On a Projective Ensemble Approach to Two Sample Test for Equality of Distributions

Zhimei Li [1] **Yaowu Zhang** [1]

Abstract

In this work, we propose a robust test for the multivariate two-sample problem through projective ensemble, which is a generalization of the Cramér-von Mises statistic. The proposed test statistic has a simple closed-form expression without any tuning parameters involved, it is easy to implement and can be computed in quadratic time. Moreover, our test is insensitive to the dimension and consistent against all fixed alternatives, it does not require the moment assumption and is robust to the presence of outliers. We study the asymptotic behaviors of the test statistic under the null and two kinds of alternative hypotheses. We also suggest a permutation procedure to approximate critical values and establish its consistency. We demonstrate the effectiveness of our test through extensive simulation studies and a real data application.

1. Introduction

We study the problem of testing for homogeneity of two random samples, i.e., testing whether two samples come from the same population, which is one of the most fundamental problems in statistics and has applications in a wide range of areas (see, e.g., Lehmann & Romano, 2006). Specifically, suppose $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}} \in \mathbb{R}^p$ and $\mathbf{y} = (Y_1, \ldots, Y_p)^{\mathrm{T}} \in \mathbb{R}^p$ are two p-dimensional random vectors whose distribution functions are F and G, respectively. $\{\mathbf{x}_i, i = 1, \ldots, m\}$ and $\{\mathbf{y}_i, i = 1, \ldots, n\}$ are two mutually independent random samples drawn from F and G, with sample size m and n, respectively. The problem of testing for whether $\mathbf{x}$ and $\mathbf{y}$ are homogeneous amounts to testing for the equality of two distributions, i.e.,

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

In the past decades, the problem of distribution testing has received much attention in the literature. Under the normality assumption, it suffices to compare the differences in mean vectors, or covariance matrices, or both. For example, the Student's t test and Hotelling's T^2 test are the most well-known tests in this circumstance. Other examples include Bai & Saranadasa (1996); Li & Chen (2012); Cai et al. (2014); Cai & Liu (2016), etc. However, it is commonly known that the first two moments are not sufficient to characterize the distribution and these kinds of methods may be inconsistent when the normality assumption violates.

To overcome this issue, several nonparametric approaches have been developed in the literature. It is natural to use a measure of difference between F_m and G_n as the test statistic, where F_m and G_n are the empirical distribution functions of F and G, respectively. For example, the Kolmogorov-Smirnov test statistic (Smirnov, 1939) is given by $\sqrt{nm/(n+m)}\sup_{t\in\mathbb{R}}|F_m(t) - G_n(t)|$, Cramér-von Mises (CvM) test statistic (Anderson, 1962) and Anderson-Darling statistic (Darling, 1957) can both be written in the following formula,

$$\frac{mn}{m+n}\int_{-\infty}^{\infty}\{F_m(t) - G_n(t)\}^2\,\omega\{H_{m+n}(t)\}\,dH_{m,n}(t),$$

with different choices of the weight function $\omega(\cdot)$, where H_{m+n} is the pooled empirical distribution function. Although these methods have several advantages when $p = 1$, i.e., consistent against any fixed alternatives, distribution free under the null, no moment conditions are required, and free of tuning parameters, it may be difficult to generalize them to multivariate cases (Kim et al., 2020). For example, Darling (1957) generalized the Kolmogorov-Smirnov test by directly using the multivariate empirical distribution function, which suffers from significant power loss when p increases. In the multivariate case, i.e., $p \geq 2$, there is a number of literature about two-sample testing procedures. We roughly classify them into two classes. The first category is graph-based tests. For example, Friedman & Rafsky (1979) constructed k minimum spanning tree graphs

[1]Research Institute for Interdisciplinary Sciences, School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai, China. Correspondence to: Yaowu Zhang <zhang.yaowu@mail.shufe.edu.cn>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

while Henze (1988) used k nearest neighbor graphs. Chen & Friedman (2017) and Chen et al. (2018) improve the test proposed by Henze (1988) to enhance the power performance under scale alternatives and imbalanced samples, respectively. Bhattacharya (2019) established a general framework for graph-based two-sample tests. However, these graph based tests are either inconsistent or rely on selecting tuning parameters delicately. The other class can be described as using reproducing kernel Hilbert space (RKHS) embeddings of probability distributions. For example, Gretton et al. (2012) proposed a class of maximum mean discrepancy (MMD) test statistic based on a reproducing kernel Hilbert approach. Székely & Rizzo (2004) introduced a class of energy statistic for testing equal distributions, and is shown to be a special case of the MMD (Sejdinovic et al., 2013). One may refer to Harchaoui et al. (2013) for a comprehensive review.

More recently, Kim et al. (2020) generalized the univariate CvM statistic to arbitrary dimension through the novel projection-averaging. Specifically, the proposed test statistic is the sample version of the squared multivariate CvM distance defined as:

$$\iint \{F_{\boldsymbol{\beta}}(t) - G_{\boldsymbol{\beta}}(t)\}^2 dH_{\boldsymbol{\beta}}(t) d\lambda(\boldsymbol{\beta}), \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$, $F_{\boldsymbol{\beta}}(t) = \mathrm{P}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} \leq t)$, and $G_{\boldsymbol{\beta}}(t) = \mathrm{P}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{y} \leq t)$, $H_{\boldsymbol{\beta}}(t)$ is chosen to be $\tau F_{\boldsymbol{\beta}}(t) + (1-\tau)G_{\boldsymbol{\beta}}(t)$, τ is the limit value of $m/(m+n)$ as $\min(m,n) \to \infty$, and $\lambda(\boldsymbol{\beta})$ is the uniform probability measure on the p-dimensional unit sphere $\mathcal{S}^{p-1} \overset{\text{def}}{=} \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\| = 1\}$. By choosing $H_{\boldsymbol{\beta}}(t) = t$, the corresponding test statistic is shown to be proportional to the energy statistic (Baringhaus & Franz, 2004). Given the advantages of the Cramér-von Mises test statistic when $p = 1$, it is supposed that these two kinds of generalizations have similar properties. For example, they are both nonnegative and equal to zero if and only if $F = G$, the integrations have a simple closed-form expression and the resulting test statistic can be easily calculated and are both free of tuning parameters. Furthermore, the projection-averaging approach proposed by Kim et al. (2020) is also robust to heavy-tailed distributions or outliers, while the energy distance is only well-defined under the moment condition (finite first moment). That is, when potential outliers exist, the energy statistic becomes unstable and the power performance of the resulting test might become very poor, while the projection-averaging statistic maintains good power. However, as a side-effect, the projection-averaging approach incorporated higher computational costs, i.e., cubic computations, which greatly diminishes the computational efficiency, especially when the sample size is large. Moreover, they focused on the case that $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ and $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{y}$ have continuous distribution functions for all $\boldsymbol{\beta} \in \mathcal{S}^{p-1}$, whereas we are targeting on a more general case and we do not need such continuous distribution assumption.

In this paper, we apply the idea of projections and develop a new projective ensemble approach for testing equality of distributions. Similar to Baringhaus & Franz (2004) and Kim et al. (2020), our proposal generalizes the Cramér-von Mises test through projection ensemble. We show that, with different choices of the ensemble approaches, i.e., $H_{\boldsymbol{\beta}}(t)$ and $\lambda(\boldsymbol{\beta})$ in (1), the integration also has a simple closed-form expression. Besides, the corresponding test statistic shares all the aforementioned appealing properties. Specifically, the explicit closed-form expression does not require any tuning parameters, making our proposed test easy to implement and insensitive to the dimension. It is robust to the presence of potential extreme values or heavily tailed observations because it does not require the moment condition. It is also a member of the MMD, nonnegative and equal to zero if and only if $F = G$, which guarantees that our test is consistent against all fixed alternatives. Moreover, our test statistic has a very simple closed-form expression and can be computed in quadratic time without the continuity assumption.

The rest of this paper is organized as follows. We give the motivation of the projective ensemble based test, suggest an index to measure the departure from the equality of distributions, and provide several appealing properties of the index at the population level in section 2.1. We use the sample version of the index as the test statistic and study the asymptotic properties of this statistic in detail in Section 2.2. In Section 3, we provide extensive numerical results to compare the finite-sample performance of our proposed tests with existing competitors. We conclude this paper with a brief discussion in Section 4. All technical proofs are provided in the Supplementary Material.

2. Projective Ensemble Test

2.1. Motivation

In this section, we develop a new projective ensemble Cramér-von Mises test for equality of distributions. We start with the rationales. It is straightforward that $\mathbf{x}$ and $\mathbf{y}$ are equally distributed if and only if $(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ and $(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{y})$ are homogeneous, for all $\boldsymbol{\beta} \in \mathbb{R}^p$. Therefore, with suitable weight functions, it is additionally equivalent to

$$\iint \{F_{\boldsymbol{\beta}}(t) - G_{\boldsymbol{\beta}}(t)\}^2 dH(\boldsymbol{\beta}, t) = 0. \tag{2}$$

When $dH(\boldsymbol{\beta}, t) = dH_{\boldsymbol{\beta}}(t)d\lambda(\boldsymbol{\beta})$, $H_{\boldsymbol{\beta}}(t)$ is chosen to be $\tau F_{\boldsymbol{\beta}}(t) + (1-\tau)G_{\boldsymbol{\beta}}(t)$, and $\lambda(\boldsymbol{\beta})$ is the uniform probability measure on the p-dimensional unit sphere $\mathcal{S}^{p-1} \overset{\text{def}}{=} \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\| = 1\}$, the integration has a closed form and corresponds to the projection-averaging approach pro-

posed by Kim et al. (2020). Similar techniques to obtain a closed-form expression are also used by Escanciano (2006) and Zhu et al. (2017). However, the pooled distribution function $H_{\boldsymbol{\beta}}(t)$ is typically unknown and needs to be estimated from the sample in practice, which incorporates heavy computational burdens and greatly limits its application in practice. When $H_{\boldsymbol{\beta}}(t) = t$, the integration also has a closed form and coincides with the energy statistic (Székely & Rizzo, 2004; Baringhaus & Franz, 2004). However, in order to ensure the integration is finite, the moment condition should be imposed, which makes the resulting test sensitive to heavy-tailed distributions. These observations motivate us to carefully choose other weight functions such that

1. The integration in (2) equals zero if and only if $\mathbf{x}$ and $\mathbf{y}$ are equally distributed;
2. The choice of $H(\boldsymbol{\beta}, t)$ does not depend on unknown functions which are difficult to estimate;
3. The integration in (2) has a closed-form expression, and is finite without any moment conditions.

To achieve these goals simultaneously, we note that $F_{\boldsymbol{\beta}}(t) = \mathrm{P}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} \le t)$ and $G_{\boldsymbol{\beta}}(t) = \mathrm{P}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{y} \le t)$, the integration in (2) can be rewritten as

$$\begin{aligned} & \iint F_{\boldsymbol{\beta}}^2(t) dH(\boldsymbol{\beta}, t) - 2 \iint F_{\boldsymbol{\beta}}(t) G_{\boldsymbol{\beta}}(t) dH(\boldsymbol{\beta}, t) \\ & \qquad + \iint G_{\boldsymbol{\beta}}^2(t) dH(\boldsymbol{\beta}, t) \\ = & \iint E\left\{I(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_1 \le t, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_2 \le t)\right\} dH(\boldsymbol{\beta}, t) \\ & -2 \iint E\left\{I(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_1 \le t, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{y}_2 \le t)\right\} dH(\boldsymbol{\beta}, t) \\ & + \iint E\left\{I(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{y}_1 \le t, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{y}_2 \le t)\right\} dH(\boldsymbol{\beta}, t), \end{aligned}$$

where $\mathbf{x}_1, \mathbf{x}_2$ and $\mathbf{y}_1, \mathbf{y}_2$ are two independent copies of $\mathbf{x}$ and $\mathbf{y}$, respectively. In order to obtain a closed-form expression, we need to evaluate the three integrations in the above display. We take the first integration for example. By adopting Fubini's theorem, it suffices to find $H(\boldsymbol{\beta}, t)$ such that the following integration

$$\iint I(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_1 \le t, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_2 \le t) dH(\boldsymbol{\beta}, t)$$

has a closed form for given $\mathbf{x}_1$ and $\mathbf{x}_2$. The following lemma implicitly shows that, by taking $H(\boldsymbol{\beta}, t)$ as the cumulative distribution function of a $p+1$ dimensional multivariate joint normal random vector with mean $\mathbf{0}$ and covariance $\mathbf{I}_{p+1}$, the above integration can be explicitly calculated.

Lemma 1. *(Gupta, 1963) Let $(Z_1, Z_2)^T$ be bivariate normal with mean $\mathbf{0}$. The correlation between Z_1 and Z_2 is ρ, then*

$$P(Z_1 \ge 0, Z_2 \ge 0) = \frac{1}{4} + \frac{1}{2\pi} \arcsin(\rho).$$

According to Lemma 1, by treating $\mathbf{x}_1$ and $\mathbf{x}_2$ as constants, $(\boldsymbol{\beta}, t)^{\mathrm{T}}$ as a $p+1$ dimensional multivariate joint normal random vector with cumulative distribution function $H(\boldsymbol{\beta}, t)$, the integration can be expressed as

$$\begin{aligned} & \iint I(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_1 \le t, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_2 \le t) dH(\boldsymbol{\beta}, t) \\ = & \ \mathrm{P}\left(t - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_1 \ge 0, t - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_2 \ge 0 \,\middle|\, \mathbf{x}_1, \mathbf{x}_2\right) \\ = & \ \frac{1}{4} + \frac{1}{2\pi} \arcsin\left(\frac{1 + \mathbf{x}_1^{\mathrm{T}}\mathbf{x}_2}{\sqrt{1 + \mathbf{x}_1^{\mathrm{T}}\mathbf{x}_1}\sqrt{1 + \mathbf{x}_2^{\mathrm{T}}\mathbf{x}_2}}\right). \end{aligned}$$

Consequently, the integration in (2) can be expressed in a closed form, which is shown in the following Theorem.

Theorem 1. *Suppose $\mathbf{x} = (X_1, \ldots, X_p)^T \in \mathbb{R}^p$ and $\mathbf{y} = (Y_1, \ldots, Y_p)^T \in \mathbb{R}^p$ are two p-dimensional random vectors whose distribution functions are F and G, respectively. $\mathbf{x}_1, \mathbf{x}_2$ and $\mathbf{y}_1, \mathbf{y}_2$ are two independent copies of $\mathbf{x}$ and $\mathbf{y}$, respectively. Let $H(\boldsymbol{\beta}, t)$ be the cumulative distribution function of a $p+1$ dimensional multivariate joint normal random vector with mean $\mathbf{0}$ and covariance $\mathbf{I}_{p+1}$. Then*

$$\begin{aligned} T &= 2\pi \iint \{F_{\boldsymbol{\beta}}(t) - G_{\boldsymbol{\beta}}(t)\}^2 dH(\boldsymbol{\beta}, t) \\ &= T_1 - 2T_2 + T_3, \end{aligned} \tag{3}$$

where T_1, T_2 and T_3 are defined as

$$\begin{aligned} T_1 &\overset{def}{=} E \arcsin\left(\frac{1 + \mathbf{x}_1^T\mathbf{x}_2}{\sqrt{1 + \mathbf{x}_1^T\mathbf{x}_1}\sqrt{1 + \mathbf{x}_2^T\mathbf{x}_2}}\right), \\ T_2 &\overset{def}{=} E \arcsin\left(\frac{1 + \mathbf{x}_1^T\mathbf{y}_2}{\sqrt{1 + \mathbf{x}_1^T\mathbf{x}_1}\sqrt{1 + \mathbf{y}_2^T\mathbf{y}_2}}\right), \\ T_3 &\overset{def}{=} E \arcsin\left(\frac{1 + \mathbf{y}_1^T\mathbf{y}_2}{\sqrt{1 + \mathbf{y}_1^T\mathbf{y}_1}\sqrt{1 + \mathbf{y}_2^T\mathbf{y}_2}}\right). \end{aligned}$$

In addition, T is nonnegative and equals zero if and only if $F = G$.

Theorem 1 clearly indicates that T defined in (3) can be served as an index to distinguish whether two random vectors, $\mathbf{x}$ and $\mathbf{y}$, are equally distributed without moment condition or continuity assumption. It is nonnegative and equals zero if and only if $\mathbf{x}$ and $\mathbf{y}$ are equally distributed.

2.2. Asymptotic Properties

According to Theorem 1, it is natural to use the sample version of T as the test statistic when testing for equality of distributions. Suppose $\{\mathbf{x}_i, i = 1, \ldots, m\}$ and $\{\mathbf{y}_i, i = 1, \ldots, n\}$ are two mutually independent random samples drawn from F and G, with sample size m and n,

respectively. At the sample level, we estimate T_1, T_2, and T_3 by

$$\widehat{T}_1 \overset{\text{def}}{=} \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\arcsin\left(\frac{1+\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j}{\sqrt{1+\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i}\sqrt{1+\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_j}}\right),$$

$$\widehat{T}_2 \overset{\text{def}}{=} \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\arcsin\left(\frac{1+\mathbf{x}_i^{\mathrm{T}}\mathbf{y}_j}{\sqrt{1+\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i}\sqrt{1+\mathbf{y}_j^{\mathrm{T}}\mathbf{y}_j}}\right),$$

$$\widehat{T}_3 \overset{\text{def}}{=} \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\arcsin\left(\frac{1+\mathbf{y}_i^{\mathrm{T}}\mathbf{y}_j}{\sqrt{1+\mathbf{y}_i^{\mathrm{T}}\mathbf{y}_i}\sqrt{1+\mathbf{y}_j^{\mathrm{T}}\mathbf{y}_j}}\right).$$

Then, the test statistic is given by

$$\widehat{T} = \widehat{T}_1 - 2\widehat{T}_2 + \widehat{T}_3.$$

Because $\widehat{T}_1$, $\widehat{T}_2$, and $\widehat{T}_3$ are all standard V statistics, $\widehat{T}$ can be calculated directly with computational complexity $O\{(m+n)^2\}$, whereas the computational cost for Kim et al. (2020) is of order $O\{(m+n)^3\}$. We now give the asymptotic properties of the test statistic under the null hypothesis in Theorem 2.

Theorem 2. *Under the null hypothesis, that is, $F = G$, as $\min(m,n)\to\infty$, $\frac{mn}{m+n}\widehat{T}$ converges in distribution to*

$$2\pi\iint\{\zeta(\boldsymbol{\beta},t)\}^2 dH(\boldsymbol{\beta},t),$$

where $\zeta(\boldsymbol{\beta},t)$ is a gaussian random process with mean zero and covariance function, $\operatorname{cov}\{\zeta(\boldsymbol{\beta},t),\zeta(\boldsymbol{\alpha},s)\}$, is given by

$$P\left(\boldsymbol{\beta}^T\mathbf{x}\le t,\boldsymbol{\alpha}^T\mathbf{x}\le s\right) - P\left(\boldsymbol{\beta}^T\mathbf{x}\le t\right)P\left(\boldsymbol{\alpha}^T\mathbf{x}\le s\right). \quad (4)$$

Note that $mn/(m+n)^2\to\tau(1-\tau)$ as $\min(m,n)\to\infty$, where τ is the limit value of $m/(m+n)$. Theorem 2 indicates that our test statistic is $(m+n)$ consistent under the null hypothesis without requiring moment condition or continuity assumption. However, the covariance function defined in (4) is typically distribution dependent, which makes the limiting null distribution intractable. In practice, we propose the permutation procedure to approximate the null distribution. Before that, we study the asymptotic properties of $\widehat{T}$ under two kinds of alternative hypotheses, i.e., the global alternative and the local alternative.

Under the global alternative, $F\neq G$ and the difference between the two distribution functions does not vary with the sample size. In this case, the proposed test statistic is asymptotic normal, which is stated in the following Theorem.

Theorem 3. *Under the global alternative hypothesis, as $\min(m,n)\to\infty$, $(m+n)^{1/2}(\widehat{T}-T)$ converges in distribution to*

$$\mathcal{N}\left\{0,\frac{4(1-\tau)\mathrm{var}(Z_1)+4\tau\mathrm{var}(Z_2)}{\tau(1-\tau)}\right\}$$

where $\tau\in(0,1)$ is the limit value of $m/(m+n)$, Z_1 and Z_2 are defined in (S.3.1) and (S.3.2) in the Supplementary Material, respectively.

From Theorem 3, it is clear to see that our proposed test statistic is root-$(m+n)$ consistent under any fixed alternative. Recall that the test statistic is $(m+n)$ consistent under the null hypothesis. Then our proposed test can consistently detect any fixed alternatives with probability approaching one.

Under the local alternative, $F\neq G$ but the difference between the two distribution functions diminishes as the sample size increases. We consider a sequence of local alternatives as follows:

$$H_{1l}: \mathrm{P}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\le t) = \mathrm{P}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{y}\le t) + (m+n)^{-1/2}\ell(\boldsymbol{\beta},t),$$

where $\ell(\boldsymbol{\beta},t)$ is a function depending only on $\boldsymbol{\beta}$ and t such that $\iint\ell^2(\boldsymbol{\beta},t)dH(\boldsymbol{\beta},t)$ exists.

Theorem 4. *Under the local alternative hypothesis, as $\min(m,n)\to\infty$, $\frac{mn}{m+n}\widehat{T}$ converges in distribution to*

$$2\pi\iint\{\zeta(\boldsymbol{\beta},t)+\tau^{1/2}(1-\tau)^{1/2}\ell(\boldsymbol{\beta},t)\}^2 dH(\boldsymbol{\beta},t),$$

where τ is the limit value of $m/(m+n)$, $\zeta(\boldsymbol{\beta},t)$ is a gaussian random process defined in Theorem 2.

Theorem 4 implicitly shows that, our proposed test asymptotically has nontrivial power performance when the difference between the two distribution functions is at the rate of $(m+n)^{-1/2}$. That is, as long as the difference is larger than $O\left\{(m+n)^{-1/2}\right\}$, it can be consistently detected by our proposed test with probability tending to one.

However, it is yet unclear how to approximate the limiting null distribution because it equals the weighted sum of infinite number of chi squared variables. In addition, the weights are distribution dependent and are typically unknown. To address this issue, we now propose the permutation procedure as follows:

1. Let $\{\mathbf{z}_1,\mathbf{z}_2,\ldots,\mathbf{z}_{m+n}\} = \{\mathbf{x}_1,\ldots,\mathbf{x}_m,\mathbf{y}_1,\ldots,\mathbf{y}_n\}$ denote the pooled samples. Randomly permute the pooled samples to obtain $\{\mathbf{z}_1^*,\mathbf{z}_2^*,\ldots,\mathbf{z}_{m+n}^*\}$.

2. Select the first m observations from the pooled samples as $\{\mathbf{x}_1^*,\ldots,\mathbf{x}_m^*\}$, and the rest observations as $\{\mathbf{y}_1^*,\ldots,\mathbf{y}_n^*\}$.

3. Based on the two randomly permuted samples $\{\mathbf{x}_1^*,\ldots,\mathbf{x}_m^*\}$ and $\{\mathbf{y}_1^*,\ldots,\mathbf{y}_n^*\}$, calculate the test statistic to obtain $\widehat{T}^*$.

4. Repeat steps 1 to 3 for B times to obtain $\widehat{T}_b^*, b =$

$1, 2, \ldots, B$. The associated p-value is given by

$$B^{-1} \sum_{b=1}^{B} I\left(\widehat{T}_b^* \geq \widehat{T}\right),$$

where $I(\cdot)$ is an indicator function. Reject the null hypothesis if the p-value is smaller than the given significance level.

Intuitively, since the randomly permutated samples follow the same distribution with the pooled sample, $\mathbf{x}_i^*$ and $\mathbf{y}_j^*$ are equally distributed, it is thus reasonable to use the permutation procedure to approximate the null distribution. Theorem 5 states the consistency of the above permutation procedure.

Theorem 5. *As* $\min(m, n) \to \infty$,

$$\sup_{t \geq 0} \left| P\left\{ mn/(m+n)\widehat{T}^* \leq t \mid \mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{y}_1, \ldots, \mathbf{y}_n \right\} - P(T_\infty^* \leq t) \right|$$

converges in probability to 0, where T_∞^* *is defined as*

$$T_\infty^* = 2\pi \iint \{\zeta^*(\boldsymbol{\beta}, t)\}^2 dH(\boldsymbol{\beta}, t)$$

and $\zeta^*(\boldsymbol{\beta}, t)$ *is a gaussian random process with mean zero and covariance function,* $\operatorname{cov}\{\zeta^*(\boldsymbol{\beta}, t), \zeta^*(\boldsymbol{\alpha}, s)\}$, *given by*

$$\begin{aligned} &\tau P\left(\boldsymbol{\beta}^T\mathbf{x} \leq t, \boldsymbol{\alpha}^T\mathbf{x} \leq s\right) + (1-\tau)P(\boldsymbol{\beta}^T\mathbf{y} \leq t, \boldsymbol{\alpha}^T\mathbf{y} \leq s) \\ &- \left\{\tau P(\boldsymbol{\beta}^T\mathbf{x} \leq t) + (1-\tau)P(\boldsymbol{\beta}^T\mathbf{y} \leq t)\right\}\left\{\tau P(\boldsymbol{\alpha}^T\mathbf{x} \leq s) \right. \\ &\left. + (1-\tau)P(\boldsymbol{\alpha}^T\mathbf{y} \leq s)\right\}. \end{aligned} \tag{5}$$

We remark here that the covariance function of $\zeta^*(\boldsymbol{\beta}, t)$ in (5) coincides with that of $\zeta(\boldsymbol{\beta}, t)$ in (4) under the null hypothesis, while is different from that of $\zeta(\boldsymbol{\beta}, t)$ in (4) under the alternative. That is, $\widehat{T}^*$ has the same limiting distribution as $\widehat{T}$ based on the original samples under the null, but has a different distribution under the alternative. However, the permutation procedure can still provide an asymptotically valid inference for the proposed test. Specifically, it provides asymptotically valid critical values under the null because the limiting distribution of $\widehat{T}^*$ and $\widehat{T}$ are asymptotically the same. While under the alternative, the test is still powerful because $\widehat{T}^*$ is $(m+n)$ consistent while $\widehat{T}$ is root-$(m+n)$ consistent in this case. This implies that $\widehat{T}^*$ is of order $O_p(1)$ while $\widehat{T} \to \infty$ in probability, and the corresponding p-value converges in probability to zero.

3. Numerical Studies

3.1. Simulations

In this section, we study the finite sample performance of the proposed method through extensive numerical analysis. We also compare the performance of the projection ensemble based test ("PE") with other competing nonparametric tests. Specifically, they are the projection-averaging based Cramér-von Mises test (Kim et al., 2020, "CvM"), the k nearest neighbor test (Henze, 1988, "NN"), the modified k nearest neighbor test (Mondal et al., 2015, "MGB"), the energy statistic based test (Székely & Rizzo, 2004, "Energy"), the inter-point distance test (Biswas & Ghosh, 2014, "BG"), the cross-match test (Rosenbaum, 2005, "CM"), and ball divergence test (Pan et al., 2018, "Ball").

Table 1. The empirical powers for different methods when all the samples are generated from multivariate normal distributions at significance level $\alpha = 0.05$.

	PE	CvM	NN	MGB
Location	1.000	1.000	0.999	0.996
Scale	0.713	0.880	0.723	1.000
Location-Scale	0.966	1.000	0.997	1.000
	Energy	**BG**	**CM**	**Ball**
Location	1.000	1.000	0.994	1.000
Scale	0.989	1.000	0.169	1.000
Location-Scale	1.000	1.000	0.765	1.000

We generate the samples $\{\mathbf{x}_i, \mathbf{x}_i \in \mathbb{R}^p, i = 1, \ldots, n_x\}$, $\{\mathbf{y}_i, \mathbf{y}_i \in \mathbb{R}^p, i = 1, \ldots, n_y\}$, and $\{\mathbf{z}_i, \mathbf{z}_i \in \mathbb{R}^p, i = 1, \ldots, n_z\}$ independently from $t_d(\mu_x \mathbf{1}_p, \sigma_x^2 \mathbf{I}_p)$, $t_d(\mu_y \mathbf{1}_p, \sigma_y^2 \mathbf{I}_p)$, and $t_d(\mu_z \mathbf{1}_p, \sigma_z^2 \mathbf{I}_p)$, respectively, where $\mathbf{1}_p$ is a p-dimensional vector whose entries are all 1 and $\mathbf{I}_p$ is a p-by-p identity matrix. Here $t_d(\mu_x \mathbf{1}_p, \sigma_x^2 \mathbf{I}_p)$ stands for a multivariate t distribution with location parameter $\mu_x \mathbf{1}_p$ and scale matrix $\sigma_x^2 \mathbf{I}_p$, and the degrees of freedom is d. We set $\mu_x = 0$, $\sigma_x = 1$, $\mu_y = 1$, $\sigma_y = 1$, $\mu_z = 1$, and $\sigma_z = 2$, so that we can inspect the power performance of location shift by comparing the distribution of $\mathbf{x}$ and $\mathbf{y}$, the power performance of scale difference by comparing the distribution of $\mathbf{y}$ and $\mathbf{z}$, and the power performance of both location shift and scale difference by comparing the distribution of $\mathbf{x}$ and $\mathbf{z}$. Throughout the experiment, we set the significance level as 0.05. We repeat each experiment 1000 times and determine the critical values with 1000 permutations.

We first consider the case that all the samples are generated from multivariate normal distributions. That is, we set $d = \infty$ in this experiment. We set all the sample size to be 20, i.e., $n_x = n_y = n_z = 20$. The dimension is set to be $p = 10$. The corresponding power performances are charted in Table 1. It can be seen clearly that, all methods except the cross-match test perform fairly well in this normal case. The cross-match test is not efficient in detecting the scale difference may be mainly because it relies on some tuning parameters.

We now consider the case that the moment conditions are

Table 2. The empirical powers for different methods when all the samples are generated from Cauchy distributions at significance level $\alpha = 0.05$.

	PE	CVM	NN	MGB
LOCATION	1.000	1.000	0.139	0.056
SCALE	0.550	0.555	0.311	0.601
LOCATION-SCALE	0.998	1.000	0.335	0.582
	ENERGY	BG	CM	BALL
LOCATION	0.043	0.048	1.000	0.036
SCALE	0.302	0.217	0.108	0.577
LOCATION-SCALE	0.289	0.215	0.872	0.581

Table 3. The empirical powers for different methods when all the samples are generated from Cauchy distributions and the sample sizes are imbalanced at significance level $\alpha = 0.05$.

	PE	CVM	NN	MGB
LOCATION	0.999	0.999	0.084	0.040
SCALE	0.397	0.477	0.000	0.737
LOCATION-SCALE	1.000	1.000	0.000	0.857
	ENERGY	BG	CM	BALL
LOCATION	0.055	0.052	0.996	0.037
SCALE	0.259	0.026	0.045	0.520
LOCATION-SCALE	0.372	0.089	0.943	0.728

not fulfilled. We set $d = 1$ in this experiment. That is, all the elements follow Cauchy distributions. We set the dimension to be $p = 100$ and the samples sizes are all set to be 20. Table 2 summarizes the empirical powers for all the tests, from which we can see that some methods are not efficient in this case. For example, when detecting the location shift, the power for the energy statistic based test is only 0.043. This indicates that the energy statistic based tests may not be efficient when the moment condition is not satisfied. However, both our methods and the projection-averaging based Cramér-von Mises test are powerful even when extreme values exist.

We also consider the case that the sample sizes are imbalanced. We let $d = 1$ so that all the elements follow Cauchy distributions. We also set the dimension to be $p = 100$ and the samples sizes are $n_x = 20, n_y = 10$, and $n_z = 40$, respectively. The resultant power performances for all the tests are summarized in Table 3. Since all the settings are the same as that in the last experiment, we compare the results in Table 3 with that in Table 2. We can see that the k nearest neighbour test is sensitive to the imbalance of the sample size, while our method performs similarly compared with the balanced case.

Next, we inspect the execution time for different methods when testing for the homogeneity of the samples $\{\mathbf{x}_i, \mathbf{x}_i \in \mathbb{R}^p, i = 1, \ldots, n_x\}$ and $\{\mathbf{y}_i, \mathbf{y}_i \in \mathbb{R}^p, i = 1, \ldots, n_y\}$. We set all the settings the same as that of the multivariate normal case, except we vary the sample size n from $\{20, 50, 100\}$, where $n_x = n_y = n$. This experiment is run on a laptop (Dell XPS 13 9360, Intel Core i7-8550U CPU @ 1.80GHz 2.00 GHz). All tests are implemented with R 3.6.1 for fair comparison. The average running time of calculating the test statistic for different methods are reported in Table 4. According to Table 4, we can see that energy statistic based test, the inter-point distance test and our projection ensemble based test are superior to the others. Meanwhile, the projection-averaging based Cramér-von Mises test suffers from heavy computations, which may greatly limit its usage in practice, especially when the sample size becomes large.

Table 4. The average running time (in milliseconds) for different methods.

n	PE	CVM	NN	MGB
20	0.16	6.11	1.00	0.98
50	0.27	37.19	3.17	2.72
100	0.63	213.69	7.46	7.36
n	ENERGY	BG	CM	BALL
20	0.11	0.11	3.69	4.49
50	0.29	0.30	6.28	20.22
100	1.17	1.06	16.20	81.56

To sum up, our proposed projection ensemble based method is comparable with the projection-averaging based Cramér-von Mises test in terms of power performance, and is superior to the other tests across almost all the cases, especially in the presence of the heavy-tailed distributions. However, the projection-averaging based Cramér-von Mises test suffers from cubic computations, while our method is much more computationally efficient.

3.2. An Application

In this section, we apply the proposed test with projection ensemble to Daily Demand Forecasting Orders Data Set (Ferreira et al., 2016) from the UCI machine learning repository. This is a real database of a Brazilian company of large logistics, which was collected during 60 days. The original data consists of twelve predictive attributes and a target feature. In this study, we inspect whether the demand on Friday is significantly different from other weekdays. Specifically, we consider the following features: Non urgent order (X_1), urgent order (X_2), three order types (X_3, X_4, X_5), fiscal sector orders (X_6), orders from the traffic controller sector (X_7), three kinds of banking orders (X_8, X_9, X_{10}), total orders (X_{11}).

We adopt the projection ensemble based test and other competing nonparametric tests described in Section 3.1 on this

Table 5. The empirical p-values for different methods for the daily demand forecasting orders data set.

	PE	CvM	NN	MGB
P-VALUE	0.008	0.004	0.249	0.180
	ENERGY	BG	CM	BALL
P-VALUE	0.011	0.210	0.554	0.112

data set. The resulting p-values based on 1000 permutations are charted in Table 5. Following Liu & Xie (2019), we adopt the Cauchy combination test statistic as

$$T = \frac{1}{8}\sum_{i=1}^{8} \tan\left\{(0.5 - p_i)\,\pi\right\},$$

where p_i is the p-value of the ith test. Because each permutation p-value is approximately uniformly distributed under the null, the test statistic T is the average of 8 standard Cauchy random variables, which also has a standard Cauchy distribution when all p_is are mutually independent. As pointed by Liu & Xie (2019), although correlations among p-values will affect the null distribution of the Cauchy combination test statistic, the impact on the tail is very limited. Therefore, we combine all p-values in this way and get the final p-value as follows,

$$p = 1/2 - \pi^{-1}\arctan(T).$$

The corresponding p-value is 0.0164, which indicates that the demand on Friday is significantly different from other weekdays at significance level $\alpha = 0.05$. Meanwhile, according to Table 5, we can see that only our method, the projection-averaging based Cramér-von Mises test (Kim et al., 2020), and the energy statistic based test (Székely & Rizzo, 2004) can reject the null hypothesis at significance level $\alpha = 0.05$, which further confirms that our method is very powerful when testing for equality of distributions.

4. Conclusion and Discussion

In this work, we apply the idea of projections and propose a robust test for the multivariate two-sample problem. Through projection ensemble, the proposed test statistic is a generalization of the univariate Cramér-von Mises test statistic, which has similar properties with Baringhaus & Franz (2004) and Kim et al. (2020). It is demonstrated that with a suitable choice of the ensemble approach, the proposed test statistic has a simple closed-form expression without any tuning parameters. It is easy to implement, insensitive to the dimension and robust to the presence of potential extreme values or heavily tailed observations. Extensive numerical studies indicate that the proposed projection ensemble based test is superior to most existing tests, especially in the presence of the heavy-tailed distributions. Moreover, it is comparable with the projection-averaging based Cramér-von Mises test in terms of power performance, but much more efficient in terms of computation.

It is notable that although the proposed test statistic is more computationally efficient than that of Kim et al. (2020), it still requires $O\{(m+n)^2\}$ calculations and $O\{(m+n)^2\}$ memory for storing them. This may greatly limit its usage for large datasets, especially in the era of big data because we may encounter millions of observations in practice. Therefore, it is also crucial to develop efficient algorithms when computing the test statistic when the sample size becomes large. In univariate cases, we can adopt AVL tree-type implementation to develop an efficient algorithm with complexity $O\{(m+n)\log(m+n)\}$ when calculating the test statistic, which is also used in Huang & Huo (2017) when computing energy statistics. In the case of multivariate random variables, we can approximate the test statistic with random projections, whose computational cost can be reduced to $O\{(m+n)K\log(m+n)\}$ and memory cost $O\{\max(m+n, K)\}$, where K is the number of random projections. It can be anticipated that the resulting test can achieve nearly the same power as the direct test (Huang & Huo, 2017).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (11801349).

References

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

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

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

Bhattacharya, B. B. A general asymptotic framework for distribution-free graph-based two-sample tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(3):575–602, 2019.

Biswas, M. and Ghosh, A. K. A nonparametric two-sample test applicable to high dimensional data. *Journal of Multivariate Analysis*, 123:160–171, 2014.

Cai, T. T. and Liu, W. Large-scale multiple testing of correlations. *Journal of the American Statistical Association*, 111(513):229–240, 2016.

Cai, T. T., Liu, W., and Xia, Y. Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76 (2):349–372, 2014.

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

Chen, H., Chen, X., and Su, Y. A weighted edge-count two-sample test for multivariate and object data. *Journal of the American Statistical Association*, 113(523):1146–1155, 2018.

Darling, D. A. The kolmogorov-smirnov, cramer-von mises tests. *The Annals of Mathematical Statistics*, 28(4):823–838, 1957.

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

Ferreira, R. P., Martiniano, A., Ferreira, A., Ferreira, A., and Sassi, R. J. Study on daily demand forecasting orders using artificial neural network. *IEEE Latin America Transactions*, 14(3):1519–1525, 2016.

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

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

Gupta, S. S. Probability integrals of multivariate normal and multivariate t1. *The Annals of mathematical statistics*, pp. 792–828, 1963.

Harchaoui, Z., Bach, F., Cappe, O., and Moulines, E. Kernel-based methods for hypothesis testing: A unified view. *IEEE Signal Processing Magazine*, 30(4):87–97, 2013.

Henze, N. A multivariate two-sample test based on the number of nearest neighbor type coincidences. *The Annals of Statistics*, 16(2):772–783, 1988.

Huang, C. and Huo, X. An efficient and distribution-free two-sample test based on energy statistics and random projections. *arXiv preprint arXiv:1707.04602*, 2017.

Kim, I., Balakrishnan, S., and Wasserman, L. Robust multivariate nonparametric tests via projection-averaging. *The Annals of Statistics*, 2020.

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

Li, J. and Chen, S. X. Two sample tests for high-dimensional covariance matrices. *The Annals of Statistics*, 40(2):908–940, 2012.

Liu, Y. and Xie, J. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, pp. 1–18, 2019.

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

Pan, W., Tian, Y., Wang, X., and Zhang, H. Ball divergence: Nonparametric two sample test. *Annals of statistics*, 46 (3):1109, 2018.

Rosenbaum, P. R. An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4):515–530, 2005.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013.

Smirnov, N. V. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Math. Univ. Moscou*, 2(2):3–14, 1939.

Székely, G. J. and Rizzo, M. L. Testing for equal distributions in high dimension. *InterStat*, 5(16.10):1249–1272, 2004.

Zhu, L., Xu, K., Li, R., and Zhong, W. Projection correlation between two random vectors. *Biometrika*, 104(4):829–843, 2017.