lesssim \mathbb{E}[U_{\text{CVM}}]$. Similarly we have $\sigma_{2,1}^2 \lesssim \mathbb{E}[U_{\text{CVM}}]$. The rest of $\sigma_{c,d}^2$ can be uniformly bounded since $\psi$ is uniformly bounded. Hence the result follows.

**Lemma C.7** (Two moments of the permutation distribution). The first and second moments of the permutation distribution of $U_{\text{CVM}}$ are
\[
\mathbb{E}_\pi [U_{\text{CVM}}] = 0 \quad \text{and} \quad \mathbb{E}_\pi \left[ U_{\text{CVM}}^2 \right] \leq C \cdot \left( \frac{1}{m} + \frac{1}{n} \right)^2,
\]
where $C$ is a universal constant independent of $d, m, n$.

**Proof.** Working directly with the kernel $h$ is less intuitive to understand the moments of $U_{\text{CVM}}$ under permutations. So we consider the three order kernel $h^\ast$ in (32). Then from Lemma B.2, we have
\[
U_{\text{CVM}} = \frac{1}{(m)_3(n)_3} \sum_{i_1 \neq i_2 \neq i_3} \sum_{j_1 \neq j_2 \neq j_3} \sum_{i_1 \neq i_2 \neq i_3 \neq j_1 \neq j_2 \neq j_3} \mathbb{E} \left[ h^*(X_{i_1}, X_{i_2}, X_{i_3}; Y_{j_1}, Y_{j_2}, Y_{j_3}) \right].
\]
Let $\{Z_1, \ldots, Z_{m+n}\} = \{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$ be the pooled samples and let $\mathbb{E}_\pi$ be the expectation over permutations given $\{Z_1, \ldots, Z_{m+n}\}$. Then the first moment of $U_{\text{CVM}}$ becomes
\[
\mathbb{E}_\pi [U_{\text{CVM}}] = \mathbb{E}_\pi \left[ h^*(Z_{\pi(1)}, Z_{\pi(2)}, Z_{\pi(3)}; Z_{\pi(m+1)}, Z_{\pi(m+2)}, Z_{\pi(m+3)}) \right],
\]
and
where \( \pi := (\pi(1), \ldots, \pi(m+n)) \) is a random permutation of \( \{1, \ldots, m+n\} \). Notice that

\[
\tilde{h}^*(x_1, x_2, x_3; y_1, y_2, y_3) = -h^*(y_1, x_2, x_3; x_1, y_2, y_3).
\]

This observation shows that the conditional expectation of \( h^* \) given the permutations \( \mathcal{P}(\pi) = \{\pi(2), \pi(3), \pi(m+2), \pi(m+3)\} \) becomes zero, i.e.

\[
\mathbb{E}_{\pi(1), \pi(m+1)} \left[ h^*(Z_{\pi(1)}, Z_{\pi(2)}, Z_{\pi(3)}; Z_{\pi(m+1)}, Z_{\pi(m+2)}, Z_{\pi(m+3)}) \big| \mathcal{P}(\pi) \right] = 0,
\]

for all \( \mathcal{P}(\pi) \). Hence, \( \mathbb{E}_\pi [U_{\text{CVM}}] = 0 \) by the law of total expectation.

Next we calculate the second moment of \( U_{\text{CVM}} \) under permutations where

\[
U_{\text{CVM}}^2 = \frac{1}{(m)_3(n)_3} \sum_{i_1 \neq i_2 \neq i_3} \sum_{j_1 \neq j_2 \neq j_3} \sum_{i_1', i_2', i_3'} \sum_{j_1', j_2', j_3'} \{ h^*(Z_{i_1}, Z_{i_2}, Z_{i_3}; Z_{j_1+m}, Z_{j_2+m}, Z_{j_3+m}) h^*(Z_{i_1'}, Z_{i_2'}, Z_{i_3'}; Z_{j_1'+m}, Z_{j_2'+m}, Z_{j_3'+m}) \}.
\]

Consider the three cases in which

- Case 1: \( \{i_1, i_2, i_3\} \cap \{i_1', i_2', i_3'\} = 0 \) and \( \{j_1, j_2, j_3\} \cap \{j_1', j_2', j_3'\} = 0 \),
- Case 2: \( \{i_1, i_2, i_3\} \cap \{i_1', i_2', i_3'\} = 1 \) and \( \{j_1, j_2, j_3\} \cap \{j_1', j_2', j_3'\} = 0 \),
- Case 3: \( \{i_1, i_2, i_3\} \cap \{i_1', i_2', i_3'\} = 0 \) and \( \{j_1, j_2, j_3\} \cap \{j_1', j_2', j_3'\} = 1 \).

For the first case, there is no overlap between \( \{Z_{\pi(i_1)}, Z_{\pi(i_2)}, Z_{\pi(i_3)}; Z_{\pi(j_1+m)}, Z_{\pi(j_2+m)}, Z_{\pi(j_3+m)}\} \) and \( \{Z_{\pi(i_1')}, Z_{\pi(i_2')}, Z_{\pi(i_3')}; Z_{\pi(j_1'+m)}, Z_{\pi(j_2'+m)}, Z_{\pi(j_3'+m)}\} \). Then by conditioning on \( \mathcal{P}'(\pi) = \{\pi(i_2), \pi(i_3), \pi(j_2 + m), \pi(j_3 + m), \pi(i_1'), \pi(i_2'), \pi(i_3'), \pi(j_1' + m), \pi(j_2' + m), \pi(j_3' + m)\} \), we see from \( f(x_1, x_2, x_3; y_1, y_2, y_3) = -f(y_1, x_2, x_3; x_1, y_2, y_3) \) that

\[
\mathbb{E}_{\pi(i_1), \pi(j_1+m)} \left[ h^*(Z_{\pi(i_1)}, Z_{\pi(i_2)}, Z_{\pi(i_3)}; Z_{\pi(j_1+m)}, Z_{\pi(j_2+m)}, Z_{\pi(j_3+m)}) \right] = 0,
\]

for all \( \mathcal{P}'(\pi) \). As a result, the law of total expectation shows that

\[
\mathbb{E}_{\pi} \left[ h^*(Z_{\pi(i_1)}, Z_{\pi(i_2)}, Z_{\pi(i_3)}; Z_{\pi(j_1+m)}, Z_{\pi(j_2+m)}, Z_{\pi(j_3+m)}) \right] = 0.
\]

Similarly, it is seen that (54) holds for Case 2 and Case 3. For the other cases (other than Case 1, Case 2 and Case 3), we simply use the bound

\[
\mathbb{E}_{\pi} \left[ h^*(Z_{\pi(i_1)}, Z_{\pi(i_2)}, Z_{\pi(i_3)}; Z_{\pi(j_1+m)}, Z_{\pi(j_2+m)}, Z_{\pi(j_3+m)}) \right] \leq 1.
\]

The size of the other cases are at most \( \prod_{i=0}^{4} (m-i) \times \prod_{j=0}^{6} (n-j) \) or \( \prod_{i=0}^{5} (m-i) \times \prod_{j=0}^{5} (n-j) \) or \( \prod_{i=0}^{6} (m-i) \times \prod_{j=0}^{4} (n-j) \) up to scaling factors. This concludes that

\[
\mathbb{E}_{\pi} [U_{\text{CVM}}^2] \leq C \left( \frac{1}{m} + \frac{1}{n} \right)^2
\]

as desired. \( \square \)

48
Now, we verify the condition in (16). Using the Chebyshev’s inequality and Lemma C.7,
\[ P(\pi(U_{\text{CvM}} \geq t)) \leq \frac{\mathbb{E}_\pi[U_{\text{CvM}}^2]}{t^2} \leq \frac{C_0}{t^2} \left( \frac{1}{m} + \frac{1}{n} \right)^2. \]

As a result, the permutation critical value \( c^*_{\alpha,m,n} \) is bounded by \( \sqrt{C_0/\alpha(1/m + 1/n)} \) with probability one. This implies that its \( \xi/2 \) upper quantile \( c^*_{\xi/2} \) is also bounded by
\[ c^*_{\xi/2} \leq \sqrt{\frac{C_0}{\alpha} \left( \frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}} \right)^2}. \]

From Lemma C.6, we have
\[ \sqrt{\frac{\xi}{2}} \text{Var}[U_{\text{CvM}}] \leq \sqrt{\frac{\xi}{2}} \left\{ C_1 \mathbb{E}[U_{\text{CvM}}] \cdot \left( \frac{1}{m} + \frac{1}{n} \right) + \frac{C_2}{m^2} + \frac{C_3}{n^2} + \frac{C_4}{mn} \right\} \]
\[ \leq C_5 \left( \frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}} \right)^2. \]

By choosing a sufficiently large \( c > 0 \) in (15), we conclude that
\[ \mathbb{E}_1[U_{\text{CvM}}] \geq c^*_{\xi/2} + \sqrt{\frac{\xi}{2}} \text{Var}[U_{\text{CvM}}]. \]

**C.9 Proof of Lemma 4.3**

Let \( \sigma^2_0 \) and \( \sigma^2_1 \) be the variance of
\[ \frac{1}{2} \left[ h_*(X_{2i-1}, X_{2i}; Y_{2i-1}, Y_{2i}) + h_*(X_{2i}, X_{2i-1}; Y_{2i-1}, Y_{2i}) \right], \]
respectively. From the boundedness of \( h_* \), we have \( 0 < \sigma^2_0, \sigma^2_1 < \infty \) for all \( m, n \). By the central limit theorem, the null distribution approximates
\[ \frac{\sqrt{M}L_{\text{CvM}}}{\sigma_0} \xrightarrow{d} N(0,1) \text{ under } H_0, \]
which implies that \( \sqrt{M}\sigma_0^{-1}\eta_{\alpha,m,n} \to -z_\alpha \) where \( z_\alpha \) is the \( \alpha \) quantile of the standard normal distribution and \( z_\alpha < 0 \) for \( \alpha < 1/2 \). Hence, the power function approximates
\[ \lim_{N \to \infty} \mathbb{P}_1(L_{\text{CvM}} \geq \eta_{\alpha,m,n}) = \lim_{N \to \infty} \mathbb{P}_1 \left( \sqrt{M} \frac{(L_{\text{CvM}} - W^2_d)}{\sigma_1} \geq \frac{\sqrt{M}\eta_{\alpha,m,n}}{\sigma_1} - \frac{\sqrt{M}W^2_d}{\sigma_1} \right) \]
\[ = \lim_{N \to \infty} \mathbb{P}_1 \left( \sqrt{M} \frac{(L_{\text{CvM}} - W^2_d)}{\sigma_1} \geq -\sigma_0 \frac{z_\alpha}{\sigma_1} - \frac{\sqrt{M}W^2_d}{\sigma_1} \right) \]
\[ \leq \lim_{N \to \infty} \mathbb{P}_1 \left( \sqrt{M} \frac{(L_{\text{CvM}} - W^2_d)}{\sigma_1} \geq -\frac{\sqrt{M}W^2_d}{\sigma_1} \right) \]
where the last equality uses
\[ \sqrt{M} \frac{L_{CvM} - W_d^2}{\sigma_1} \xrightarrow{d} N(0,1) \text{ under } H_1 \]
and \( \sqrt{M} W_d^2 \to 0 \) by the assumption. This completes the proof.

C.10 Proof of Theorem 5.1

Chakraborty and Chaudhuri (2017) show that the power based on the CQ and WMW tests are asymptotically identical under some location models, which include ours (see Theorem 2.2 of Chakraborty and Chaudhuri, 2017). Here, we will establish the rest of identities. Since \( T_{CQ} \) is easier to deal with than \( T_{WMW} \), we will establish 1) the relationship between the CvM statistic and the CQ statistic and 2) the relationship between the energy statistic and the CQ statistic in order and will show their asymptotic identities.

Remark C.1. Ramdas et al. (2015) observe that the tests based on \( T_{CQ} \), the Gaussian MMD and a variant of the energy statistic involving a tuning parameter are the same under high-dimensional location models (Theorem 3 of Ramdas et al., 2015). Our approach is different to theirs in that we do not consider the variant of the energy statistic but the original energy statistic without tuning parameters.

The proof is lengthy and involves Taylor approximations repeatedly. We believe that the same result can be obtained under much weaker assumptions (as evidenced by our simulations). For example, one can relax the independent assumption among covariates to some mixing conditions as in Chakraborty and Chaudhuri (2017). However, it will make the proof much longer and less straightforward. Hence, we prove the result admittedly under restricted assumptions and provide simulation results to support our claim.

Overview of the proof.

1. In Lemma C.8, we establish the asymptotic identity between the CvM statistic and the CQ statistic (after properly centering and scaling). We prove the result by applying Taylor approximations repeatedly and in Lemma C.9, we show that the remainder term of the approximation is negligible.

2. In Lemma C.10, we establish the asymptotic identity between the energy statistic and the CQ statistic (after properly centering and scaling) by following similar procedures applied for Lemma C.8.

3. At the end of this section, we collect the results and establish the desired identities.

We begin with establishing the relationship between the CvM statistic and the CQ statistic.

Lemma C.8. Consider the same location-shifted model assumption used in Theorem 5.1 with \( \mu_X - \mu_Y = \mu \). Then for every fixed \( m, n \) and as \( d \to \infty \),
\[ \frac{2\pi\sqrt{3d}}{\gamma_{m,n}} (U_{CvM} - \frac{\mu^2}{2\pi\sqrt{3}\sigma^2}) = \frac{1}{\sqrt{d\sigma^2\gamma_{m,n}}} \left( T_{CQ} - d\mu^2 \right) + o_P(1) \xrightarrow{d} N(0,1), \]

50
where
\[
\gamma_{m,n} = \sqrt{\frac{2}{m(m-1)}} + \frac{2}{n(n-1)} + \frac{4}{mn}.
\] (55)

Proof. To simplify the notations, let us denote the inner products by
\[
\begin{align*}
\mathcal{K}_1 &= \frac{(X_1 - Y_1)\mathbb{T}(X_2 - Y_1)}{||X_1 - Y_1||||X_2 - Y_1||}, & \mathcal{K}_2 &= \frac{(X_1 - Y_2)\mathbb{T}(X_2 - Y_2)}{||X_1 - Y_2||||X_2 - Y_2||} \\
\mathcal{K}_3 &= \frac{(Y_1 - X_1)\mathbb{T}(Y_2 - X_1)}{||Y_1 - X_1||||Y_2 - X_1||}, & \mathcal{K}_4 &= \frac{(Y_1 - X_2)\mathbb{T}(Y_2 - X_2)}{||Y_1 - X_2||||Y_2 - X_2||}.
\end{align*}
\]

Then the symmetrized kernel of \( U_{\text{CVM}} \) becomes
\[
2\pi\tilde{h}(X_1, X_2; Y_1, Y_2) = \frac{2\pi}{3} - \frac{1}{2} \left( \arccos(\mathcal{K}_1) + \arccos(\mathcal{K}_2) + \arccos(\mathcal{K}_3) + \arccos(\mathcal{K}_4) \right).
\]

Since \( \tilde{h} \) is location invariant, without loss of generality, we assume \( \mathbb{E} [X] = 0 \) and \( \mathbb{E} [Y] = \mu 1^\mathbb{T} \).

Using the Taylor expansion of \( \arccos(x) \) around \( x = 1/2 \) and the central limit theorem (CLT), one can show that
\[
\Psi (X_1 - Y_1, X_2 - Y_1) = \frac{\pi}{3} + \frac{1}{\sqrt{3}} - \frac{2}{\sqrt{3}||X_1 - Y_1||||X_2 - Y_1||} + o_P(d^{-1/2})
\] (56)
under the assumption on \( \mu_X - \mu_Y = \mu \) in (19). Hence, the kernel \( \tilde{h} \) is approximated by
\[
2\pi\tilde{h}(X_1, X_2; Y_1, Y_2) \approx \frac{1}{\sqrt{3}} (\mathcal{K}_1 + \mathcal{K}_2 + \mathcal{K}_3 + \mathcal{K}_4) - \frac{2}{\sqrt{3}} + o_P(d^{-1/2}).
\]

For the rest of the proof, we will show that \( \mathcal{K}_1 + \mathcal{K}_2 + \mathcal{K}_3 + \mathcal{K}_4 \) is dominated by
\[
h_{\text{CQ}}(X_1, X_2; Y_1, Y_2) + h_{\text{CQ}}(X_1, X_2; Y_2, Y_1)
\]
\[
= (X_1 - Y_1)\mathbb{T}(X_2 - Y_2) + (X_1 - Y_2)\mathbb{T}(X_2 - Y_1)
\]
and therefore \( U_{\text{CVM}} \) is essentially the same as \( T_{\text{CQ}} \) as \( d \to \infty \). In order to show this, further denote
\[
\varrho_{x_1y_1} = ||X_1 - Y_1||, \quad \varrho_{x_1y_2} = ||X_1 - Y_2||,
\]
\[
\varrho_{x_2y_1} = ||X_2 - Y_1||, \quad \varrho_{x_2y_2} = ||X_2 - Y_2||.
\]

Note then by the law of large number and the continuous mapping theorem, we have
\[
\frac{1}{\varrho_{x_1y_1}\varrho_{x_2y_1}} \overset{p}{\to} 2\sigma^2 + \mu^2.
\] (57)

In addition, the central limit theorem provides
\[
\frac{1}{\sqrt{d}}X_1\mathbb{T}X_2 = O_P(1),
\]
\[
\frac{1}{\sqrt{d}}X_1\mathbb{T}Y_1 = O_P(1),
\] (58)
\[
\frac{1}{\sqrt{d}}(Y_1\mathbb{T}Y_2 - d\mu^2) = \frac{1}{\sqrt{d}}Y_1\mathbb{T}Y_2 + o(1) = O_P(1) \quad \text{from the condition in (19)}.
\]
By (57) and (58), we can approximate
\[ \mathcal{K}_1 + \mathcal{K}_2 + \mathcal{K}_3 + \mathcal{K}_4 \]
= \( \frac{1}{2d\sigma^2 + d\mu^2} \left\{ X_1^T X_2 - X_1^T Y_1 - X_2^T Y_1 + X_1^T X_2 - X_1^T Y_2 \right. \\
- X_2^T Y_2 + Y_1^T Y_2 - Y_1^T X_1 - Y_2^T X_1 + Y_1^T Y_2 - Y_1^T X_2 - Y_2^T X_2 \bigg\} \\
+ \frac{Y_1^T Y_1}{\ell_{x_1y_1} \ell_{x_2y_1}} + \frac{Y_2^T Y_2}{\ell_{x_1y_2} \ell_{x_2y_2}} + \frac{X_1^T X_1}{\ell_{x_1y_1} \ell_{x_1y_1}} + \frac{X_2^T X_2}{\ell_{x_2y_2} \ell_{x_2y_2}} + o_P(d^{-1/2}) \\
= \frac{1}{2d\sigma^2 + d\mu^2} \left\{ (X_1 - Y_1)^T (X_2 - Y_2) + (X_1 - Y_2)^T (X_2 - Y_1) \right\} \\
- \frac{1}{2d\sigma^2 + d\mu^2} \left\{ Y_1^T X_1 + Y_2^T X_1 + Y_1^T X_2 + Y_2^T X_2 \right\} \\
+ \frac{Y_1^T Y_1}{\ell_{x_1y_1} \ell_{x_2y_1}} + \frac{Y_2^T Y_2}{\ell_{x_1y_2} \ell_{x_2y_2}} + \frac{X_1^T X_1}{\ell_{x_1y_1} \ell_{x_1y_1}} + \frac{X_2^T X_2}{\ell_{x_2y_2} \ell_{x_2y_2}} + o_P(d^{-1/2}). \\
\]
Hence,
\[ \mathcal{K}_1 + \mathcal{K}_2 + \mathcal{K}_3 + \mathcal{K}_4 \]
= \( \frac{1}{2d\sigma^2 + d\mu^2} \left\{ h_{CL}(X_1, X_2; Y_1, Y_2) + h_{CL}(X_1, X_2; Y_2, Y_1) \right\} + \mathcal{R}_d + o_P(d^{-1/2}), \)
where
\[ \mathcal{R}_d = \frac{Y_1^T Y_1}{\ell_{x_1y_1} \ell_{x_2y_1}} + \frac{Y_2^T Y_2}{\ell_{x_1y_2} \ell_{x_2y_2}} + \frac{X_1^T X_1}{\ell_{x_1y_1} \ell_{x_1y_1}} + \frac{X_2^T X_2}{\ell_{x_2y_2} \ell_{x_2y_2}} \\
- \frac{1}{2d\sigma^2 + d\mu^2} \left\{ Y_1^T X_1 + Y_2^T X_1 + Y_1^T X_2 + Y_2^T X_2 \right\}. \]
Now if the centered residual is asymptotically negligible,
\[ \mathcal{R}_d - 2 = o_P(d^{-1/2}), \]
then we can arrive at
\[ 2\pi \tilde{h}(X_1, X_2; Y_1, Y_2) \]
= \( \frac{1}{\sqrt{3}} (\mathcal{K}_1 + \mathcal{K}_2 + \mathcal{K}_3 + \mathcal{K}_4) - \frac{2}{\sqrt{3}} + o_P(d^{-1/2}) \)
= \( \frac{1}{\sqrt{3}d(2\sigma^2 + \mu^2)} \left\{ h_{CL}(X_1, X_2; Y_1, Y_2) + h_{CL}(X_1, X_2; Y_2, Y_1) \right\} + \frac{1}{\sqrt{3}}(\mathcal{R}_d - 2) + o_P(d^{-1/2}) \)
= \( \frac{1}{\sqrt{3}d(2\sigma^2 + \mu^2)} \left\{ h_{CL}(X_1, X_2; Y_1, Y_2) + h_{CL}(X_1, X_2; Y_2, Y_1) \right\} + o_P(d^{-1/2}). \)
However, it is hard to deal with \((X_i^T X_i)/(\ell_{x_iy_1} \ell_{x_iy_2})\) and \((Y_i^T Y_i)/(\ell_{x_1y_1} \ell_{x_2y_2})\) in \(\mathcal{R}_d\) directly. So we need to further approximate \(\mathcal{R}_d\). To do so, we apply the CLT and the weak law of
large number to have

\[
\frac{Y_1^\top Y_1}{\varrho_{y_1} \varrho_{y_2}} = \frac{Y_1^\top Y_1}{d(2\sigma^2 + \mu^2)} - \frac{\sigma^2 + \mu^2}{2\sigma^2 + \mu^2} + \frac{d(\sigma^2 + \mu^2)}{\varrho_{y_1} \varrho_{y_2}} + o_P(d^{-1/2}).
\]

In order to deal with the inverse of \(\varrho_{x_1y_1} \varrho_{x_2y_1}\), we use the Taylor expansion of \(f(x) = 1/\sqrt{x}\) around at \(x = 4\) to show

\[
\frac{d(\sigma^2 + \mu^2)}{\varrho_{x_1y_1} \varrho_{x_2y_1}} = \frac{1}{2} + \frac{1}{4} - \frac{\varrho_{x_1y_1} \varrho_{x_2y_1}^2}{16d^2(\sigma^2 + \mu^2)^2} + o_P(d^{-1/2}).
\]

Hence, we have

\[
\frac{Y_1^\top Y_1}{\varrho_{x_1y_1} \varrho_{x_2y_1}} = \frac{3}{4} - \frac{\sigma^2 + \mu^2}{2\sigma^2 + \mu^2} - \frac{\varrho_{x_1y_1} \varrho_{x_2y_1}^2}{16d^2(\sigma^2 + \mu^2)^2} + \frac{Y_1^\top Y_1}{d(\sigma^2 + \mu^2)} + o_P(d^{-1/2}).
\]

Similarly,

\[
\frac{Y_2^\top Y_2}{\varrho_{x_1y_1} \varrho_{x_2y_1}} = \frac{3}{4} - \frac{\sigma^2 + \mu^2}{2\sigma^2 + \mu^2} - \frac{\varrho_{x_1y_2} \varrho_{x_2y_2}^2}{16d^2(\sigma^2 + \mu^2)^2} + \frac{Y_2^\top Y_2}{d(\sigma^2 + \mu^2)} + o_P(d^{-1/2}).
\]

\[
\frac{X_1^\top X_1}{\varrho_{x_1y_1} \varrho_{x_2y_1}} = \frac{3}{4} - \frac{\sigma^2 + \mu^2}{2\sigma^2 + \mu^2} - \frac{\varrho_{x_1y_1} \varrho_{x_2y_1}^2}{16d^2(\sigma^2 + \mu^2)^2} + \frac{X_1^\top X_1}{d(\sigma^2 + \mu^2)} + o_P(d^{-1/2}).
\]

\[
\frac{X_2^\top X_2}{\varrho_{x_1y_1} \varrho_{x_2y_1}} = \frac{3}{4} - \frac{\sigma^2 + \mu^2}{2\sigma^2 + \mu^2} - \frac{\varrho_{x_2y_1} \varrho_{x_2y_2}^2}{16d^2(\sigma^2 + \mu^2)^2} + \frac{X_2^\top X_2}{d(\sigma^2 + \mu^2)} + o_P(d^{-1/2}).
\]

Let us denote

\[
\varpi(X_1, X_2, Y_1, Y_2) := Y_1^\top Y_1 + Y_2^\top Y_2 + X_1^\top X_1 + X_2^\top X_2 - Y_1^\top X_1 - Y_2^\top X_1 - Y_1^\top X_2 - Y_2^\top X_2.
\]

By combining the above results and letting \(\mathcal{R}_d^*\) be

\[
\mathcal{R}_d^* = \frac{1}{d(2\sigma^2 + \mu^2)} \varpi(X_1, X_2, Y_1, Y_2) + \frac{2\sigma^2 - \mu^2}{2\sigma^2 + \mu^2} \frac{1}{16d^2(\sigma^2 + \mu^2)^2} \left\{ \varrho_{x_1y_1} \varrho_{x_2y_1}^2 + \varrho_{x_1y_2} \varrho_{x_2y_2}^2 + \varrho_{x_1y_2} \varrho_{x_1y_2}^2 + \varrho_{x_2y_1} \varrho_{x_2y_2}^2 \right\},
\]

the residual \(\mathcal{R}_d\) can be approximated by

\[
\mathcal{R}_d = \mathcal{R}_d^* + o_P(d^{-1/2}).
\]

Let \(\eta_k\) be the \(k\)th moment of each coordinate of \(V\), that is

\[
\eta_k = E\left[ V_1^k \right].
\]

(60)

After straightforward calculation, one can obtain the following moments:

\[
\begin{align*}
E\left[ \varrho_{x_1y_1} \varrho_{x_2y_1}^2 \right] &= E\left[ \varrho_{x_1y_2} \varrho_{x_2y_2}^2 \right] \\
&= d^2 \mu^4 + 4d^2 \mu^2 \sigma^2 + 4d^2 \sigma^4 + 4d^2 \mu^2 \sigma^2 - d \sigma^4 + 4d \eta_3 \mu + d \eta_4,
\end{align*}
\]

53
and similarly
\[
E \left[ \varrho_{x_1y_1}^2 \varrho_{x_1y_2}^2 \right] = E \left[ \varrho_{x_2y_1}^2 \varrho_{x_2y_2}^2 \right] = d^2 \mu^4 + 4d^2 \mu^2 \sigma^2 + 4d^2 \sigma^4 + 4d^2 \mu^2 \sigma^2 - d \mu^4 - 4d \eta_2 \mu + d \eta_4.
\]

Based on these moments, the expected value of \( R_d^* \) is calculated by
\[
E[R_d^*] = 2 \left( \frac{2 \sigma^2 - \mu^2}{2 \sigma^2 + \mu^2} - \frac{1}{16d^2(\sigma^2 + \mu^2)^2} \right)
\times \left\{ 4d^2 \mu^4 + 16d^2 \mu^2 \sigma^2 + 16d^2 \sigma^4 + 16d \mu^2 \sigma^2 - 4d \sigma^4 + 4d \eta_4 \right\}.
\]

Then, under the condition in (19), one can show that
\[
\sqrt{d} \left( E[R_d^*] - 2 \right) = o_P(1) = \sqrt{d} \left( 2 \sigma^2 - \mu^2 \right) - \frac{\sqrt{d}}{16d^2(\sigma^2 + \mu^2)^2} \left\{ 16d^2 \mu^2 \sigma^2 + 16d^2 \sigma^4 \right\} + o(1)
\]
\[
= \frac{\sqrt{d}(2 \sigma^2 - \mu^2)}{2 \sigma^2 + \mu^2} - \frac{\sqrt{d}}{16d^2(\sigma^2 + \mu^2)^2} \left\{ 16d^2 \mu^2 \sigma^2 + 16d^2 \sigma^4 \right\} + o(1)
\]
\[
= -\frac{\sqrt{d} \mu^6 - \sqrt{d} \mu^4 \sigma^2}{(2 \sigma^2 + \mu^2)(\sigma^2 + \mu^2)^2} + o(1)
\]
\[
= o(1).
\]

Now, from Lemma C.9, the remainder term becomes negligible:
\[
\sqrt{d} (R_d^* - 2) = o_P(1).
\]

In the end, we have established the following approximation:
\[
2 \pi \sqrt{d} h(X_1, X_2; Y_1, Y_2)
\]
\[
= \frac{1}{\sqrt{3d}(2 \sigma^2 + \mu^2)} \left\{ h_{CQ}(X_1, X_2; Y_1, Y_2) + h_{CQ}(X_1, X_2; Y_2, Y_1) \right\} + o_P(1)
\]
\[
= \frac{1}{2\sqrt{3d} \sigma^2} \left\{ h_{CQ}(X_1, X_2; Y_1, Y_2) + h_{CQ}(X_1, X_2; Y_2, Y_1) \right\} + o_P(1).
\]

Combining this result with Theorem 2.1 in Chakraborty and Chaudhuri (2017), we finish the proof by obtaining
\[
\frac{2 \pi \sqrt{3d}}{\gamma_{m,n}} \left( U_{CV} - \frac{\mu^2}{2 \pi \sqrt{3d} \sigma^2} \right) = \frac{1}{\sqrt{d \sigma^2 \gamma_{m,n}}} \left( T_{CQ} - d \mu^2 \right) + o_P(1) \xrightarrow{d} N(0, 1).
\]

The next lemma shows that the variance of \( R_d^* \) is smaller order of \( d^{-1} \); thereby, the residual term in (61) is asymptotically negligible. We first upper-bound the variance of \( R_d^* \) by using the Efron-Stein inequality and then establish the result.

54
Lemma C.9. Consider the same model assumption in Theorem 5.1 with $\mu = o(1)$. Then
\[ \text{Var}[R_d^*] = o(d^{-1}), \]
where $R_d^*$ is given in (59).

Proof. First, define
\[ Z = R_d^* - \frac{2\sigma^2 - \mu^2}{2\sigma^2 + \mu^2} \]
\[ = f(X_{11}, \ldots, X_{1d}, X_{21}, \ldots, X_{2d}, Y_{11}, \ldots, Y_{1d}, Y_{21}, \ldots, Y_{2d}). \]

We will show the variance of $Z$ is of $o(d^{-1})$.

Based on the Efron-Stein inequality, the variance of $R_d^*$ is upper bounded by
\[ \text{Var}(Z) \leq \sum_{i=1}^{4d} E\left[\left(Z - E^{(i)}Z\right)^2\right] \]
\[ = 2dE\left[\left(Z - E^{(1)}Z\right)^2\right] + 2dE\left[\left(Z - E^{(2d+1)}Z\right)^2\right], \]
where $E^{(i)}[\cdot]$ is the conditional expectation conditioned on everything but $i$th element of $(X_{11}, \ldots, X_{1d}, X_{21}, \ldots, X_{2d}, Y_{11}, \ldots, Y_{1d}, Y_{21}, \ldots, Y_{2d})$. Without loss of generality, assume $E[X_{1i}] = E[X_{2i}] = 0$ and $E[Y_{1i}] = E[Y_{2i}] = \mu$.

For the rest of the proof, we will upper bound
\[ (I) \ 2dE\left[\left(Z - E^{(1)}Z\right)^2\right] \text{ and } (II) \ 2dE\left[\left(Z - E^{(2d+1)}Z\right)^2\right] \]
in sequence.

Bounding the first term $(I)$.
Let us begin with dealing with $Z - E^{(1)}Z$. We define
\[ \varphi_d = 16d^3(\sigma^2 + \mu^2)^2(2\sigma^2 + \mu^2) \]
so that
\[ \varphi_d^2 E\left[\left(Z - E^{(1)}Z\right)^2\right] = E\left[(I_{(1)} - I_{(2)})^2\right] = E\left[I_{(1)}^2 + I_{(2)}^2 - 2I_{(1)}I_{(2)}\right] \]
where
\[ I_{(1)} = 16d^2(\sigma^2 + \mu^2)^2\left\{X_{11}^2 - \sigma^2 - X_{11}Y_{11} - X_{11}Y_{21}\right\}, \]
\[ I_{(2)} = d(2\sigma^2 + \mu^2)\left\{\Delta_1 + \Delta_2 \left(\sum_{i=2}^{d}(Y_{2i} - X_{1i})^2 + \sum_{i=1}^{d}(Y_{1i} - X_{2i})^2\right) \right. \]
\[ \left. + \Delta_3 \left(\sum_{i=2}^{d}(Y_{1i} - X_{1i})^2 + \sum_{i=1}^{d}(Y_{2i} - X_{2i})^2\right) \right\}, \]
\[ \Delta_1 = Y_{11}^2 X_{11}^2 - 2Y_{11}^2 Y_{21} X_{11} + 2Y_{11} Y_{21}^2 X_{11} - 2Y_{11}^2 Y_{21}^2 X_{11} + 4Y_{11} Y_{21} X_{11}^2 - 2Y_{11} X_{11}^3 + Y_{21}^2 X_{11} - 2Y_{21} X_{11}^2 + X_{11}^4 - \sigma^2 Y_{11}^2 - \sigma^2 Y_{21}^2 - \eta_4, \]

\[ \Delta_2 = X_{11}^2 - 2Y_{11} X_{11} - \sigma^2, \]

\[ \Delta_3 = X_{11}^2 - 2Y_{21} X_{11} - \sigma^2. \]

Since the goal is to show

\[ 2d \mathbb{E} \left[ \left( Z - \mathbb{E}(1) Z \right)^2 \right] = o(d^{-1}), \]

and \( \vartheta_d^2 \) is order of \( d^6 \), it is enough to show that

\[ \mathbb{E} \left[ (\mathcal{I}_{(1)} - \mathcal{I}_{(2)})^2 \right] = o(d^4). \]

For this reason, we ignore the terms that are of smaller order than \( d^4 \). Based on the fact that \( \mu = o(1) \), the expected value of \( \mathcal{I}_{(1)}^2 \) is approximated by

\[ \mathbb{E} \left[ \mathcal{I}_{(1)}^2 \right] = (256\sigma^12 + 256\eta_4\sigma^8)d^4 + o(d^{-4}). \]

Similarly, one can have

\[ \mathbb{E} \left[ \mathcal{I}_{(2)}^2 \right] = (256\sigma^12 + 256\eta_4\sigma^8)d^4 + o(d^{-4}). \]

On the other hand,

\[ -2 \mathbb{E} \left[ \mathcal{I}_{(1)} \mathcal{I}_{(2)} \right] = -(512\sigma^12 + 512\eta_4\sigma^8)d^4 + o(d^{-4}), \]

therefore, we conclude

\[ 2d \mathbb{E} \left[ \left( Z - \mathbb{E}(1) Z \right)^2 \right] = o(d^{-1}). \] (62)

Bounding the second term (II).

Next, we will prove

\[ 2d \mathbb{E} \left[ \left( Z - \mathbb{E}(2d+1) Z \right)^2 \right] = o(d^{-1}). \]

Similar to the previous step, define

\[ \vartheta_d = 16d^3(\sigma^2 + \mu^2)^2(2\sigma^2 + \mu^2) \]

so that

\[ \vartheta_d^2 \mathbb{E} \left[ \left( Z - \mathbb{E}(2d+1) Z \right)^2 \right] = \mathbb{E} \left[ (\mathcal{J}_{(1)} - \mathcal{J}_{(2)})^2 \right] = \mathbb{E} \left[ \mathcal{J}_{(1)}^2 + \mathcal{J}_{(2)}^2 - 2\mathcal{J}_{(1)} \mathcal{J}_{(2)} \right], \]
where
\[
J_{(1)} = 16d^2(\sigma^2 + \mu^2)^2 \left\{ Y_{11}^2 - (\sigma^2 + \mu^2) - X_{11}Y_{11} + X_{11}\mu - X_{21}Y_{11} + X_{21}\mu \right\}
\]
\[
J_{(2)} = d(2\sigma^2 + \mu^2) \left\{ \Delta'_{1} + \Delta'_{2} \left( \sum_{i=2}^{d} (X_{2i} - Y_{1i})^2 + \sum_{i=1}^{d} (X_{1i} - Y_{2i})^2 \right) \right. \\
\left. + \Delta'_{3} \left( \sum_{i=2}^{d} (X_{1i} - Y_{1i})^2 + \sum_{i=1}^{d} (X_{2i} - Y_{2i})^2 \right) \right\},
\]
and
\[
\Delta'_{1} = X_{11}^2 Y_{11}^2 - 2X_{11}^2 X_{21} Y_{11} - 2X_{11}X_{21}^2 Y_{11} + 4X_{11}X_{21} Y_{11}^2 - 2X_{11}Y_{11}^3 \\
+ X_{21}^2 Y_{11}^2 - 2X_{21} Y_{11}^3 + Y_{11}^4 - \mu^4 - \mu^2 X_{11}^2 - \mu^2 X_{21}^2 - 6\mu^2 \sigma^2 \\
- 4\mu \eta_4 - \sigma^2 X_{11}^2 - \sigma^2 X_{21}^2 - \eta_4,
\]
\[
\Delta'_{2} = Y_{11}^2 - 2X_{11} Y_{11} + 2X_{11}\mu - \mu^2 - \sigma^2,
\]
\[
\Delta'_{3} = Y_{11}^2 - 2X_{21} Y_{11} + 2X_{21}\mu - \mu^2 - \sigma^2.
\]
Once again, after ignoring small order terms, we have
\[
E[ J_{(1)}^2 ] = (256\sigma^{12} + 256\eta_4 \sigma^8) d^4 + o(d^{-4})
\]
\[
E[ J_{(2)}^2 ] = (256\sigma^{12} + 256\eta_4 \sigma^8) d^4 + o(d^{-4})
\]
\[-2E[ J_{(1)} J_{(2)} ] = -(512\sigma^{12} + 512\eta_4 \sigma^8) d^4 + o(d^{-4}),
\]
and therefore conclude
\[
2dE \left[ \left( Z - E^{(2d+1)} Z \right)^2 \right] = o(d^{-1}). \quad (63)
\]
Combining (62) with (63),
\[
\text{Var} \left[ R_d^m \right] = \text{Var} \left[ Z \right] \leq 2dE \left[ \left( Z - E^{(1)} Z \right)^2 \right] + 2dE \left[ \left( Z - E^{(2d+1)} Z \right)^2 \right] = o(d^{-1}),
\]
which completes the proof. \hfill \Box

The following lemma is about the high-dimensional approximation of the energy statistic to the CQ statistic:

**Lemma C.10.** Suppose \( X \) and \( Y \) satisfy the assumptions in Theorem 5.1 with \( \mu_X - \mu_Y = \mu \). Then for every fixed \( m, n \) and \( d \to \infty \),
\[
\sqrt{\frac{\gamma_{m,n}^2}{\gamma_{m,n}}} \left( E_{m,n} - \sqrt{\frac{d \mu^2}{2 \sigma}} \right) = \frac{1}{\sqrt{\text{Var}^2 \gamma_{m,n}}} \left( T_{CQ} - d \mu^2 \right) + o_P(1) \overset{d}{\to} N(0, 1),
\]
where \( E_{m,n} \) is the energy statistic given in (10) and \( \gamma_{m,n} \) is given in (55).
Proof. Consider the symmetrized kernel of the energy statistic $E_{m,n}$:

\[
\chi^*_{\text{Energy}}(X_1, X_2; Y_1, Y_2)
\]

\[
= \frac{1}{2}\left\{||X_1 - Y_1|| + ||X_2 - Y_1|| + ||X_1 - Y_2|| + ||X_2 - Y_2||\right\}
\]

\[- ||X_1 - X_2|| - ||Y_1 - Y_2||.\]

Based on the CLT and Taylor expansion of $f(x) = \sqrt{x}$ at $x = d(2\sigma^2 + \mu^2)$,

\[
||X_1 - Y_1|| = \frac{1}{2}\sqrt{d(2\sigma^2 + \mu^2)} + \frac{1}{2\sqrt{d(2\sigma^2 + \mu^2)}}||X_1 - Y_1||^2 + o_P(1).
\]

Similarly,

\[
||X_1 - X_2|| = \frac{1}{2}\sqrt{2d\sigma^2} + \frac{1}{2\sqrt{2d\sigma^2}}||X_1 - X_2||^2 + o_P(1)
\]

\[
||Y_1 - Y_2|| = \frac{1}{2}\sqrt{2d\sigma^2} + \frac{1}{2\sqrt{2d\sigma^2}}||Y_1 - Y_2||^2 + o_P(1).
\]

Therefore,

\[
\chi^*_{\text{Energy}}(X_1, X_2; Y_1, Y_2)
\]

\[
= \frac{1}{4\sqrt{d(2\sigma^2 + \mu^2)}}\left\{||X_1 - Y_1||^2 + ||X_2 - Y_1||^2 + ||X_1 - Y_2||^2 + ||X_2 - Y_2||^2\right\}
\]

\[+ \sqrt{d(2\sigma^2 + \mu^2)} - 2d\sigma^2 - \frac{1}{2\sqrt{2d\sigma^2}}\left\{||X_1 - X_2||^2 + ||Y_1 - Y_2||^2\right\} + o_P(1)
\]

\[
= \frac{1}{2\sqrt{d(2\sigma^2 + \mu^2)}}\left\{h_{CQ}(X_1, X_2; Y_1, Y_2) + h_{CQ}(X_1, X_2; Y_2, Y_1)\right\}
\]

\[+ R_d + o_P(1),\]

where

\[
R_d = \sqrt{d(2\sigma^2 + \mu^2)} + \sqrt{2d\sigma^2} - \frac{1}{2\sqrt{d(2\sigma^2 + \mu^2)}} - \frac{1}{2\sqrt{2d\sigma^2}}
\]

\[\times \left\{||X_1 - X_2||^2 + ||Y_1 - Y_2||^2\right\}.\]

Note that

\[
\frac{1}{\sqrt{d}}\left\{||X_1 - X_2||^2 + ||Y_1 - Y_2||^2 - 4d\sigma^2\right\} = O_P(1),
\]

and thus the remainder term becomes negligible under the assumption in (19), i.e.

\[
R_d = \sqrt{d(2\sigma^2 + \mu^2)} - \sqrt{2d\sigma^2} + 4\sqrt{d\sigma^2} - \frac{1}{2\sqrt{2d\sigma^2}} - \frac{1}{2\sqrt{2d\sigma^2}} + o_P(1)
\]

58
\[
= \sqrt{d} \left( \frac{4\sigma^2 + \mu^2}{2\sigma^4 + \mu^2\sigma^2} \right) \frac{\sqrt{2\sigma^2 - 4\sigma^2} \left[ \sqrt{2\sigma^2} + \frac{\sqrt{3\sigma^2}}{2\sigma^2} \mu^2 + O(\mu^4) \right]}{2\sqrt{\sigma^2 - 4\sigma^2} \left[ \sqrt{2\sigma^2} + \frac{\sqrt{3\sigma^2}}{2\sigma^2} \mu^2 \right]} + o_P(1)
\]

Furthermore, it is clear to see that

\[
\frac{1}{2\sqrt{d}(2\sigma^2 + \mu^2)} \left\{ h_{CQ}(X_1, X_2; Y_1, Y_2) + h_{CQ}(X_1, X_2; Y_2, Y_1) \right\}
\]

Therefore,

\[
\chi_{\text{Energy}}^*(X_1, X_2; Y_1, Y_2)
\]

and after centering and scaling properly, we can obtain the result.

Finally, we collect the previous results and prove the main result in Theorem 5.1.

**Main Proof of Theorem 5.1**

Suppose $X$ and $Y$ satisfy the model assumptions in Theorem 5.1 with

\[
\mu = \frac{\psi}{d^{1/4}\sqrt{m+n}}.
\]

Then for every fixed $m, n$ as $d \to \infty$, followed by Lemma C.8, Lemma C.10 and Theorem 2.1 of Chakraborty and Chaudhuri (2017), we have

\[
\lim_{d \to \infty} \beta_{\text{CVM}}(\delta) = \lim_{d \to \infty} \beta_{\text{Energy}}(\delta) = \lim_{d \to \infty} \beta_{CQ}(\delta) = \lim_{d \to \infty} \beta_{\text{WMW}}(\delta),
\]

and they converge to

\[
\Phi \left( -z_\alpha + \frac{\psi^2}{\mu} \frac{1}{(m+n)\gamma_{m,n}\sigma^2} \right).
\]

Lastly, we finish the proof by noting that

\[
(m+n)\gamma_{m,n} \to \frac{\sqrt{2}}{\pi_1\pi_2} \text{ as } m, n \to \infty.
\]

**C.11 Proof of Lemma 6.1**

For given $w \in \mathbb{R}^d$, it is seen that

\[
\int_{S^{d-1}} \left| I(\beta^T z \leq \beta^T w) - I(\beta^T z' \leq \beta^T w) \right| d\lambda(\beta) \tag{64}
\]
\[
\begin{align*}
&= \int_{S_{d-1}} I(\beta^T z \leq \beta^T w < \beta^T z') + I(\beta^T z' \leq \beta^T w < \beta^T z) d\lambda(\beta) \\
&= \frac{1}{2} - \frac{1}{2\pi} \arccos \left\{ \frac{(z - w)^T (w - z')}{||z - w|| ||w - z'||} \right\} + \frac{1}{2} - \frac{1}{2\pi} \arccos \left\{ \frac{(z' - w)^T (w - z)}{||z' - w|| ||w - z||} \right\} \\
&= 1 - \frac{1}{\pi} \arccos \left\{ \frac{(z - w)^T (w - z')}{||z - w|| ||w - z'||} \right\} \\
&= \frac{1}{\pi} \left( \pi - \arccos \left\{ \frac{(z - w)^T (w - z')}{||z - w|| ||w - z'||} \right\} \right) \\
&= \frac{1}{\pi} \arccos \left\{ \frac{(z - w)^T (z' - w)}{||z - w|| ||z' - w||} \right\} := \rho_{\text{Angle}}(z, z'; w),
\end{align*}
\]

where (i) is due to \(\arccos(x) + \arccos(-x) = \pi\). Then \(\rho_{\text{Angle}}(z, z')\) is the expected value of \(\rho_{\text{Angle}}(z, z'; W_*)\) over \(W_* \sim \frac{1}{2} F_X + \frac{1}{2} G_Y\), i.e.
\[
\rho_{\text{Angle}}(z, z') = \mathbb{E}_{W_*} \left[ \rho_{\text{Angle}}(z, z'; W_*) \right] = \frac{1}{\pi} \mathbb{E}_{W_*} \left[ \arccos \left\{ \frac{(z - W_*^T (z' - W_*)}{||z - W_*|| ||z' - W_*||} \right\} \right].
\]

Now, if \(z = z'\), it is trivial to see \(\rho_{\text{Angle}}(z, z') = 0\). In addition, if \(\rho_{\text{Angle}}(z, z') = 0\), then we have \(z = z'\). In order to show the second direction, note that \(\arccos\) is a positive and monotone decreasing function and so \(\rho_{\text{Angle}}(z, z') = 0\) implies that
\[
\frac{(z - W_*)^T (z' - W_*)}{||z - W_*|| ||z' - W_*||} = 1,
\]
almost surely with respect to \((1/2) F_X + (1/2) G_Y\). By the Cauchy-Schwarz, the inner product becomes one if and only if \((z - W_*)\) or \((z' - W_*)\) is a multiple of the other. This is only possible when \(z - W_* = z' - W_*\) almost surely, which implies \(z = z'\). The symmetry property follows easily by the definition of \(\rho_{\text{Angle}}\). In addition, from triangle inequality, we have
\[
\begin{align*}
&\int_{S_{d-1}} \left| I(\beta^T z \leq \beta^T w) - I(\beta^T z' \leq \beta^T w) \right| d\lambda(\beta) \\
&\leq \int_{S_{d-1}} \left| I(\beta^T z \leq \beta^T w) - I(\beta^T z'' \leq \beta^T w) \right| d\lambda(\beta) \\
&\quad + \int_{S_{d-1}} \left| I(\beta^T z'' \leq \beta^T w) - I(\beta^T z' \leq \beta^T w) \right| d\lambda(\beta),
\end{align*}
\]
and therefore by the equality in (64), we can establish
\[
\rho_{\text{Angle}}(z, z'; w) \leq \rho_{\text{Angle}}(z, z''; w) + \rho_{\text{Angle}}(z', z''; w).
\]
Now, by taking the expectation over \(W_*\), we conclude that
\[
\rho_{\text{Angle}}(z, z') \leq \rho_{\text{Angle}}(z, z'') + \rho_{\text{Angle}}(z', z''),
\]

60
Next, we will show that for $\forall n \geq 2$, $z_1, \ldots, z_n \in S$, and $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, with $\sum_{i=1}^{n} \alpha_i = 0$, 
\[
\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \rho_{\text{Angle}}(z_i, z_j) \leq 0.
\]
The result follows from Section 6 of Bogomolny et al. (2007) who showed that for each fixed $w$, 
\[
\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \rho_{\text{Angle}}(z_i, z_j; w) \leq 0,
\] (65)
for any $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, with $\sum_{i=1}^{n} \alpha_i = 0$. Therefore, by taking the expected value over $W_*$ in (65), we conclude that $\rho_{\text{Angle}}$ is of negative-type.

Regarding Remark 6.1, note that 
\[
\mathbb{E}_\nu \left[ \rho_{\text{Angle}}(z, z'; \nu) \right] = \int_{S^{d-1}} \int_{\mathbb{R}} I(\beta^T z \leq \nu < \beta^T z') + I(\beta^T z' \leq \nu < \beta^T z) d\nu d\lambda(\beta)
\]
\[
\overset{(i)}{=} \int_{S^{d-1}} |\beta^T (z - z')| d\lambda(\beta)
\]
\[
\overset{(ii)}{=} \gamma_d ||z - z'||,
\]
where (i) and (ii) are due to Lemma 2.1 and Lemma 2.3 of Baringhaus and Franz (2004) and 
\[
\gamma_d = \frac{\sqrt{\pi} (d - 1) \Gamma (\frac{(d - 2)}{2})}{2 \Gamma(d/2)}.
\]
Therefore, the generalized angular distance with the Lebesgue measure corresponds to the Euclidean distance.

C.12 Proof of Proposition 6.1

From the definition of $\rho_{\text{Angle}}$, it is seen that 
\[
2 \mathbb{E}_{X,Y} \left[ \rho_{\text{Angle}}(X,Y) \right] - \mathbb{E}_{X,X'} \left[ \rho_{\text{Angle}}(X, X') \right] - \mathbb{E}_{Y,Y'} \left[ \rho_{\text{Angle}}(Y, Y') \right]
\]
\[
= \frac{1}{\pi} \mathbb{E} \left[ \Psi(X - X', Y - X') \right] + \frac{1}{\pi} \mathbb{E} \left[ \Psi(X - Y', Y - Y') \right] - \frac{1}{2\pi} \mathbb{E} \left[ \Psi(X - X'', X' - X'') \right]
\]
\[
- \frac{1}{2\pi} \mathbb{E} \left[ \Psi(X - Y, X' - Y) \right] - \frac{1}{2\pi} \mathbb{E} \left[ \Psi(Y - X', Y' - X') \right] - \frac{1}{2\pi} \mathbb{E} \left[ \Psi(Y - Y'', Y' - Y'') \right].
\]
Then the result follows by Lemma B.1.

C.13 Proof of Theorem 7.1

Given $\alpha \in S^{p-1}, \beta \in S^{q-1}$, expand the square term to have 
\[
\left\{ 4 \mathbb{P} \left( \alpha^T (X_1 - X_2) < 0, \beta^T (Y_1 - Y_2) < 0 \right) - 1 \right\}^2
\]
\[
16 \mathbb{E} \left[ I(\alpha^\top (X_1 - X_2) < 0, \alpha^\top (X_3 - X_4) < 0) \right. \\
\quad \times I(\beta^\top (Y_1 - Y_2) < 0, \beta^\top (Y_3 - Y_4) < 0) \left. \right] \\
- 8 \mathbb{E} \left[ I(\alpha^\top (X_1 - X_2) < 0) \times I(\beta^\top (Y_1 - Y_2) < 0) \right] + 1.
\]

By applying Lemma 2.2, the first term becomes
\[
\mathbb{E} \left[ \left( 2 - \frac{2}{\pi} \Psi(X_1 - X_2, X_3 - X_4) \right) \left( 2 - \frac{2}{\pi} \Psi(Y_1 - Y_2, Y_3 - Y_4) \right) \right]
\]
and the remainder terms become \(-1\), which yields the expression.

### C.14 Proof of Theorem 7.2

Given \(\alpha \in S^{p-1}\) and \(\beta \in S^{q-1}\),
\[
\int_{\mathbb{R}^2} \left[ F_{\alpha^\top X, \beta^\top Y}(u, v) - F_{\alpha^\top X}(u) F_{\beta^\top Y}(v) \right]^2 dF_{\alpha^\top X}(u) dF_{\beta^\top Y}(v)
\]

equals
\[
\mathbb{E} \left[ I(\alpha^\top (X_1 - X_3) \leq 0, \alpha^\top (X_2 - X_3) \leq 0) \right. \\
\quad \times I(\beta^\top (Y_1 - Y_4) \leq 0, \beta^\top (Y_2 - Y_4) \leq 0) \left. \right]
\]

\[
+ \mathbb{E} \left[ I(\alpha^\top (X_1 - X_5) \leq 0, \alpha^\top (X_2 - X_5) \leq 0) \right. \\
\quad \times I(\beta^\top (Y_3 - Y_6) \leq 0, \beta^\top (Y_4 - Y_6) \leq 0) \left. \right]
\]

\[
- 2 \mathbb{E} \left[ I(\alpha^\top (X_1 - X_4) \leq 0, \alpha^\top (X_2 - X_4) \leq 0) \right. \\
\quad \times I(\beta^\top (Y_1 - Y_5) \leq 0, \beta^\top (Y_3 - Y_5) \leq 0) \left. \right].
\]

Then apply Lemma 2.2 to obtain the expression.

### C.15 Proof of Lemma 7.1

The proof is based on the inclusion–exclusion principle. For the sets \(A, B, C\), the principle presents
\[
\left| A \cap B \cap C \right| = \left| A \cup B \cup C \right| - \left| A \right| - \left| B \right| - \left| C \right| + \left| A \cap B \right| + \left| A \cap C \right| + \left| B \cap C \right|
\] (66)

\[
= |S| - |A^c \cap B^c \cap C^c| - |A| - |B| - |C| + |A \cap B| + |A \cap C| + |B \cap C|,
\]

where the second equality is due to De Morgan’s laws with the entire set \(S\). For our purpose, define the sets \(A_\beta, B_\beta, C_\beta\) by
\[
A_\beta = \left\{ \beta^\top U_1 \leq 0 \right\}, \quad B_\beta = \left\{ \beta^\top U_2 \leq 0 \right\} \quad \text{and} \quad C_\beta = \left\{ \beta^\top U_3 \leq 0 \right\}.
\]
Let $\beta$ have the uniform distribution over $S^{d-1}$. Since $\beta$ and $-\beta$ have the same distribution, we have
\[
\int_{A_\beta \cap B_\beta \cap C_\beta} d\lambda(\beta) = \int_{A_\beta \cap B_\beta \cap C_\beta} d\lambda(\beta).
\]
Furthermore, since
\[
\int_{S_\beta} d\lambda(\beta) = 1, \quad \int_{A_\beta} d\lambda(\beta) = \int_{B_\beta} d\lambda(\beta) = \int_{B_\beta} d\lambda(\beta) = \frac{1}{2}
\]
and
\[
\int_{A_\beta \cap B_\beta} d\lambda(\beta) = \frac{1}{2} - \frac{1}{2\pi} \Psi(U_1, U_2),
\]
\[
\int_{A_\beta \cap C_\beta} d\lambda(\beta) = \frac{1}{2} - \frac{1}{2\pi} \Psi(U_1, U_3),
\]
\[
\int_{B_\beta \cap C_\beta} d\lambda(\beta) = \frac{1}{2} - \frac{1}{2\pi} \Psi(U_2, U_3),
\]
the identity of (66) presents
\[
\int_{S^{d-1}} I(\beta^\top U_1 \leq 0) I(\beta^\top U_2 \leq 0) I(\beta^\top U_3 \leq 0) d\mu(\beta)
\]
\[
= \frac{1}{2} - \frac{1}{4\pi} \left[ \Psi(U_1, U_2) + \Psi(U_1, U_3) + \Psi(U_2, U_3) \right],
\]
which completes the proof.

C.16 Proof of Theorem 7.3

From Bergsma and Dassios (2014), the univariate $\tau^*$ can be written as
\[
\tau^* = 4 \mathbb{P}(X_1, X_2 < X_3, X_4 \& Y_1, Y_2 < Y_3, Y_4)
\]
\[
+ 4 \mathbb{P}(X_1, X_2 < X_3, X_4 \& Y_1, Y_2 > Y_3, Y_4)
\]
\[
- 8 \mathbb{P}(X_1, X_2 < X_3, X_4 \& Y_1, Y_3 < Y_2, Y_4),
\]
where $Z_1, Z_2 < Z_3, Z_4$ denotes $\max\{Z_1, Z_2\} < \min\{Z_3, Z_4\}$.

Note that
\[
I(X_1, X_2 < X_3, X_4)
\]
\[
= I(X_1 < X_2 < X_3 < X_4) + I(X_2 < X_1 < X_3 < X_4)
\]
\[
+ I(X_1 < X_2 < X_4 < X_3) + I(X_2 < X_1 < X_4 < X_3)
\]
\[
= I(X_1 < X_2) I(X_2 < X_3) I(X_3 < X_4) + I(X_2 < X_1) I(X_1 < X_3) I(X_3 < X_4)
\]
\[
+ I(X_1 < X_2) I(X_2 < X_4) I(X_4 < X_3) + I(X_2 < X_1) I(X_1 < X_4) I(X_4 < X_3).
\]
Similarly, we have
\[
I(Y_1, Y_2 < Y_3, Y_4) = I(Y_1 < Y_2)I(Y_2 < Y_3)I(Y_3 < Y_4) + I(Y_2 < Y_1)I(Y_1 < Y_3)I(Y_3 < Y_4)
+ I(Y_1 < Y_2)I(Y_2 < Y_4)I(Y_4 < Y_3) + I(Y_2 < Y_1)I(Y_1 < Y_4)I(Y_4 < Y_3).
\]

Therefore, the product \(I(X_1, X_2 < X_3, X_4)I(Y_1, Y_2 < Y_3, Y_4)\) can be expressed as the linear combination of
\[
I(X_{i_1} < X_{i_2})I(X_{i_2} < X_{i_3})I(X_{i_3} < X_{i_4})I(Y_{j_1} < Y_{j_2})I(Y_{j_2} < Y_{j_3})I(Y_{j_3} < Y_{j_4}).
\]

Using Lemma 7.1,
\[
\int_{S^{p-1}} I(\alpha^T X_{i_1} < \alpha^T X_{i_2})I(\alpha^T X_{i_2} < \alpha^T X_{i_3})I(\alpha^T X_{i_3} < \alpha^T X_{i_4})d\lambda(\alpha)
= \frac{1}{2} - \frac{1}{4\pi} [\Psi(U_1, U_2) + \Psi(U_1, U_3) + \Psi(U_2, U_3)],
\]
where \(U_1 = X_{i_1} - X_{i_2}, U_2 = X_{i_2} - X_{i_3}\) and \(U_3 = X_{i_3} - X_{i_4}\).

Similarly,
\[
\int_{S^{q-1}} I(\beta^T Y_{j_1} < \beta^T Y_{j_2})I(\beta^T Y_{j_2} < \beta^T Y_{j_3})I(\beta^T Y_{j_3} < \beta^T Y_{j_4})d\lambda(\beta)
= \frac{1}{2} - \frac{1}{4\pi} [\Psi(V_1, V_2) + \Psi(V_1, V_3) + \Psi(V_2, V_3)],
\]
where \(V_1 = Y_{j_1} - Y_{j_2}, V_2 = Y_{j_2} - Y_{j_3}\) and \(V_3 = Y_{j_3} - Y_{j_4}\).

As a result, we have
\[
\int_{S^{p-1}} \int_{S^{q-1}} \mathbb{P}(\alpha^T X_1, \alpha^T X_2 < \alpha^T X_3, \alpha^T X_4 \& \beta^T Y_1, \beta^T Y_2 < \beta^T Y_3, \beta^T Y_4)d\lambda(\alpha)d\lambda(\beta)
= \mathbb{E}[h_p(X_1, X_2, X_3, X_4)h_q(Y_1, Y_2, Y_3, Y_4)].
\]

Applying the same argument to the rest, we can obtain
\[
\tau_{p,q} = \mathbb{E}[h_p(X_1, X_2, X_3, X_4)h_q(Y_1, Y_2, Y_3, Y_4)]
+ \mathbb{E}[h_p(X_1, X_2, X_3, X_4)h_q(Y_3, Y_4, Y_1, Y_2)]
- 2\mathbb{E}[h_p(X_1, X_2, X_3, X_4)h_q(Y_1, Y_3, Y_2, Y_4)],
\]
which finishes the proof.

**C.17 Proof of Theorem A.1**

Let us write
\[
U_{m,n}^* (Z_{m,n}) := U_{m,n}^* (Z_1, \ldots, Z_N)
\]

64
\begin{align*}
N\{U_{m,n}(Z_1, \ldots, Z_N) - \mathbb{E}[U_{m,n}(Z_1, \ldots, Z_N)]\}
and denote \( U_{m,n}^*(Z_{\pi(1)}, \ldots, Z_{\pi(N)}) \) by \( \tilde{U}_{m,n}(Z_\pi) \). Our goal is to show for two independent random permutations \( \pi, \pi' \) that
\begin{align}
(U_{m,n}^*(Z_\pi), \ U_{m,n}^*(Z_{\pi'})) & \xrightarrow{d} (T, T'),
\end{align}
where \( T, T' \) are independent and identically distributed with the unconditional limiting null distribution \( R(t) \). Then the desired result follows by Lemma B.4. The proof consists of several pieces and closely follows the proof of the limiting distribution of a two-sample degenerating \( U \)-statistic in Chapter 3 of Bhat (1995).

We start with the projection of the two-sample \( U \)-statistic via Hoffding’s decomposition. Consider the projection of the two-sample degenerate \( U \)-statistic based on \( Z_{m,n} \):
\begin{align*}
\hat{U}_{m,n}(Z_{m,n}) &= \frac{r(r-1)}{m(m-1)} \sum_{1 \leq i_1 < i_2 \leq m} g_{2,0}^*(Z_{i_1}, Z_{i_2}) + \frac{r(r-1)}{n(n-1)} \sum_{1 \leq j_1 < j_2 \leq n} g_{0,2}^*(Z_{j_1+m}, Z_{j_2+m}) \\
&\quad + \frac{r^2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} g_{1,1}^*(Z_{i}, Z_{j+m}).
\end{align*}
Then it can be seen that
\begin{align*}
\mathbb{E} \left[ (U_{m,n}(Z_{m,n}) - \hat{U}_{m,n}(Z_{m,n})) \right] = 0 \text{ and } \text{Var}(U_{m,n}(Z_{m,n}) - \hat{U}_{m,n}(Z_{m,n}) = O(N^{-3}),
\end{align*}
which implies
\begin{align}
N(U_{m,n}(Z_{m,n}) - \theta) = N(\hat{U}_{m,n}(Z_{m,n}) - \theta) + o_p(1).
\end{align}
Under the finite second moment of the kernel \( h \), we may have the following decompositions:
\begin{align*}
g_{2,0}^*(x, y) &= \sum_{i=1}^{\infty} \lambda_i \phi_i(x)\phi_i(y), \\
g_{0,2}^*(x, y) &= \sum_{i=1}^{\infty} \gamma_i \psi_i(x)\psi_i(y), \\
g_{1,1}^*(x, y) &= \sum_{i=1}^{\infty} \alpha_i \phi_i^*(x)\psi_i^*(y),
\end{align*}
where \( \{\phi_i(\cdot)\}, \{\psi_i(\cdot)\}, \{\phi^*(\cdot), \psi^*(\cdot)\} \) are orthonormal eigenfunctions and the corresponding eigenvalues \( \{\lambda_i\}, \{\gamma_i\}, \{\alpha_i\} \), associated with \( g_{2,0}^*, g_{0,2}^* \) and \( g_{1,1}^* \), respectively (see e.g., Bhat, 1995, for details). From the given conditions of the theorem, the eigenvalues and the eigenfunctions are related as follows:
\begin{align*}
\phi_i(z) &= \psi_i(z) = \phi_i^*(z) = \psi_i^*(z), \\
\gamma_i &= \lambda_i \text{ and } \alpha_i = \frac{1-r}{r} \lambda_i.
\end{align*}
Therefore,

\[ N\hat{U}_{m,n}(Z_{m,n}) = \hat{a}_1 \left[ \frac{1}{m} \sum_{1 \leq i_1 \neq i_2 \leq m} \sum_{i=1}^{\infty} \lambda_i \phi_i(Z_{i1}) \phi_i(Z_{i2}) \right] + \hat{a}_2 \left[ \frac{1}{n} \sum_{1 \leq j_1 \neq j_2 \leq n} \sum_{j=1}^{\infty} \lambda_j \phi_j(Z_{j1+m}) \phi_j(Z_{j2+m}) \right] + \hat{a}_3 \left[ \frac{1}{\sqrt{mn}} \sum_{i_1=1}^{m} \sum_{j_1=1}^{n} \sum_{k=1}^{\infty} \lambda_k \phi_k(Z_{i1}) \phi_k(Z_{j1+m}) \right] = \hat{a}_1 T_m + \hat{a}_2 T'_n + \hat{a}_3 T''_{mn}, \]

where

\[ \hat{a}_1 = \frac{r(r-1)}{2} \frac{N}{m-1}, \quad \hat{a}_2 = \frac{r(r-1)}{2} \frac{N}{n-1} \quad \text{and} \quad \hat{a}_3 = -r(r-1) \frac{N}{\sqrt{mn}}. \]

Denote the centered and scaled projection of the \(U\)-statistic by \(\bar{U}_{m,n} := N(\bar{U}_{m,n}(Z_\pi) - \theta)\) and \(\bar{U}'_n := N(\bar{U}_{m,n}(Z_{\pi'}) - \theta).\)

Then due to (68),

\[ (U^*_{m,n}(Z_\pi), U^*_{m,n}(Z_{\pi'})) = (\bar{U}_{m,n}(Z_\pi), \bar{U}'_n(Z_{\pi'})) + o_P(1). \]

Therefore it suffices to show

\[ (\bar{U}_{m,n}, \bar{U}'_{m,n}) \overset{d}{\rightarrow} (T, T') \]

to complete the main proof. Having this goal in mind, we start with a truncation of the degenerating \(U\)-statistic.

**Truncation of the \(U\)-statistics.**

Now, define a truncated version of \(N(\bar{U}_{m,n}(Z_{m,n}) - \theta)\) by

\[ N(\bar{U}_{m,n,K}(Z_{m,n}) - \theta) = \hat{a}_1 \left[ \frac{1}{m} \sum_{1 \leq i_1 \neq i_2 \leq m} \sum_{i=1}^{K} \lambda_i \phi_i(Z_{i1}) \phi_i(Z_{i2}) \right] + \hat{a}_2 \left[ \frac{1}{n} \sum_{1 \leq j_1 \neq j_2 \leq n} \sum_{j=1}^{K} \lambda_j \phi_j(Z_{j1+m}) \phi_j(Z_{j2+m}) \right] + \hat{a}_3 \left[ \frac{1}{\sqrt{mn}} \sum_{i_1=1}^{m} \sum_{j_1=1}^{n} \sum_{k=1}^{K} \lambda_k \phi_k(Z_{i1}) \phi_k(Z_{j1+m}) \right] = \hat{a}_1 T_{mK} + \hat{a}_2 T'_{nK} + \hat{a}_3 T''_{mnK}. \]
Write
\[
\hat{a}_1T_{nK} + \hat{a}_2T_{nK}' + \hat{a}_3T_{mnK}' = \hat{a}_1\left[ \sum_{k=1}^{K} \lambda_k \left( W_{km}^2 - V_{km} \right) \right] + \hat{a}_2\left[ \sum_{k=1}^{K} \lambda_k \left( W_{kn}'^2 - V_{kn}' \right) \right] + \hat{a}_3\left[ \sum_{k=1}^{K} \lambda_k W_{km} W_{kn}' \right]
\]

\[
= \frac{r(r-1)}{2} \left\{ \sum_{k=1}^{K} \lambda_k \left( \sqrt{\frac{N}{m}} W_{km} - \sqrt{\frac{N}{n}} W_{kn}' \right)^2 - \sum_{k=1}^{K} \lambda_k \left( \frac{N}{m} V_{km} + \frac{N}{n} V_{kn}' \right) \right\},
\]

where
\[
W_{km} = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \phi_k(Z_i), \quad W_{kn}' = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \phi_k(Z_{j+m}),
\]

\[
V_{km} = \frac{1}{m} \sum_{i=1}^{m} \phi_k^2(Z_i), \quad V_{kn}' = \frac{1}{n} \sum_{j=1}^{n} \phi_k^2(Z_{j+m}),
\]

for \(k = 1, \ldots, K\).

By strong law of large numbers,
\[
V_{mn}^\top := (V_{1m}, \ldots, V_{Km}, V_{1n}', \ldots, V_{Kn}')^\top \quad \text{a.s.} \quad V^\top := (V_1, \ldots, V_K, V_1', \ldots, V_K')^\top
\]

and by the assumption that \(m/N \to \pi_1, n/N \to \pi_2\),
\[
N(\hat{U}_{m,n,K} - \theta) = \frac{r(r-1)}{2} \left\{ \sum_{k=1}^{K} \lambda_k \left( \sqrt{\frac{N}{m}} W_{km} - \frac{r(r-1)}{2} \sqrt{\frac{N}{n}} W_{kn}' \right)^2 - \frac{1}{\pi_1 \pi_2} \sum_{k=1}^{K} \lambda_k \right\} + o_P(1)
\]

\[
= \frac{r(r-1)}{2} \left\{ N \sum_{k=1}^{K} \lambda_k \left( \frac{1}{m} \sum_{i=1}^{m} \phi_k(Z_i) - \frac{1}{n} \sum_{j=1}^{n} \phi_k(Z_{j+m}) \right)^2 - \frac{1}{\pi_1 \pi_2} \sum_{k=1}^{K} \lambda_k \right\} + o_P(1)
\]

\[
= \frac{r(r-1)}{2} \left\{ N \sum_{k=1}^{K} \lambda_k \left( \sum_{i=1}^{N} \epsilon_i \phi_k(Z_i) \right)^2 - \frac{1}{\pi_1 \pi_2} \sum_{k=1}^{K} \lambda_k \right\} + o_P(1)
\]

where
\[
(\epsilon_1, \ldots, \epsilon_m, \epsilon_{m+1}, \ldots, \epsilon_{m+n}) = \left( \frac{m^{-1}, \ldots, m^{-1}, -n^{-1}, \ldots, -n^{-1}}{m \text{ terms}} \right).
\]

**Proving independence of the truncated U-statistics.**

Consider the truncated permutation statistics
\[
\tilde{U}_{m,n,K} := N(\hat{U}_{m,n,K}(Z_\pi) - \theta)
\]
where $T$ apply the continuous mapping theorem together with Slutsky’s theorem to have

$$
\tilde{U}_{nK} := N(\hat{U}_{m,n,K}(Z_{\pi'}) - \theta)
$$

and thus the components of the vector are asymptotically independent to each other. Then

$$
= \frac{r(r-1)}{2} \left\{ N \sum_{k=1}^{K} \lambda_k \left( \sum_{i=1}^{N} \epsilon_{\pi(i)} \phi_k(Z_i) \right)^2 - \frac{1}{\pi_1 \pi_2} \sum_{k=1}^{K} \lambda_k \right\} + o_p(1)
$$

Note that $\epsilon_{\pi(i)}$ and $\epsilon_{\pi'(i)}$ are independent random variables by the assumption having either $1/m$ or $-1/n$ with $m/N$ and $n/N$ probabilities; hence

$$
\text{Cov} (\epsilon_{\pi(i)} \phi_k(Z_i), \epsilon_{\pi'(i)} \phi_k(Z_i)) = \mathbb{E} [\epsilon_{\pi(i)}] \mathbb{E} [\epsilon_{\pi'(i)}] \mathbb{E} [\phi_k^2(Z_i)] = 0.
$$

By the Cramér-Wold device and the Lindeberg condition, we see that

$$
\sqrt{N} \left( \sum_{i=1}^{N} \epsilon_{\pi(i)} \phi_1(Z_i), \ldots, \sum_{i=1}^{N} \epsilon_{\pi(i)} \phi_K(Z_i) \right) \xrightarrow{d} N(0, \pi_1^{-1} \pi_2^{-1} I_{2K})
$$

and thus the components of the vector are asymptotically independent to each other. Then apply the continuous mapping theorem together with Slutsky’s theorem to have

$$
(\tilde{U}_{m,n,K}, \tilde{U}'_{m,n,K}) \xrightarrow{d} (T_K, T'_K)
$$

where $T_K$ and $T'_K$ are independent and have the same distribution as

$$
\frac{r(r-1)}{2\pi_1 \pi_2} \sum_{k=1}^{K} \lambda_k (W_k^2 - 1),
$$

where $W_k \overset{i.i.d.}{\sim} N(0,1)$.

**Bounding the difference between characteristic functions.**

We will use the characteristic functions to show

$$
(\tilde{U}_{m,n}, \tilde{U}'_{m,n}) \xrightarrow{d} (T, T').
$$

More specifically, we will show that for any $x, y \in \mathbb{R}$ and any $\epsilon > 0$ and sufficiently large $N$,

$$
\left| \mathbb{E} \left[ e^{i(x\tilde{U}_{m,n} + y\tilde{U}'_{m,n})} \right] - \mathbb{E} \left[ e^{i(xT + yT')} \right] \right| \leq (I) + (II) + (III) < \epsilon
$$

where

$$
(I) = \left| \mathbb{E} \left[ e^{i(x\tilde{U}_{m,n} + y\tilde{U}'_{m,n})} \right] - \mathbb{E} \left[ e^{i(x\tilde{U}_{m,n,K} + y\tilde{U}'_{m,n,K})} \right] \right|,
$$

$$
(II) = \left| \mathbb{E} \left[ e^{i(x\tilde{U}_{m,n,K} + y\tilde{U}'_{m,n,K})} \right] - \mathbb{E} \left[ e^{i(xT_K + yT'_K)} \right] \right|,
$$

$$
(III) = \left| \mathbb{E} \left[ e^{i(xT_K + yT'_K)} \right] - \mathbb{E} \left[ e^{i(xT + yT')} \right] \right|.
$$

68
We bound these terms in sequence.

1. **Bounding (I).**

Based on $|e^{iz}| = 1$ and $|e^{iz} - 1| \leq |z|$, we bound (I) by

\[
(I) = \left| \mathbb{E} \left[ e^{i(x\tilde{U}_{m,n} + y\tilde{U}_{m,n})} \right] - \mathbb{E} \left[ e^{i(x\tilde{U}_{m,n,K} + y\tilde{U}_{m,n,K})} \right] \right|
\]

\[
\leq |x| \left[ \mathbb{E} \left( \tilde{U}_{m,n,K} - \tilde{U}_{m,n} \right)^2 \right]^{1/2} + |y| \left[ \mathbb{E} \left( \tilde{U}_{m,n,K} - \tilde{U}_{m,n} \right)^2 \right]^{1/2}
\]

\[
\leq (|x| + |y|) \left\{ \frac{r(r - 1)}{2\hat{\pi}_1} \left( 2 \sum_{k=K+1}^{\infty} \lambda_k^2 \right)^{1/2} + \frac{r(r - 1)}{2\hat{\pi}_2} \left( 2 \sum_{k=K+1}^{\infty} \lambda_k^2 \right)^{1/2} \right\}
\]

\[
= (|x| + |y|) \frac{r(r - 1)}{\sqrt{2}} \left( \frac{1}{\sqrt{\hat{\pi}_1}} - \frac{1}{\sqrt{\hat{\pi}_2}} \right)^2 \left( \sum_{k=K+1}^{\infty} \lambda_k^2 \right)^{1/2}
\]

\[
\leq (|x| + |y|) \frac{r(r - 1)}{2\hat{\pi}_1 \hat{\pi}_2} \left( \sum_{k=K+1}^{\infty} \lambda_k^2 \right)^{1/2}
\]

where $\hat{\pi}_1 = m/N$ and $\hat{\pi}_2 = n/N$.

Now, for fixed $x$ and $y$ and any given $\epsilon > 0$, we choose $K$ large enough to bound

\[
(|x| + |y|) \frac{r(r - 1)}{\sqrt{2}} \left( \frac{1}{\sqrt{\hat{\pi}_1}} - \frac{1}{\sqrt{\hat{\pi}_2}} \right)^2 \left( \sum_{k=K+1}^{\infty} \lambda_k^2 \right)^{1/2} < \frac{\epsilon}{3}, \quad (70)
\]

Since $\hat{\pi}_1 \to \pi_1$ and $\hat{\pi}_2 \to \pi_2$ as $N \to \infty$, we have

\[
(I) \leq (|x| + |y|) \frac{r(r - 1)}{2\hat{\pi}_1 \hat{\pi}_2} \left( \sum_{k=K+1}^{\infty} \lambda_k^2 \right)^{1/2} < \frac{\epsilon}{3},
\]

for all sufficiently large $N$.

2. **Bounding (II).**

From the result established in (69), we have

\[
(II) = \left| \mathbb{E} \left[ e^{i(x\tilde{U}_{m,n,K} + y\tilde{U}_{m,n,K}')} \right] - \mathbb{E} \left[ e^{i(xT_K + yT_K')} \right] \right| < \frac{\epsilon}{3} \quad \text{for all sufficiently large } N.
\]

3. **Bounding (III).**

From Chapter 3 of Bhat (1995) with the given conditions on the kernel, the asymptotic
distribution of a degenerate $U$-statistic converges to

$$N \left( U_{m,n} - \theta \right) \overset{d}{\to} \frac{r(r-1)}{2\pi_1} \sum_{k=1}^{\infty} \lambda_k (W^2_k - 1) + \frac{r(r-1)}{2\pi_2} \sum_{k=1}^{\infty} \lambda_k (W'^2_k - 1)$$

(71)

$$- \frac{r(r-1)}{\sqrt{\pi_1 \pi_2}} \sum_{k=1}^{\infty} \lambda_k W_k W'_k$$

where $\{W_k\}$ and $\{W'_k\}$ are independent standard normal random variables and $\{\lambda_k\}$ are eigenvalues associated with the kernel. Note that the right-side of (71) can be re-written as

$$\frac{r(r-1)}{2\pi_1 \pi_2} \sum_{k=1}^{\infty} \lambda_k \left( (\sqrt{\pi_2} W_k - \sqrt{\pi_1} W'_k)^2 - 1 \right),$$

where $\sqrt{\pi_2} W_k - \sqrt{\pi_1} W'_k \sim N(0, 1)$. Therefore, $T, T'$ are identically distributed as

$$\frac{r(r-1)}{2\pi_1 \pi_2} \sum_{k=1}^{\infty} \lambda_k (W^2_k - 1).$$

Recall that $T_K, T'_K$ have the same distribution as

$$\frac{r(r-1)}{2\pi_1 \pi_2} \sum_{k=1}^{K} \lambda_k (W^2_k - 1).$$

Consequently,

$$\left| \mathbb{E} \left[ e^{i(x T_K + y T'_K)} \right] - \mathbb{E} \left[ e^{i(x T + y T')} \right] \right| \leq |x| \left[ \mathbb{E} \left( (T_K - T)^2 \right)^{1/2} + |y| \left[ \mathbb{E} \left( (T'_K - T')^2 \right)^{1/2} \right] \right.$$  

$$\leq (|x| + |y|) \frac{r(r-1)}{\sqrt{2\pi_1 \pi_2}} \left( \sum_{k=K+1}^{\infty} \lambda^2_k \right)^{1/2} < \frac{\epsilon}{3},$$

with the same choice of $x, y, \epsilon, K$ in (70).

• Combining the bounds.

From the previous results, we conclude that for any $x, y \in \mathbb{R}$ and any $\epsilon > 0$ with sufficiently large $N$,

$$\left| \mathbb{E} \left[ e^{i(x \tilde{U}_{m,n} + y \tilde{U}'_{m,n})} \right] - \mathbb{E} \left[ e^{i(x T + y T')} \right] \right| < \epsilon,$$

and therefore

$$\left( \tilde{U}_{m,n}, \tilde{U}'_{m,n} \right) \overset{d}{\to} (T, T').$$

This completes the proof.
D Additional Simulations

In this section, we provide additional simulation results to support our claim in Section 5. The simulation setting is the same as that in the first part of Section 8, but now the dimension is set at $d = 1,000$. We generate samples from $X,Y \sim N(\mu^{(0)}, \Sigma)$ under $H_0$, and $X \sim N(\mu^{(0)}, \Sigma)$ and $Y \sim N(\mu^{(1)}, \Sigma)$ under $H_1$. Here, we provide scatter plots between the considered statistics including the CvM, Energy and CQ statistics to demonstrate their asymptotic identities. The results are presented in Figure 3 and 4. From the results, it is seen that pairwise observations of (i) the CvM statistic vs. the CQ statistic or (ii) the Energy statistic vs. the CQ statistic approximately lie on a linear line, meaning that after centering and scaling, the two statistics are identical with small stochastic errors. These results confirm Lemma C.8 and Lemma C.10, and further illustrate that the resulting tests have approximately the same power.
Figure 3: The scatter plots between the CvM statistic $U_{m,n}$ and the CQ statistic $T_{CQ}$ when $m = n = 20$ and $d = 1,000$. The red lines are based on the least square estimates.
Figure 4: The scatter plots between the Energy statistic $E_{m,n}$ and the CQ statistic $T_{CQ}$ when $m = n = 20$ and $d = 1,000$. The red lines are based on the least square estimates.