| Statistica Sinica Preprint No: SS-2023-0069 | |
|---|---|
| **Title** | High-Dimensional Behaviour of Some Two-Sample Tests Based on Ball Divergence |
| **Manuscript ID** | SS-2023-0069 |
| **URL** | http://www.stat.sinica.edu.tw/statistica/ |
| **DOI** | 10.5705/ss.202023.0069 |
| **Complete List of Authors** | Bilol Banerjee and Anil K. Ghosh |
| **Corresponding Authors** | Anil K. Ghosh |
| **E-mails** | akghosh@isical.ac.in |
| Notice: Accepted version subject to English editing. | |

# High Dimensional Behaviour of Some Two-Sample Tests Based on Ball Divergence

Bilol Banerjee and Anil K. Ghosh

*Indian Statistical Institute, Kolkata*

*Abstract:* We propose some two-sample tests based on ball divergence and investigate their high dimensional behaviour. First, we consider the High Dimension, Low Sample Size (HDLSS) setup. Under appropriate regularity conditions, we establish the consistency of these tests in the HDLSS regime, where the dimension grows to infinity while the sample sizes from the two distributions remain fixed. Next, we show that these conditions can be relaxed when the sample sizes also increase with the dimension, and in such cases, consistency can be proved even for shrinking alternatives. We use a simple example to show that even when there are no consistent tests in the HDLSS regime, the proposed tests can be consistent if the sample sizes increase with the dimension at an appropriate rate. This rate is obtained by establishing the minimax rate optimality of these tests over a certain class of alternatives. Several simulated and benchmark data sets are analyzed to compare the empirical performance of these tests with some state-of-the-art methods available for testing the equality of two high-dimensional distributions.

*Key words and phrases:* Ball divergence, Energy statistics, High dimensional asymptotics, Minimax rate optimality, Permutation tests, Shrinking alternatives.

## 1. Introduction

In a two-sample problem, we test for the equality of two $d$-dimensional distributions $F$ and $G$ based on $n$ independent copies $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of $\mathbf{X} \sim F$ and $m$ independent copies $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ of $\mathbf{Y} \sim G$. This problem is well investigated in the literature, and several tests are available for it. In the parametric regime, we often assume $F$ and $G$ to be Gaussian and test for the equality of their location and/or scale parameters. Several nonparametric tests are also available, especially for $d = 1$. While the Wilcoxon-Mann-Whitney test is used for the univariate two-sample location problem, the Wald-Wolfowitz run test, the Kolmogorov-Smirnov (KS) test, and the Camer-von-Mises (CVM) test (see. e.g., Hollander et al., 2014; Gibbons and Chakraborti, 2011) are applicable to general two-sample problems.

Using the idea of a minimum spanning tree, Friedman and Rafsky (1979) generalized the run test and the KS test to higher dimensions. Baringhaus and Franz (2004) proposed a test based on inter-point distances, which can be viewed as a multivariate generalization of the CVM test. Aslan and Zech (2005) also used inter-point distances to develop tests based on energy statistics. Schilling (1986) and Henze (1988) developed multivariate two-sample tests based on nearest-neighbors. Rosenbaum (2005) proposed a distribution-free test based on optimal non-bipartite matching. Gretton

et al. (2012) used the notion of maximum mean discrepancy (MMD) to construct a test based on kernel mean embedding of two probability distributions. These multivariate two-sample tests are consistent in the classical asymptotic regime. For any fixed $d$, the powers of these tests converge to unity as $n$ and $m$ increase. Since these tests are based on pairwise distances among the observations, they can be conveniently used for high-dimensional data even when the dimension is much larger than the combined sample size. But, most of them often perform poorly in the high dimension, low sample size (HDLSS) situations, especially when the scale difference between $F$ and $G$ dominates their location difference (see, e.g., Biswas and Ghosh, 2014).

Following the seminal paper by Hall et al. (2005), recently the HDLSS regime has received increasing attention. Over the last ten years, several two-sample tests have been proposed for HDLSS data. Wei et al. (2016); Ghosh and Biswas (2016); Srivastava et al. (2016) proposed some tests based on linear projections, which are mainly useful for two-sample location problems. Biswas and Ghosh (2014) and Tsukada (2019) proposed some general two-sample tests based on averages of inter-point distances. Under some appropriate assumptions, these two tests turn out to be consistent in both classical and HDLSS asymptotic regimes but nothing is known about their asymptotic behaviour when the sample sizes increase with

the dimension. Moreover, they are not robust against outliers generated from heavy-tailed distributions. Kim et al. (2020) developed a robust multivariate test based on projection averaging but it is applicable only when the distances between the observations are measured using the Euclidean metric. Some graph-based high-dimensional two-sample tests have also been proposed in the literature. This includes the test based on nearest neighbors (Mondal et al., 2015), multivariate run test based on the shortest Hamiltonian path (Biswas et al., 2014) and the test based on triangles (Liu and Modarres, 2011). Under appropriate regularity conditions, these graph-based tests are consistent in the HDLSS asymptotic regime. But in classical asymptotic regime, they usually have poor powers against local alternatives (see, Bhattacharya, 2019). Even the large sample consistency of the SHP-based run test and the triangle test is yet to be proved. Also, it is not known how these tests perform when the dimension and the sample sizes grow simultaneously. This type of asymptotic behaviour has been studied for some location (see, e.g., Bai and Saranadasa, 1996; Chen and Qin, 2010; Aoshima and Yata, 2018) and scale (see,e.g., Li and Chen, 2012; Cai et al., 2013) tests, but for the general two-sample test, the literature is scarce.

In this article, we propose some two-sample tests based on ball divergence and study their high dimensional behaviour in the HDLSS setup as well as

in situations, where the dimension and sample sizes grow simultaneously. It is organized as follows. In Section 2, we construct a permutation test based on ball divergence and prove its large-sample consistency. In Section 3, we study the performance of this test in the HDLSS regime. We observe that the test based on the $\ell_2$ distance may fail to discriminate between two distributions differing outside the first two moments. To take care of this problem, we propose tests based on other appropriate distance functions and prove their high dimensional consistency for a larger class of alternatives. In Section 4, we consider the case where the sample sizes also increase with the dimension. We establish the minimax rate optimality of the proposed tests and prove their consistency under shrinking alternatives. Some simulated and real data sets are analyzed in Section 5 to compare the performance of our tests with some state-of-the-art methods. Finally, Section 6 contains a brief summary of the work and a discussion on some possible directions for future research. All proofs and mathematical details and some additional numerical results are given in the supplementary material.

## 2. The proposed test based on ball divergence

Let $\mathbf{X} \sim F$ and $\mathbf{Y} \sim G$ be two random variables taking values on a separable metric space $(\mathbb{R}^d, \rho)$, where $\rho$ is a metric on $\mathbb{R}^d$. We know that $F$ and $G$

differ if and only if there exists a ball $\mathbb{B}(\mathbf{u}, \epsilon) := \{\mathbf{v} \in \mathbb{R}^d \mid \rho(\mathbf{v}, \mathbf{u}) \leq \epsilon\}$ such that $F(\mathbb{B}(\mathbf{u}, \epsilon)) \neq G(\mathbb{B}(\mathbf{u}, \epsilon))$. So, for any $\epsilon > 0$, $|F(\mathbb{B}(\mathbf{u}, \epsilon)) - G(\mathbb{B}(\mathbf{u}, \epsilon))|$ gives a measure of difference between $F$ and $G$ in a neighborhood of $\mathbf{u} \in \mathbb{R}^d$. Therefore, for two sets $\mathcal{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\mathcal{Y} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$ of independent realizations of $\mathbf{X}$ and $\mathbf{Y}$, respectively, we can choose $\mathbf{U}_i$ and $\mathbf{U}_j$ $(i \neq j = 1, \ldots, N)$ from the pooled sample $\mathcal{U} = \{\mathbf{U}_1 = \mathbf{X}_1, \ldots, \mathbf{U}_n = \mathbf{X}_n, \mathbf{U}_{n+1} = \mathbf{Y}_1, \ldots, \mathbf{U}_N = \mathbf{Y}_m\} = \mathcal{X} \cup \mathcal{Y}$ of size $N = m+n$ to construct the balls $\mathbb{B}_{ij} := \mathbb{B}(\mathbf{U}_i, \rho(\mathbf{U}_j, \mathbf{U}_i))$ and compute the differences $D_{ij} = \left|\hat{F}_{ij}(\mathbb{B}_{ij}) - \hat{G}_{ij}(\mathbb{B}_{ij})\right|$. Here $\hat{F}_{ij}$ and $\hat{G}_{ij}$ are the empirical analogs of $F$ and $G$, obtained from $\mathcal{U}$ after removing $\mathbf{U}_i$ and $\mathbf{U}_j$ from the respective samples. One can use these differences to construct a statistic $T = [N(N-1)]^{-1} \sum_{i \neq j} D_{ij}^2$ and reject the null hypothesis $H_0 : F = G$ for higher values of it. However, to reduce the computing cost, here we consider only those cases, where $\mathbf{U}_i$ and $\mathbf{U}_j$ come from the same distribution, and use the test statistic

$$T_{n,m}^{\rho} = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \left\{ \frac{1}{n-2} \sum_{k=1, k \neq i,j}^{n} \delta(\mathbf{X}_k, \mathbf{X}_j, \mathbf{X}_i) - \frac{1}{m} \sum_{k=1}^{m} \delta(\mathbf{Y}_k, \mathbf{X}_j, \mathbf{X}_i) \right\}^2$$
$$+ \frac{1}{m(m-1)} \sum_{1 \leq i \neq j \leq m} \left\{ \frac{1}{n} \sum_{k=1}^{n} \delta(\mathbf{X}_k, \mathbf{Y}_j, \mathbf{Y}_i) - \frac{1}{m-2} \sum_{k=1, k \neq i,j}^{m} \delta(\mathbf{Y}_k, \mathbf{Y}_j, \mathbf{Y}_i) \right\}^2,$$

where $\delta(\mathbf{s}, \mathbf{u}, \mathbf{v}) = \mathbb{1}\{\rho(\mathbf{s}, \mathbf{v}) \leq \rho(\mathbf{u}, \mathbf{v})\}$, and $\mathbb{1}\{\cdot\}$ is the indicator function. Pan et al. (2018) proposed a similar test statistic, where they also considered the case $i = j$, while $\mathbf{U}_i$ and $\mathbf{U}_j$ were not removed from $\mathcal{U}$ for computing

$\hat{F}(\mathbb{B}_{ij})$ and $\hat{G}(\mathbb{B}_{ij})$. One can show that $T^{\rho}_{n,m}$ is a consistent estimator (follows from Lemmas A.1 and A.2) of the ball divergence measure

$$\Theta^2_{\rho}(F,G) = \int\int \{F(\mathbb{B}(\mathbf{u},\rho(\mathbf{v},\mathbf{u}))-G(\mathbb{B}(\mathbf{u},\rho(\mathbf{v},\mathbf{u}))\}^2[dF(\mathbf{u})dF(\mathbf{v})+dG(\mathbf{u})dG(\mathbf{v})],$$

between $F$ and $G$ defined in Pan et al. (2018). Clearly, a large value of $T^{\rho}_{n,m}$ gives an evidence against $H_0 : F = G$, and we reject $H_0$ when $T^{\rho}_{n,m}$ exceeds the critical value. For a given level of significance $\alpha$ $(0 < \alpha < 1)$, this critical value (cut-off) is computed using the permutation method described below.

- Consider a permutation $\pi$ of $\{1,\ldots,N\}$ and the corresponding permutation $\mathcal{U}_{\pi} = \{\mathbf{U}_{\pi(1)},\ldots,\mathbf{U}_{\pi(N)}\}$ of the pooled data $\mathcal{U}$.
- Use $\mathcal{X}_{n,\pi} = \{\mathbf{U}_{\pi(1)},\ldots,\mathbf{U}_{\pi(n)}\}$ and $\mathcal{Y}_{m,\pi} = \{\mathbf{U}_{\pi(n+1)},\ldots,\mathbf{U}_{\pi(n+m)}\}$ as the two samples to calculate $T^{\rho}_{n,m,\pi}$, the permutation analog of $T^{\rho}_{n,m}$.
- Repeat this method for all possible permutations. If $\mathcal{S}_N$ denotes the set of all permutations of $\{1,\ldots,N\}$, the critical value is given by

$$c_{1-\alpha} = \inf\{t \in \mathbb{R} : \frac{1}{N!}\sum_{\pi\in\mathcal{S}_N} \mathbb{1}[T^{\rho}_{n,m,\pi} \le t] \ge 1-\alpha\}.$$

We reject $H_0$ if $T^{\rho}_{n,m} > c_{1-\alpha}$ or the corresponding p-value $\frac{1}{N!}\sum_{\pi\in\mathcal{S}_N} \mathbb{1}[T^{\rho}_{n,m,\pi} \ge T^{\rho}_{n,m}] < \alpha$. The following lemma shows that this permutation method leads to a valid level $\alpha$ test irrespective of the choice of $n, m$ and $d$.

**Lemma 2.1.** *Let $T_{n,m}$ be a two-sample test statistic computed based on $n$ and $m$ independent observations from $d$-dimensional distributions $F$ and*

*$G$, respectively. If $p_{n,m}$ denotes the corresponding permutation p-value, then under $H_0 : F = G$, we have $\mathbb{P}[p_{n,m} < \alpha] \leq \alpha$ irrespective of $n, m$ and $d$.*

Note that Lemma 2.1 holds for any permutation test. Interestingly, the cut-off of the proposed permutation test $c_{1-\alpha}$ can be upper bounded by a function of $n$ and $m$ that converges to zero as $n$ and $m$ diverge to infinity. This is asserted by the following lemma.

**Lemma 2.2.** *Let $\mathcal{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\mathcal{Y} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$ be two sets of independent random vectors from two $d$-dimensional distributions $F$ and $G$, respectively. For any $\alpha$ $(0 < \alpha < 1)$, the inequality $0 < c_{1-\alpha} \leq 2/\big(3\alpha(\min\{n, m\} - 2)\big)$ holds with probability one.*

Note that this upper bound of $c_{1-\alpha}$ is of the order $O(1/(\min\{n, m\}))$, and it does not depend on $d$. So, $c_{1-\alpha}$ converges to 0 as $\min\{n, m\}$ diverges to infinity. Therefore, under the alternative hypothesis $H_1 : F \neq G$, if $T^{\rho}_{m,n}$ converges to a positive constant, the power of the test converges to one. Theorem 2.1 shows this large sample consistency of the permutation test.

**Theorem 2.1.** *If $\Theta^2_{\rho}(F, G) > 0$, the power of the level $\alpha$ $(0 < \alpha < 1)$ test based on $T^{\rho}_{m,n}$ converges to 1 as $\min\{n, m\}$ increases to infinity.*

If $(R^d, \rho)$ is a finite dimensional separable metric space, $\Theta^2_{\rho}(F, G) = 0$ if and if only $F = G$. So, under $H_1 : F \neq G$, we have $\Theta^2_{\rho}(F, G) > 0$. For

computing $c_{1-\alpha}$, instead of considering all $N!$ permutations of $\{1, \ldots, N\}$, it is enough to consider all possible subsets of $\mathcal{U}$ of size $n$. But, if $n$ and $m$ are moderately large, it may not be computationally feasible to consider all $\binom{N}{n}$ subsets or all $N!$ permutations. In such cases, we generate $B$ random permutations $\pi_1, \ldots, \pi_B$ and reject $H_0$ if the corresponding p-value $p_{n,m,B} = \frac{1}{B+1}\Big\{\sum_{i=1}^{B} \mathbb{1}[T^{\rho}_{n,m,\pi_i} \geq T^{\rho}_{n,m}] + 1\Big\}$ is smaller than $\alpha$. Note that the use of all $N!$ permutations leads to the p-value $p_{n,m} = \frac{1}{N!}\Big\{\sum_{\pi \in \mathcal{S}_N} \mathbb{1}[T^{\rho}_{n,m,\pi_i} \geq T^{\rho}_{n,m}]\Big\}$. As $B$ increases, $p_{n,m,B} - p_{n,m}$ converges to 0 almost surely (see Lemma 2.3). This justifies the implementation of the test based on random permutations.

**Lemma 2.3.** *Given the pooled sample* $\mathcal{U}$, $|p_{n,m,B} - p_{n,m}| \xrightarrow{a.s.} 0$ *as* $B \to \infty$.

Though Pan et al. (2018) also suggested implementing their test using the permutation method, they proved the consistency of their test based on the large sample distribution of the test statistic. The consistency of the permutation test was missing. Moreover, they did not investigate the high-dimensional behaviour of their test.

## 3. behaviour of the proposed test in the HDLSS setup

In this section, we study the high dimensional behaviour of the test when the dimension of the data grows to infinity while the sample sizes remain fixed. The behaviour of the test may depend on the metric $\rho$. Since the

$\ell_2$ distance is arguably the most popular choice as the distance function on $\mathbb{R}^d$, we first consider the test based on this distance.

### 3.1 Test based on the $\ell_2$ distance

The test statistic based on the $\ell_2$ distance, denoted by $T^{\ell_2}_{n,m}$, is obtained replacing $\delta(\mathbf{s},\mathbf{u},\mathbf{v})$ in $T^{\rho}_{n,m}$ by $\mathbb{1}\{\|\mathbf{s}-\mathbf{v}\| \le \|\mathbf{u}-\mathbf{v}\|\}$. To investigate the behaviour of the resulting test, we consider the following assumptions.

(A1) If $\mathbf{X}_1, \mathbf{X}_2 \overset{iid}{\sim} F$ and $\mathbf{Y}_1, \mathbf{Y}_2 \overset{iid}{\sim} G$ are independent, for $\mathbf{W} = \mathbf{X}_1 - \mathbf{X}_2$, $\mathbf{X}_1 - \mathbf{Y}_1$ and $\mathbf{Y}_1 - \mathbf{Y}_2$, $\left|d^{-1}\|\mathbf{W}\|^2 - d^{-1}E(\|\mathbf{W}\|^2)\right| \xrightarrow{P} 0$ as $d \to \infty$.

(A2) There exist constants $\nu^2$, $\sigma_F^2$ and $\sigma_G^2$ such that $d^{-1}\|\boldsymbol{\mu}_F - \boldsymbol{\mu}_G\|^2 \to \nu^2$, $d^{-1}\text{trace}(\boldsymbol{\Sigma}_F) \to \sigma_F^2$ and $d^{-1}\text{trace}(\boldsymbol{\Sigma}_G) \to \sigma_G^2$ as $d \to \infty$. (Here $\boldsymbol{\mu}_F = E_F(\mathbf{X})$, $\boldsymbol{\mu}_G = E_G(\mathbf{Y})$, $\boldsymbol{\Sigma}_F = Var_F(\mathbf{X})$ and $\boldsymbol{\Sigma}_G = Var_G(\mathbf{Y})$.)

These assumptions are quite common in the HDLSS literature. While (A1) gives the weak law of large numbers (WLLN) for the sequence $\{(\mathbf{W}^{(q)})^2 : q \ge 1\}$ (i.e., $\left|\frac{1}{d}\sum_{q=1}^{d}(W^{(q)})^2 - E\frac{1}{d}\sum_{q=1}^{d}(W^{(q)})^2\right| \xrightarrow{P} 0$ as $d \to \infty$), (A2) gives the limiting value of $d^{-1}E\|\mathbf{W}\|^2$ and hence that of $d^{-1}\|\mathbf{W}\|^2$ depending on whether $\mathbf{W} = \mathbf{X}_1 - \mathbf{X}_2, \mathbf{Y}_1 - \mathbf{Y}_2$ or $\mathbf{X}_1 - \mathbf{Y}_1$. In addition to (A2), Hall et al. (2005) assumed uniformly bounded fourth moments and a $\rho$-mixing property for the component variables to investigate the high dimensional

behaviour of some popular classifiers. The weak law (A1) holds under those conditions. However, instead of $\rho$-mixing, it is enough to assume $\sum_{i\neq j} Cov(W^{(i)}, W^{(j)}) = o(d^2)$ for WLLN (Sarkar et al., 2020). As one of the reviewers pointed out, one can also assume (A2) and sufficient moment conditions like $Var(\|\mathbf{X}-\boldsymbol{\mu}_F\|^2) = o(d^2)$, $Var(\|\mathbf{Y}-\boldsymbol{\mu}_G\|^2) = o(d^2)$, $\text{trace}(\boldsymbol{\Sigma}_F^2) = o(d^2)$ and $\text{trace}(\boldsymbol{\Sigma}_G^2) = o(d^2)$. In addition to assuming uniformly bounded fourth moments and a $\rho$-mixing condition on the standardized variables, following Ahn et al. (2007); Jung and Marron (2009) one can assume a sphericity condition (i.e. $\text{trace}(\boldsymbol{\Sigma}^2)/(\text{trace}(\boldsymbol{\Sigma}))^2 \to 0$ as $d \to \infty$ both for $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_F$ and $\boldsymbol{\Sigma}_G$) for WLLN. The non-strongly spiked eigenvalue (NSSE) model (see Aoshima and Yata, 2018) (where $\lambda_{max}^2(\boldsymbol{\Sigma})/\text{trace}(\boldsymbol{\Sigma}^2) \to 0$ as $d \to \infty$) satisfies the sphericity condition. Yata and Aoshima (2012, 2020) also assumed similar conditions. Under those conditions, (A1) holds for $\mathbf{W} = \mathbf{X}_1 - \mathbf{X}_2$ and $\mathbf{W} = \mathbf{Y}_1 - \mathbf{Y}_2$. Under (A1) and (A2), we have the following lemma.

**Lemma 3.1.** *Suppose that* $\mathbf{X}_1, \mathbf{X}_2 \overset{iid}{\sim} F$ *and* $\mathbf{Y}_1, \mathbf{Y}_2 \overset{iid}{\sim} G$ *are independent. If* $F$ *and* $G$ *satisfy (A1) and (A2), then* $d^{-1/2}\|\mathbf{X}_1-\mathbf{X}_2\| \xrightarrow{P} \sigma_F\sqrt{2}$, $d^{-1/2}\|\mathbf{Y}_1-\mathbf{Y}_2\| \xrightarrow{P} \sigma_G\sqrt{2}$ *and* $d^{-1/2}\|\mathbf{X}_1 - \mathbf{Y}_1\| \xrightarrow{P} \sqrt{\sigma_G^2 + \sigma_F^2 + \nu^2}$ *as* $d \to \infty$.

Lemma 3.2 uses these distance convergence results to show that if $\nu^2 > 0$ or $\sigma_F^2 \neq \sigma_G^2$, $P(T_{n,m}^{\ell_2} > 1/3)$ converges to 1 as $d$ grows to infinity.

**Lemma 3.2.** *Assume that the two distributions $F$ and $G$ satisfy (A1)-(A2). If $\nu^2 + (\sigma_F - \sigma_G)^2 > 0$, we have $\lim_{d\to\infty} P\{T^{\ell_2}_{n,m} > 1/3\} = 1$.*

In Lemma 2.2, we have already seen that the critical value of the permutation test $c_{1-\alpha}$ is smaller than $\frac{2}{3\alpha}\big(\frac{1}{\min\{n,m\}-2}\big)$ with probability one. So, the resulting test has the high-dimensional consistency if $\min\{n, m\} - 2 > 2/\alpha$. This result is stated as Theorem 3.1 below.

**Theorem 3.1.** *Assume that $F$ and $G$ satisfy (A1)-(A2). If $\nu^2 + (\sigma_F - \sigma_G)^2 > 0$ and $\min\{n, m\} \geq 2 + 2/\alpha$, the power of the level $\alpha$ $(0 < \alpha < 1)$ test based on $T^{\ell_2}_{n,m}$ converges to one as $d$ increases to infinity.*

Theorem 3.1 gives a sufficient condition for the consistency of the proposed test in HDLSS regime. If $F$ and $G$ differ in their locations ($\nu^2 > 0$) and/or scales ($\sigma_F^2 \neq \sigma_G^2$), the test based on $T^{\ell_2}_{n,m}$ turns out to be consistent.

Now, we consider three examples each involving two multivariate normal distributions to study the empirical performance of the proposed test.

**Example 1:** Two distributions $F = \mathcal{N}_d(\mathbf{0}_d, \mathbf{I}_d)$ and $G = \mathcal{N}_d(0.15\,\mathbf{1}_d, \mathbf{I}_d)$ differ only in their means. Here $\mathbf{0}_d = (0, \ldots, 0)^\top, \mathbf{1}_d = (1, \ldots, 1)^\top \in \mathbb{R}^d$, $\mathbf{I}_d$ is the $d \times d$ identity matrix and $\mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the $d$-variate normal distribution with the mean vector $\boldsymbol{\mu}$ and the dispersion matrix $\boldsymbol{\Sigma}$.

**Example 2:** Two distributions $F = \mathcal{N}_d(\mathbf{0}_d, \mathbf{I}_d)$ and $G = \mathcal{N}_d(\mathbf{0}_d, 1.1\mathbf{I}_d)$ have the same location but they differ in their scales.

**Example 3:** Both $F = \mathcal{N}_d(\mathbf{0}_d, \mathbf{\Sigma}_{1,d})$ and $G = \mathcal{N}_d(\mathbf{0}_d, \mathbf{\Sigma}_{2,d})$ have diagonal dispersion matrices. The first $d/2$ diagonal elements of $\mathbf{\Sigma}_{i,d}$ $(i = 1, 2)$ are $i$ and the rest are $3 - i$.

For each example, we considered 10 different choices of $d$ ($d = 2^i$ for $i = 1, \ldots, 10$). In each case, we used the test based on 100 observations (50 from each distribution). This process was repeated 500 times to estimate the power of the test by the proportion of times it rejected $H_0$. In Examples 1 and 2, we have $\nu^2 + (\sigma_F - \sigma_G)^2 > 0$. So, as one would expect in view of Theorem 3.1, the power of the proposed test increased with the dimension (see Figure 1). However, in Example 3, we have $\nu^2 = 0$ and $\sigma_F = \sigma_G$. So, the sufficient conditions for consistency (see Theorem 3.1) do not hold. In this example, the proposed test had a poor performance. To overcome this limitation of the test based on the $\ell_2$ distance, in the next subsection, we use different distance functions to construct the test statistic and study the high dimensional performance of the resulting tests.

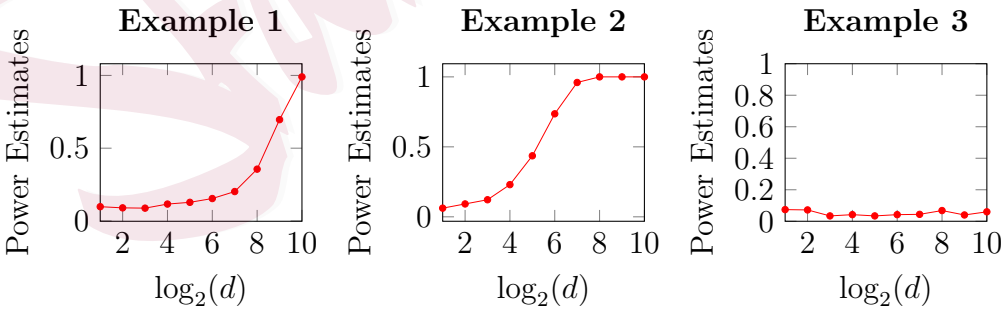


Figure 1: Power of the permutation test based on $T^{\ell_2}_{n,m}$ in Examples 1-3.

### 3.2 Tests based on generalized distances

Instead of $\ell_2$ distance, here we consider the generalized distance function proposed by Sarkar and Ghosh (2018). The generalized distance between two observations $\mathbf{x} = (x^{(1)}, \ldots, x^{(d)})^\top$ and $\mathbf{y} = (y^{(1)}, \ldots, y^{(d)})^\top$ is given by $\varphi_{h,\psi}(\mathbf{x}, \mathbf{y}) = h\big\{\frac{1}{d}\sum_{q=1}^{d} \psi(|x^{(q)} - y^{(q)}|^2)\big\}$, where $h, \psi : \mathbb{R}_+ \to \mathbb{R}_+$ are continuous, monotonically increasing functions with $h(0) = \psi(0) = 0$. Note that all $\ell_p$ distances ($p \geq 1$) are special cases of $\varphi_{h,\psi}$ (up to a multiplicative constant). Using $\varphi_{h,\psi}$, we get the generalized test statistic $T_{n,m}^{h,\psi}$ (replace $\delta(\mathbf{s}, \mathbf{u}, \mathbf{v})$ in $T_{n,m}^{\rho}$ by $\mathbb{1}\{\varphi_{h,\psi}(\mathbf{s}, \mathbf{v}) \leq \varphi_{h,\psi}(\mathbf{u}, \mathbf{v})\}$) and reject $H_0$ for large values of it. The cut-off is chosen using the permutation method as before.

High dimensional behaviour of this test can be investigated under assumptions similar to (A1)-(A2). Recall that (A1) gives WLLN for $\{(W^{(q)})^2 : \ q \geq 1\}$ with $\mathbf{W} = \mathbf{X}_1 - \mathbf{X}_2, \mathbf{X}_1 - \mathbf{Y}_1$ and $\mathbf{Y}_1 - \mathbf{Y}_2$. Here, we consider the following assumption regarding WLLN for the random variables $\{\psi(W^{(q)})^2 : \ q \geq 1\}$.

(A3) If $\mathbf{X}_1, \mathbf{X}_2 \overset{iid}{\sim} F$ and $\mathbf{Y}_1, \mathbf{Y}_2 \overset{iid}{\sim} G$ are independent, for $W = \mathbf{X}_1 - \mathbf{X}_2$, $\mathbf{X}_1 - \mathbf{Y}_1$ and $\mathbf{Y}_1 - \mathbf{Y}_2$, $(i)$ $\limsup_{d\to\infty} d^{-1}\sum_{q=1}^{d} E\psi(|W^{(q)}|^2) < \infty$ and $(ii)$ $d^{-1}\sum_{q=1}^{d}\{\psi(|W^{(q)}|^2) - E\psi(|W^{(q)}|^2)\} \xrightarrow{P} 0$ as $d \to \infty$.

Note that if the $W^{(q)}$'s are independent or $m$-dependent or they satisfy the $\rho$-mixing property, (A3) holds if the $W^{(q)}$'s have uniformly bounded second moments. If $\psi$ is bounded, the moment condition gets automatically

satisfied. Now, define $\varphi^*_{h,\psi}(F,F) = h\{d^{-1}\sum_{q=1}^{d} E\psi(|X_1^{(q)}-X_2^{(q)}|^2)\}$, $\varphi^*_{h,\psi}(G,G) = h\{d^{-1}\sum_{q=1}^{d} E\psi(|Y_1^{(q)}-Y_2^{(q)}|^2)\}$ and $\varphi^*_{h,\psi}(F,G) = h\{d^{-1}\sum_{q=1}^{d} E\psi(|X_1^{(q)}-Y_1^{(q)}|^2)\}$. There is an interesting lemma due to Sarkar and Ghosh (2018) (Lemma 1) involving these three quantities. The lemma is stated below.

**Lemma 3.3.** *Suppose that $h$ is concave and $\psi$ has a non-constant completely monotone derivative. Then for any fixed value of $d$, we have $e_{h,\psi}(F,G) = 2\varphi^*_{h,\psi}(F,G) - \varphi^*_{h,\psi}(F,F) - \varphi^*_{h,\psi}(G,G) \geq 0$, where the equality holds if and only if $F$ and $G$ have the same one-dimensional marginal distributions.*

Note that $e_{h,\psi}(F,G)$ can be viewed as an energy distance (see, e.g., Aslan and Zech, 2005) between $F$ and $G$. In view of Lemma 3.3, for appropriate choices of $h$ and $\psi$, it is somewhat reasonable to assume that under $H_1 : F \neq G$, the limiting energy distance between $F$ and $G$ remains bounded away from 0 (i.e., $\liminf_{d\to\infty} e_{h,\psi}(F,G) > 0$). Under this assumption, we can establish the consistency of the test based on $T^{h,\psi}_{m,n}$ in the HDLSS asymptotic regime. This result is given by the following theorem.

**Theorem 3.2.** *Suppose that $F$ and $G$ satisfy (A3) and $\liminf_{d\to\infty} e_{h,\psi}(F,G) > 0$. If $\min\{n,m\} \geq 2 + 2/\alpha$, the power of the level $\alpha$ $(0 < \alpha < 1)$ test based on $T^{h,\psi}_{n,m}$ increases to one as $d$ increases to infinity.*

For $h(t) = \sqrt{t}$ and $\psi(t) = t$, $\varphi_{h,\psi}$ turns out to be the $\ell_2$ metric (up to a multiplicative constant), but this choice of $\psi$ does not have a non-constant

completely monotone derivative as mentioned in Lemma 3.3. But there are other choices of $\psi$ that satisfy this property. For instance, one can use $\psi_1(t) = \sqrt{t}$, $\psi_2(t) = 1 - e^{-t/2}$ or $\psi_3(t) = \log(1+t)$ with $h(t) = t$ in all these cases. For $\psi_1$ and $\psi_2$, $\varphi_{h,\psi}$ turns out to be a distance function ($\psi_1$ leads to a scalar multiple of $\ell_1$ distance), but that is not the case for $\psi_3$. In that case, we can call it a dissimilarity measure. Note that for the $\ell_2$ distance, under (A2), we have $\liminf_{d\to\infty} e_{h,\psi}(F,G) = 2\sqrt{\nu^2 + \sigma_F^2 + \sigma_G^2} - \sigma_F\sqrt{2} - \sigma_G\sqrt{2}$, which is positive if and only if either $\nu^2 > 0$ and/or $\sigma_F \neq \sigma_G$, i.e. $\nu^2 + (\sigma_F - \sigma_G)^2 > 0$. In Example 3, we have $\nu^2 + (\sigma_F - \sigma_G)^2 = 0$ but $\liminf_{d\to\infty} e_{h,\psi}(F,G) > 0$ for $\psi = \psi_1, \psi_2, \psi_3$ with $h(t) = t$. In this example, while the test based on $T^{\ell_2}_{n,m}$ had poor performance, those based on $T^{h,\psi}_{n,m}$ with other three choices of $\psi$ (henceforth referred to as $T^{\ell_1}_{n,m}$, $T^{\exp}_{n,m}$ and $T^{\log}_{n,m}$, respectively) performed well (see Figure 2). For further investigation on the high dimensional behaviour of these tests, we consider two other examples.

**Example 4:** Both $F$ and $G$ have independent and identically distributed coordinate variables. While in $F$, they follow the standard univariate Cauchy distribution, in $G$ they have a location shift of one unit.

**Example 5:** Two distributions $F = \mathcal{N}_d(\boldsymbol{\mu}_d, \mathbf{I}_d)$ and $G = \mathcal{N}_d(-\boldsymbol{\mu}_d, \mathbf{I}_d)$ differ only in their means. Here $\|\boldsymbol{\mu}_d\| = \|d^{-1/2}\mathbf{1}_d\| = 1$ for all values of $d$.

Here also, we generated 50 observations from each distribution, and the

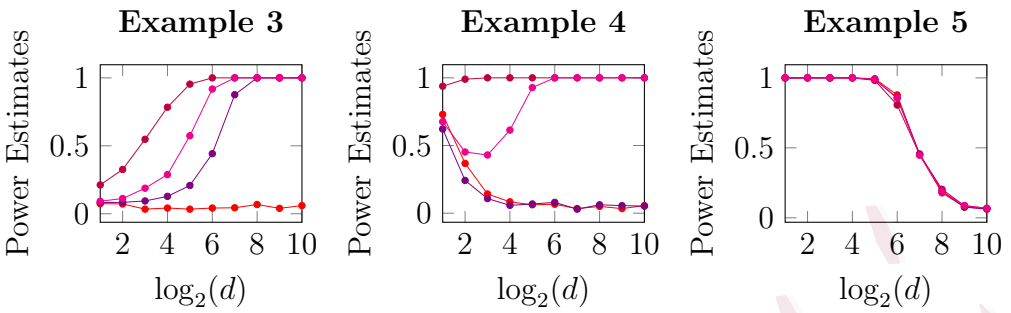


Figure 2: Powers of BD-$\ell_2$(●) BD-$\ell_1$(●) BD-exp(●) and BD-log(●) in Ex. 3-5.

process was repeated 500 times to estimate the power of the tests.

In Example 4, the ball divergence tests based on the test statistics $T^{\ell_1}_{n,m}$ and $T^{\ell_2}_{n,m}$ (henceforth referred to as BD-$\ell_1$ and BD-$\ell_2$ tests) did not work well, but those based on $T^{\exp}_{n,m}$ and $T^{\log}_{n,m}$ (henceforth referred to as BD-exp and BD-log tests) had excellent performance (see Figure 2). Among them, BD-exp had an edge. Note that in this example, the coordinate variables in $F$ and $G$ do not have finite moments. That affected the performance of BD-$\ell_1$ and BD-$\ell_2$. Since $\psi_2(t) = 1 - e^{-t/2}$ is a bounded function, we do not have such problems for BD-exp. Assumption (A3) holds for this choice of $\psi$. It holds for $\psi_3(t) = \log(1+t)$ as well. This was the reason behind the good performance of the tests based on these two choices of $\psi$.

In Example 5, the Mahalanobis distance between $F$ and $G$ remains the same for all values of $d$, and we have $\lim_{d\to\infty} e_{h,\psi}(F,G) = 0$ for all choices of $h$ and $\psi$ considered here. So, as expected, the powers of all tests gradually

dropped as $d$ increased. In such situations, for good performance of the proposed tests, we need to increase the sample sizes appropriately with the dimension. We consider such cases in the next section.

## 4. What happens if the sample sizes increase with the dimension?

In this section, we deal with the cases, where $F$ and $G$ gradually become close as $d$ increases. For such shrinking alternatives, the power of any test based on fixed sample sizes is expected to go down as $d$ increases. Therefore, to achieve better performance in high-dimension, one needs to increase the sample sizes as well. Now, one may be curious to know whether it is possible to increase $n$ and $m$ with $d$ at an appropriate rate such that one can construct a valid level $\alpha$ $(0 < \alpha < 1)$ test with power converging to 1 as the dimension increases. We shall show that this is possible for our tests as long as $\Theta^2_\rho(F, G)$ shirks to 0 at an appropriately slower rate. This is obtained by establishing the minimax rate optimality of our tests.

### 4.1 Minimax rate optimality

Consider the hypotheses $H_0' : \Theta^2_{\ell_2}(F, G) = 0$ and $H_1' : \Theta^2_{\ell_2}(F, G) > \epsilon$ for some $\epsilon > 0$. Define $\mathbb{P}^{(n,m)}_{F,G}$ as the joint distribution of $\mathbf{X}_1, \ldots, \mathbf{X}_n$, $\mathbf{Y_1}, \ldots, \mathbf{Y}_m$, where $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{iid}{\sim} F$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_m \overset{iid}{\sim} G$. Let $\mathcal{F}(\epsilon) := \{(F, G) \mid \Theta^2_{\ell_2}(F, G) > \epsilon\}$ denote the class of alternatives, and for a given

$\alpha \in (0,1)$, $\mathbb{T}_{n,m,d}(\alpha)$ denote the class of all level $\alpha$ test functions $\phi : \mathcal{U} \to \{0,1\}$. The minimax type II error rate for this class is defined as

$$R_{n,m,d}(\epsilon) = \inf_{\phi \in \mathbb{T}_{n,m,d}(\alpha)} \sup_{(F,G)\in\mathcal{F}(\epsilon)} \mathbb{P}_{F,G}^{(n,m)}(\phi = 0).$$

We want to find an $\epsilon_0 = \epsilon_0(n,m,d)$, for which the following conditions hold.

(a) For any $0 < \zeta < 1-\alpha$, there exists a constant $c(\alpha,\zeta) > 0$ such that for all $0 < c < c(\alpha,\zeta)$, we have $\liminf_{n,m,d\to\infty} R_{n,m,d}(c\ \epsilon_0(n,m,d)) \geq \zeta$ .

(b) There exists a level $\alpha$ test $\phi_0$ such that for any $0 < \zeta < 1-\alpha$, we can find $C(\alpha,\zeta) > 0$ for which $\limsup_{n,m,d\to\infty} \sup_{(F,G)\in\mathcal{F}(c\ \epsilon_0(n,m,d))} \mathbb{P}_{F,G}^{(n,m)}\{\phi_0 = 0\} \leq \zeta\ \forall c > C(\alpha,\zeta)$, i.e., $\limsup_{n,m,d\to\infty} R_{n,m,d}(c\ \epsilon_0(n,m,d)) \leq \zeta\ \forall c > C(\alpha,\zeta)$.

The rate $\epsilon_0(n,m,d)$ (which is unique up to a constant) is called the minimax rate of separation, and $\phi_0$ is called the minimax rate optimal test. Theorem 4.1 shows that if $\epsilon$ is of smaller order than $(1/\sqrt{n} + 1/\sqrt{m})^2$, for all level $\alpha$ tests the maximum type II error is bounded away from 0.

**Theorem 4.1.** *For $0 < \zeta < 1-\alpha$, there exists a constant $c_0(\alpha,\zeta)$ such that for $\lambda(n,m) = (1/\sqrt{n}+1/\sqrt{m})^2$, the minimax type II error $R_{n,m,d}(c\lambda(n,m))$ is lower bounded by $\zeta$ for all $0 < c < c_0(\alpha,\zeta)$.*

The above result shows that the minimax rate of separation cannot be of order smaller than $(1/\sqrt{n}+1/\sqrt{m})^2$. Now, we show that for $\epsilon_0(n,m,d) = \lambda(n,m)$, the test based on $T_{n,m}^{\ell_2}$ satisfies the condition (b) stated above.

**Theorem 4.2.** *For $0 < \zeta < 1 - \alpha$, there exists a constant $C_0(\alpha, \zeta)$ such that asymptotically the maximum type II error of the test based on $T^{\ell_2}_{n,m}$ over $\mathcal{F}(c\lambda(n,m))$ is uniformly bounded above by $\zeta$ for all $c > C_0(\alpha, \zeta)$, i.e.,*

$$\limsup_{n,m,d\to\infty} \sup_{(F,G)\in\mathcal{F}(c\lambda(n,m))} P^{(n,m)}_{F,G}(T_{n,m} \le c_{1-\alpha}) \le \zeta \quad \text{for all } c > C_0(\alpha,\zeta).$$

Theorems 4.1 and 4.2 together show that the minimax rate of separation $\epsilon_0(n,m,d) = (1/\sqrt{n}+1/\sqrt{m})^2$ does not depend on the dimension, and they also establish the minimax rate optimality of the permutation test based on $T^{\ell_2}_{n,m}$ for the class of alternatives $\mathcal{F}(\epsilon)$.

## 4.2 Performance under shrinking alternatives

Theorem 4.2 gives us a lower bound $\lambda(n,m)$ on the rate of $\Theta^2_{\ell_2}(F,G)$ that enables us to detect the difference between $F$ and $G$ using the permutation test based on $T^{\ell_2}_{n,m}$. If we increase the sample sizes with the dimension such that $\lambda(n,m)$ converges to 0 at a faster rate than $\Theta^2_{\ell_2}(F,G)$ (i.e., $\Theta^2_{\ell_2}(F,G)/\lambda(n,m) \to \infty$ as $d \to \infty$), the test based on $T^{\ell_2}_{n,m}$ turns out to be consistent. This result is asserted by the following theorem.

**Theorem 4.3.** *If $n$ and $m$, the sample sizes from $F$ and $G$, grow as a function of the dimension $d$ in such a way that $\lim_{d\to\infty} \Theta^2_{\ell_2}(F,G)/\lambda(n,m) = \infty$, then, the power of the level $\alpha$ $(0 < \alpha < 1)$ test based on $T^{\ell_2}_{n,m}$ converges to one as dimension increases to infinity.*

If $\liminf_{d\to\infty} \Theta^2_{\ell_2}(F,G) > 0$, the assumption in Theorem 4.3 holds even if $n$ and $m$ grow very slowly with $d$. In such cases, one can expect good results even in the HDLSS setup, and we have observed the same in our numerical studies. Recall that if $F$ and $G$ satisfy the conditions of Theorem 3.1, we have $\liminf_{d\to\infty} \Theta^2_{\ell_2}(F,G) \geq 1/3$. As expected, in such cases we have the consistency of the test when $m$ and $n$ also increase with $d$.

So far, we have discussed the minimax rate of optimality of the test based on $T^{\ell_2}_{n,m}$ and established its consistency for shrinking alternatives. The results similar to Theorems 4.1-4.3 hold even when the test is constructed based on other distance functions considered in Section 3. We state the result below as Theorem 4.4, but we skip the details of the proof since they are exactly the same as in the case of the test based on the $\ell_2$ distance.

**Theorem 4.4.** *Let $h, \psi : \mathbb{R}_+ \to \mathbb{R}_+$ be continuous, monotonically increasing functions with $h(0) = \psi(0) = 0$. Assume that $n$ and $m$, the sample sizes from $F$ and $G$, grow as functions of the dimension $d$ in such a way that $\lim_{d\to\infty} \Theta^2_{\varphi_{h,\psi}}(F,G)/\lambda(n,m) = \infty$. Then the power of the level $\alpha$ $(0 < \alpha < 1)$ test based on $T^{h,\psi}_{n,m}$ converges to one as $d$ increases to infinity.*

**Remark 1.** *In the HDLSS setup, we need Assumptions (A1)-(A2) for the consistency of the tests. But when $n$ and $m$ grow with $d$, we do not need such assumptions. Also, unlike the HDLSS setup, here we have consistency*

*for the test based on the $\ell_p$-distance for all $p \geq 1$.*

**Remark 2.** *Theorems 4.3 and 4.4 remain silent about the asymptotic behaviour of the proposed test when $\lim_{d\to\infty} \Theta^2_{\ell_2}(F,G)/\lambda(n,m) = c$ (or $\lim_{d\to\infty} \Theta^2_{h,\psi}(F,G)/\lambda(n,m) = c$) for some $c \in (0,\infty)$. However, in such cases, one can show that the asymptotic power of the test has a lower bound $1 - (C_1 c + C_2)/\big(c - \frac{1}{3\alpha}\big)^2$, where $C_1$ and $C_2$ are two universal constants (see the proof of Theorem 4.2).*

Now, consider a simple example involving two multivariate normal distributions $F = \prod_{i=1}^d \mathcal{N}_1(1/d^\beta, 1)$ and $G = \prod_{i=1}^d \mathcal{N}_1(-1/d^\beta, 1)$, where $\beta$ is a positive constant. Note that as $d$ grows to infinity, here we have $\nu^2 + (\sigma_F - \sigma_G)^2 = 0$ and $\lim_{d\to\infty} e_{h,\psi}(F,G) = 0$ for all $h$ and $\psi$ considered in this article. So, the conditions for the HDLSS consistency of the tests are not satisfied. Now, we study the behaviour of the BD-$\ell_2$ test when the sample sizes increase with the dimension at the rate $O(d^\gamma)$ for some $\gamma > 0$. We find out the relation between $\gamma$ and $\beta$ that leads to the consistency of the test. Our findings are summarized in the following proposition.

**Proposition 4.1.** *Suppose that $n$ and $m$ are the sample sizes from $F = \prod_{i=1}^d \mathcal{N}_1(1/d^\beta, 1)$ and $G = \prod_{i=1}^d \mathcal{N}_1(-1/d^\beta, 1)$, respectively. If $\beta > 0$ and $n \asymp m \asymp d^\gamma$ for some $\gamma > 0$, then, for the ball divergence test based on $T^{\ell_2}_{n,m}$, we have the following results.*

*(a) If $\beta \leq 1/4$, for any $\gamma > 0$, the test is consistent.*

*(b) If* $1/4 < \beta \leq 1/2$*, the test is consistent if* $\gamma > 4\beta - 1$*.*

*(c) If* $\beta > 1/2$*, the test is consistent if* $\gamma > 2\beta$*. If* $\gamma < 2\beta - 1$*, there exist no level* $\alpha$ $(0 < \alpha < 1)$ *tests with asymptotic power more than* $\alpha$*.*

Proposition 4.1(c) says that if $\beta > 1/2$, for the consistency of any test, one needs to increase the sample size at a rate faster than $O(d^{2\beta-1})$. So, the HDLSS consistency is not possible in this case. Recall that in Example 5, we have $\beta = 1/2$. Therefore, if we increase $n$ and $m$ at a rate faster than $O(d)$, our test will be consistent. We confirmed it in our numerical study.

For this study, we used 3 different choices of $\beta$ $(0.2, 0.3$ and $0.5)$, and in each case, 7 different choices of $\gamma$ $(0, 0.4, 0.5, 0.6, 0.9.1$ and $1.1)$ and 10 different values of $d$ ($2^i$ for $i = 1, \ldots, 10$) were considered. We took $n = m = 5 + \lfloor d^\gamma \rfloor$ to ensure $n, m \geq 5$, and each experiment was repeated 500 times to compute the power of the test based on $T^{\ell_2}_{n,m}$. The results are reported in Figure 3. In this example, for higher values of $\beta$, $\Theta^2_{\ell_2}(F, G)$ converges to zero at a faster rate. Therefore, to discriminate between $F$ and $G$, we need to increase the sample sizes at a higher rate as well. Figure 3 shows that for higher values of $\beta$, the tests corresponding to lower values of $\gamma$ performed poorly. Note that here $\gamma = 0$ represents the HDLSS scenario. We can see that for $\beta = 0.2$, even for $m = n = 6$ (i.e., $\gamma = 0$), the power of our test converged to 1 in high dimensions. This was expected in view of

Proposition 4.1(a). As expected, the test had higher power for larger values of $\gamma$. For $\beta = 0.3$ and 0.5, it did not work well in the HDLSS setup, but when $m$ and $n$ increased with $d$ at an appropriate rate, the power of the test converged to unity as we expect in view of Proposition 4.1(b)-(c).

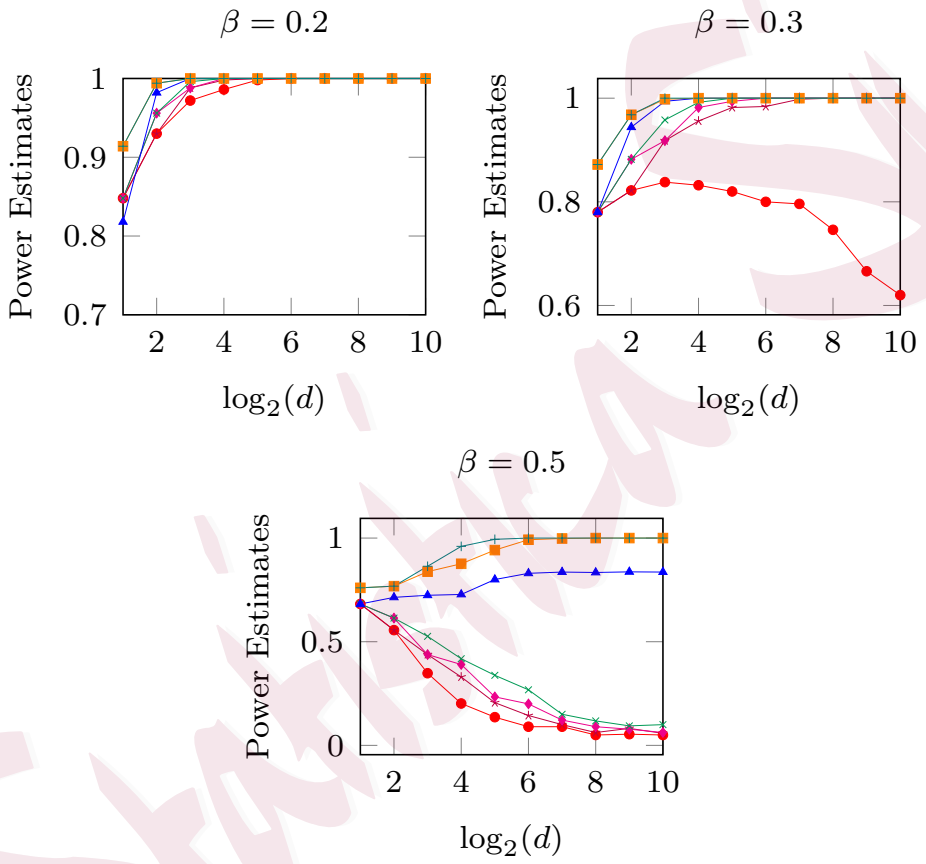


Figure 3: Powers of the BD-$\ell_2$ test for different choice of $\beta$ $(0.2, 0.3$ and $0.5)$ and $\gamma$ (0 (●), 0.4 (⋆), 0.5 (♦), 0.6 (×), 0.9 (▲), 1 (■), 1.1 (+)).

## 5. Empirical performance of the proposed tests

In this section, we compare the empirical performance of our tests with some popular tests. For this comparison, we consider the multivariate run tests based on minimum spanning tree (Friedman and Rafsky, 1979) and shortest Hamiltonian path (Biswas et al., 2014), the tests based on averages of inter-point distances proposed by Baringhaus and Franz (2004) and Biswas and Ghosh (2014), the nearest neighbor test (Schilling, 1986; Henze, 1988), and the test based on maximum mean discrepancy (Gretton et al., 2012). Henceforth, we shall refer to them as the FR test, the SHP test, the BF test, the BG test, the NN test, and the MMD test, respectively. For the NN test, we consider the test based on 3 neighbors, which has been reported to perform well (see, e.g., Schilling, 1986). Throughout this article, all tests are considered to have the 5% nominal level. The SHP test has the distribution-free property. For all other tests, the cut-off is computed based on 500 random permutations. Note that these permutation tests satisfy the level properties as mentioned in Lemma 2.1.

### 5.1 Analysis of simulated data sets

First, we study the level properties of our tests. We generated two sets of independent observations from $N_d(\mathbf{0}_d, \mathbf{I}_d)$ and used them as observations

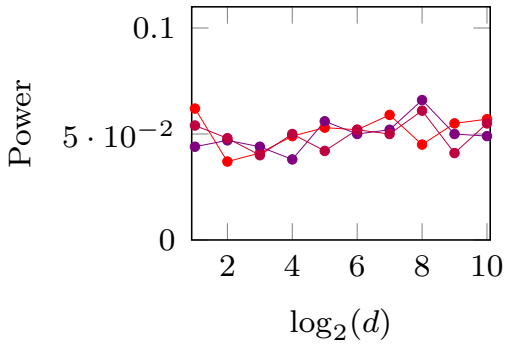


Figure 4: Observed levels of BD-$\ell_2$ test for $n = m = 20$ (•), 35 (•) and 50 (•).

from $F$ and $G$, respectively. This experiment was repeated 500 times, and for each test, we computed the proportion of times it rejected $H_0$. We carried out our experiment for different sample sizes ($n = m = 20$, 35 and 50) and dimension ($d = 2^i$ for $i = 1, 2, \ldots, 10$). Figure 4 shows that on all occasions, the BD-$\ell_2$ test rejected $H_0$ in nearly 5% of the cases. BD-$\ell_1$, BD-exp, BD-log and other competing tests also exhibited similar level properties. But, to save space, we decided not to report them.

Next, we investigate the power properties of the proposed tests. We consider two types of examples. In Section 5.1.1, we deal with examples with fixed $n$ and $m$, and study the performance of different tests as the $d$ increases. In Section 5.1.2, we consider the situations, where the conditions for HDLSS consistency of the proposed tests do not hold (i.e., we have $\nu^2 + (\sigma_F - \sigma_G)^2 = 0$ and $\lim e_{h,\psi}(F, G) = 0$). In such cases, we investigate the performance of different tests when $n$ and $m$ grow with $d$.

### 5.1.1 Dimension increases when the sample sizes remain fixed

We begin with the four examples discussed in Sections 3.1 and 3.2, and the powers of the proposed and other competing tests are reported in Figure 5.

In the location problem in Example 1, BF and MMD tests outperformed all other tests considered here. However, the powers of the proposed ball divergence tests were comparable to the rest (BG, NN, FR, and SHP tests).

In Example 2, all tests based on ball divergence and the BG test had similar performance, and they performed much better than their competitors. Among the rest, the SHP test had a relatively higher power. In high dimensions, FR and NN tests had powers close to 0. Biswas et al. (2014) and Mondal et al. (2015) explained the reasons for such poor performance of FR and NN tests in high dimensional scale problems.

In Example-3, we have $\nu^2+(\sigma_F-\sigma_G)^2 = 0$ but $\liminf_{d\to\infty} e_{h,\psi}(F,G) > 0$ for $\psi = \psi_1, \psi_2, \psi_3$ with $h(t) = t$. So, as expected, BD-$\ell_2$ did not have satisfactory performance, but BD-$\ell_1$, BD-exp and BD-log performed well in high dimensions. Among them, BD-exp had a clear edge. Unlike these three tests, the powers of other competing methods did not increase with the dimension. Note that these competing methods are based on $\ell_2$ distances. The use of a different distance function may improve their performance.

In the presence of heavy-tailed distributions in Example 4, all tests

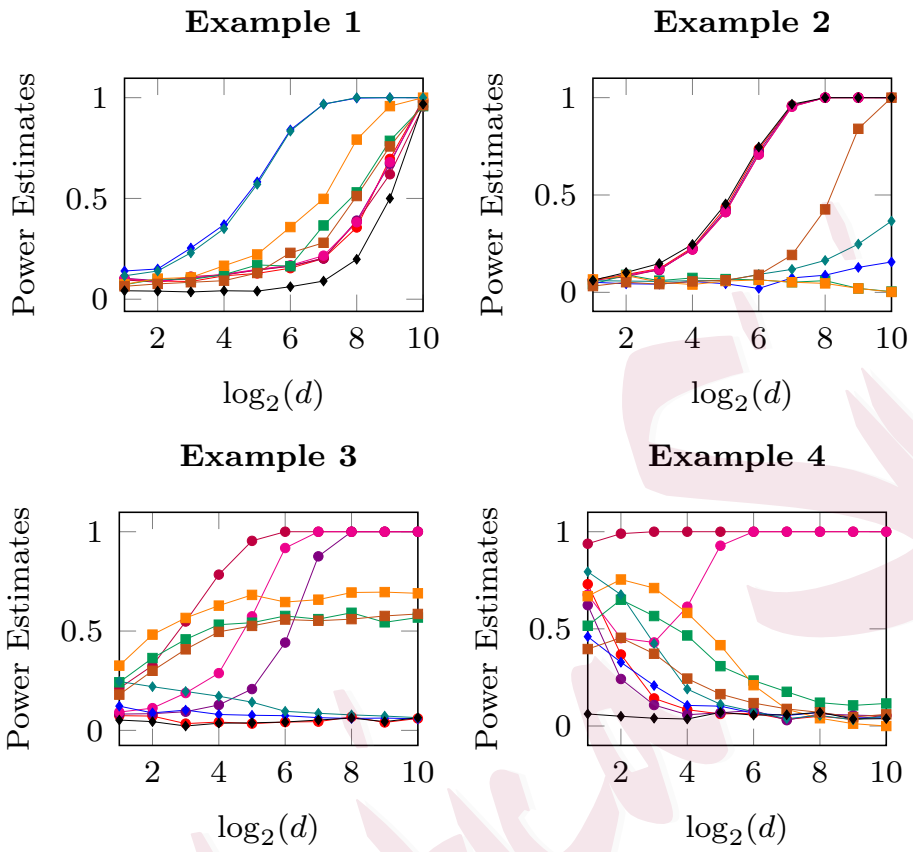


Figure 5: Powers of BD-$\ell_2$ (●), BD-$\ell_1$ (●), BD-exp (●), BD-log (●), FR (■), BF (◆), NN (■), MMD (◆), SHP (■), BG (◆) tests in Examples 1-4.

except BD-exp and BD-log had poor performance in high dimensions. Among these two tests, the one based on the bounded $\psi$-function (i.e., $\psi_2(t) = 1 - e^{-t/2}$) performed better. We also observed this in Section 3.2.

As we have discussed before, in Example 5, the power of any test based on fixed sample sizes is expected to decrease as the dimension increases. We also observed the same for all tests considered here. So, we do not report

those results. Instead, we consider three other examples (Examples 6-8).

**Example 6:** $F$ is the $d$-variate standard normal distribution while $G$ is an equal mixture of $\mathcal{N}_d(0.5\,\mathbf{1}_d, \mathbf{I}_d)$ and $\mathcal{N}_d(-0.5\,\mathbf{1}_d, \mathbf{I}_d)$.

**Example 7:** $F$ is same as in Example 6 but $G$ is mixture of $\mathcal{N}_d(\mathbf{1}_d, \mathbf{I}_d)$ and $\mathcal{N}_d(-0.25\,\mathbf{1}_d, \mathbf{I}_d)$ with mixing proportions 0.2 and 0.8, respectively.

**Example 8:** Both $F$ and $G$ have independent and identically distributed coordinate variables. The coordinate variables in $F$ follow $\mathcal{N}_1(0, 2)$ distribution, but in $G$, they follow the standard $t$ distribution with 4 degrees of freedom.

For each example, we generated 50 observations from each distribution and considered 10 different choices of $d$ ($2^i$ for $i = 1, \ldots, 10$) as before. Each experiment was repeated 500 times to estimate the power of different tests, and they are shown in Figure 6. This figure clearly shows that in the examples involving mixture normal distributions, the BG test, and the ball divergence tests performed better than their competitors. In Example 8, the ball divergence tests outperformed all other competing tests considered here. Like Example 3, here also, we have $\nu^2 + (\sigma_F - \sigma_G)^2 = 0$ but $\liminf_{d\to\infty} e_{h,\psi}(F, G) > 0$ for other three choices of $\psi$. So, as expected, BD-$\ell_1$, BD-exp and BD-log performed much better than BD-$\ell_2$.

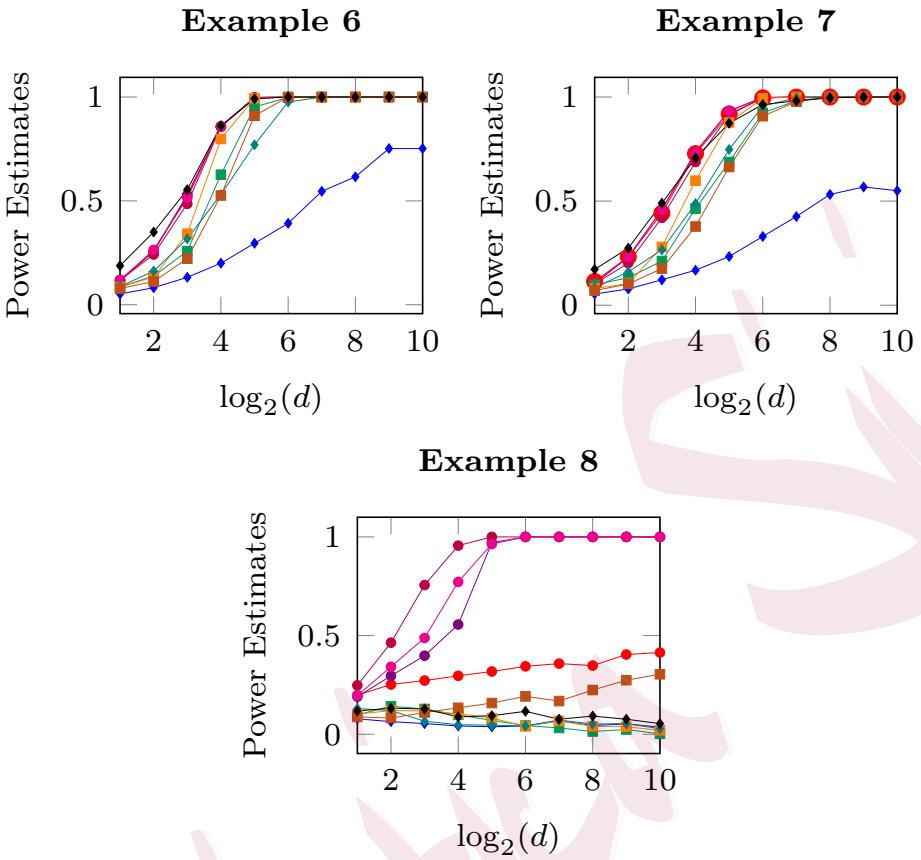


Figure 6: Powers of BD-$\ell_2$ (●), BD-$\ell_1$ (●), BD-exp (●), BD-log (●), FR (■), BF (◆), NN (■), MMD (◆), SHP (■), BG (◆) tests in Examples 6-8.

### 5.1.2 Sample sizes grow with the dimension

In this section, we deal with some examples, where we do not have a theoretical guarantee for the consistency of the ball divergence tests in the HDLSS regime, and we investigate how the proposed tests and their competitors perform when the sample sizes also grow with the dimension.

As before, the powers of all tests are computed based on 500 replications.

We consider six examples in this section. Descriptions of the first three examples are given below. In Example-9, we consider the sample sizes $n = m = 5 + \lfloor \sqrt{d} \rfloor$ while in Examples 10 and 11, we have $n = m = d + 5$.

**Example 9:** The coordinate variables in $F$ and $G$ are independent and identically distributed as $\mathcal{N}_1(d^{-0.3}, 1)$ and $\mathcal{N}_1(-d^{-0.3}, 1)$, respectively.

**Example 10:** Both $F = \mathcal{N}_d(\mathbf{0}_d, \mathbf{\Sigma}^{\circ}_{1,d})$ and $G = \mathcal{N}_d(\mathbf{0}_d, \mathbf{\Sigma}^{\circ}_{2,d})$ have the same mean $\mathbf{0}_d$, but different diagonal dispersion matrices. The first $d/2$ diagonal elements of $\mathbf{\Sigma}^{\circ}_{1,d}$ are 1, and the rest are 5. On the contrary, $\mathbf{\Sigma}^{\circ}_{2,d}$ has the first $d/2$ diagonal elements equal to 5 and rest equal to 1.

**Example 11:** $F = \mathcal{N}_d(\mathbf{0}_d, \mathbf{\Sigma}^{*}_{1,d})$ and $G = \mathcal{N}_d(\mathbf{0}_d, \mathbf{\Sigma}^{*}_{2,d})$ have the same mean but different dispersion matrices $\mathbf{\Sigma}^{*}_{1,d} = ((0.1^{|i-j|}))$ and $\mathbf{\Sigma}^{*}_{2,d} = ((0.5^{|i-j|}))$.

In the location problem in Example 9, BF and MMD tests had the best performance (see Figure 7) closely followed by NN, BG, and ball divergence tests. The SHP test had relatively low power. Interestingly, the powers of all competing tests converged to unity as the dimension increased.

Example 10 is similar to Example 3 though the parameters are different. In this example, NN, FR, and BD-exp tests had better performance than their competitors, with the BD-exp test having an edge (see Figure 7). Interestingly, the powers of all tests barring the BG test showed a tendency

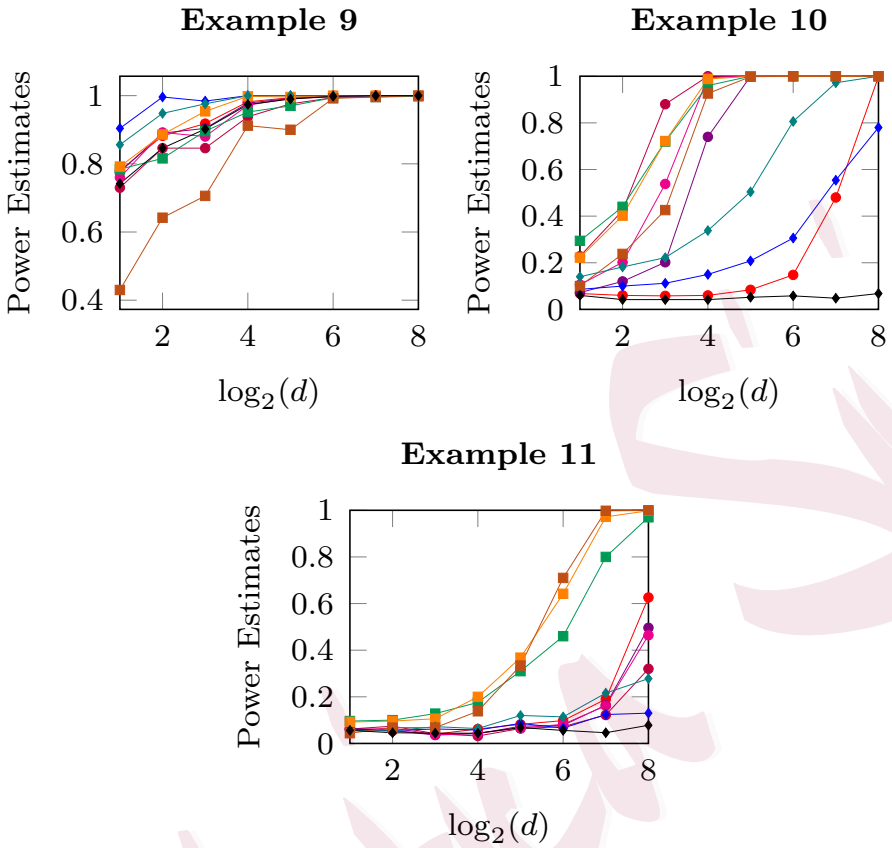


Figure 7: Powers of BD-$\ell_2$ (●), BD-$\ell_1$ (●), BD-exp (●), BD-log (●), FR (■), BF (◆), NN (■), MMD (◆), SHP (■), BG (◆) tests in Examples 9-11.

to converge to unity as the dimension increased. This was not the case in Example 3 when samples of fixed sizes were used.

In Example 11, two distributions have the same mean and marginal variances but they differ in their correlation structures. Here we have $\nu^2 + (\sigma_F - \sigma_G)^2 = 0$ and $\lim_{d\to\infty} e_{h,\phi}(F, G) = 0$ for all three choices of $\psi$ (i.e., $\psi_1$, $\psi_2$ and $\psi_3$). In this example, all tests had poor performance in the HDLSS

setup, but that was not the case when the sample sizes increased with the dimension. Here graph-based tests performed much better than average distance-based tests. However, unlike BG, BF, and MMD tests, the powers of the ball divergence tests had a sharp rise in higher dimensions.

Finally, we consider three examples involving sparse alternatives, where $F$ and $G$ differ only in $\lfloor d^{\beta} \rfloor$ many coordinates for $\beta \in (0, 1)$. Clearly, in these examples, we have $e_{h,\psi}(F, G) \asymp d^{\beta-1}$ for all choices of $\psi$ considered in this article. For our numerical study, we use $\beta = 0.7$ and $n = m = 5 + \lfloor \sqrt{d} \rfloor$.

**Example 12:** $F = \mathcal{N}_d(\boldsymbol{\mu}_d, \mathbf{I}_d)$ and $G = \mathcal{N}_d(\mathbf{0}_d, \mathbf{I}_d)$ differ only in their locations. The first $[d^{\beta}]$ coordinates of $\boldsymbol{\mu}_d$ are 2, and the rest are zero.

**Example 13:** Two distributions $F = \mathcal{N}_d(\mathbf{0}_d, \mathbf{I}_d)$ and $G = \mathcal{N}_d(\mathbf{0}_d, \boldsymbol{\Sigma}_d)$ differ only in their scales. Here $\boldsymbol{\Sigma}_d$ is a diagonal matrix with first $[d^{\beta}]$ entries equal to 5, and the rest equal to 1.

**Example 14:** The distribution $G$ differs from $F = \mathcal{N}_d(\mathbf{0}_d, 2\mathbf{I}_d)$ only in the first $[d^{\beta}]$ many coordinates. These coordinate variables are independent and they follow $t$ distribution with 4 degrees of freedom.

Figure 8 shows the powers of different tests in these three examples. In the location and scale problems in Examples 12 and 13, our findings were similar to those observed in Examples 1 and 2, respectively. In Example 12, BF and MMD tests had an edge, but the performances of the proposed tests

were competitive with the rest. The SHP test had relatively low power. In Example 13, the BG test and the ball divergence tests outperformed their competitors in higher dimensions. In this scale problem, FR and NN tests had poor performance. Example 14 can be viewed as a sparse version of Example 8, where the two distributions differ only in their shapes. In this example, all tests based on the $\ell_2$ distance failed to perform well but the ball

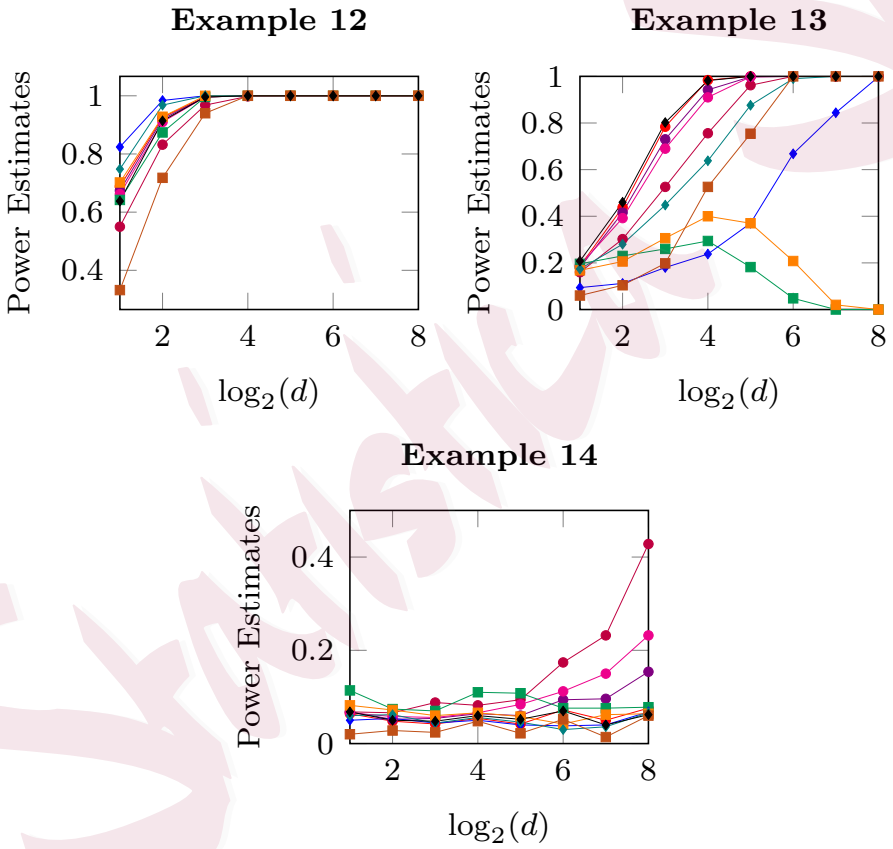


Figure 8: Powers of BD-$\ell_2$ (●), BD-$\ell_1$ (●), BD-exp (●), BD-log (●), FR (■), BF (◆), NN (■), MMD (◆), SHP (■), BG (◆) tests in Examples 12-14.

divergence tests based on other distance functions performed better. The powers of these tests showed an upward trend with increasing dimension.

## 5.2 Analysis of benchmark data sets

For further evaluation of the performance of different tests, we analyzed two real data sets, namely the Colon data and the Lightning-2 data. Colon data set contains expression levels of 2000 genes in 40 'tumor' and 22 'normal' colon tissue samples that were analyzed with an Affymetrix oligonucleotide array. This data set is available in the R package 'rda' and its description can be found in Alon et al. (1999). Lightning 2 data set contains 637-dimensional observations from two populations with respective sample sizes 48 and 73. It is available at the UCR Time Series Classification Archive (https://www.cs.ucr.edu/ eamonn/time_series_data _2018/), and its description can also be found in Sarkar et al. (2020). These two data sets have been extensively studied in the classification literature, and it is well known that in each of these data sets, there is a reasonable separation between the two distributions. So, we can assume the alternative hypothesis $H_1 : F \neq G$ to be true, and different tests can be compared based on their powers. However, when we used the full data set for testing, all tests rejected $H_0$ both in Colon and Lightning-2 data. Using that single experiment based on

the full data set, it was not possible to compare different test procedures. So, we generated random sub-samples from the entire data set, keeping the sample proportions from the two distributions approximately the same as they were in the original data. Different tests were implemented using these sub-samples, and this procedure was repeated 500 times to estimate their powers. The results for different sub-sample sizes are reported in Figure 9.

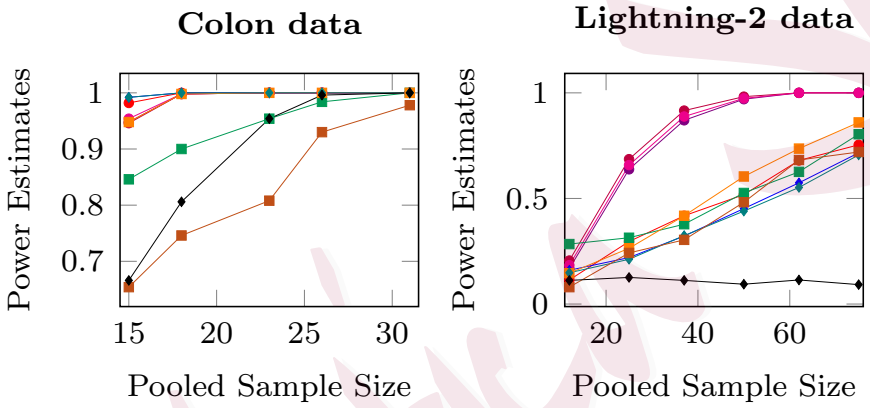


Figure 9: Powers of BD-$\ell_2$ (●), BD-$\ell_1$ (●), BD-exp (●), BD-log (●), FR (■), BF(◆), NN(■), MMD(◆), SHP(■), BG(◆) tests in benchmark data sets.

In the case of Colon data, BF, MMD, and BD-$\ell_2$ tests had very high power even when the pooled sample size was 15. These three tests had comparable performance, and they performed better than others. BD-$\ell_1$, BD-exp and BD-log tests also had competitive performances. Like BF, MMD, and NN tests, they also had unit power for samples of size 18 or higher. FR, BG, and SHP tests had relatively low powers in this data set.

Figure 9 clearly shows the superiority of the ball divergence tests BD-$\ell_1$, BD-exp and BD-log in the case of Lightning-2 data. These three tests had much higher powers compared to the rest for samples with a combined sample size larger than 20. Among the other methods, the NN test had the best overall performance. BF, MMD, FR, SHP, and BD-$\ell_2$ tests also had similar powers. The BG test performed poorly in this data set.

## 6. Concluding remarks

This article investigates high dimensional behaviour of some tests based on ball divergence and proves their consistency not only in the HDLSS regime but also in situations when the sample sizes increase with the dimension. If the distributions have exponential tails and they differ in their locations and/or scales, the tests based on $\ell_2$ and $\ell_1$ distance usually perform better. But in the case of heavy-tailed distributions, we recommend using tests based on a bounded $\psi$ function. Unlike many existing tests, the ball divergence tests based on suitable choices of $\psi$ (e.g., $\psi_1$, $\psi_2$ and $\psi_3$) can discriminate between two high-dimensional distributions differing outside the first two moments. We have proved the minimax rate optimality of the proposed tests, and that helped us to establish their consistency even for shrinking alternatives. Analyzing several simulated and real data sets, we

have shown that the proposed tests can outperform the start-of-the-art tests in a wide variety of high-dimensional two-sample problems. All simulated examples considered in this article belong to the NSSE model (see Aoshima and Yata, 2018). But the proposed tests may have good power for some SSE models as well (see Appendix B in the supplementary document).

The proposed tests can be generalized to multivariate $k$-sample problems. The variance of the probability measures of a ball corresponding to different distributions $F_1, \ldots, F_k$ gives us some idea about the difference among the $F_i$'s. The average of this variance over balls with random centre and radius can be used as a generalized ball divergence measure. A suitable estimate of this measure can be used to construct a $k$-sample test.

Recently, Pan et al. (2020) used the notion of ball divergence to develop a measure of dependence among several Banach-valued random variables. This measure, known as the ball covariance, was used for testing independence among those variables. The theory presented in this article can be used to study the high dimensional behaviour of that test, particularly when the sample sizes increase with the dimension. However, it is not clear whether the test is minimax rate optimal. This can be investigated in a future work.

## Supplementary Materials

The proofs of the theorems, lemmas and propositions are included in the supplementary material. It also contains some additional numerical results.

## References

Ahn, J., J. Marron, K. M. Muller, and Y.-Y. Chi (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika 94*(3), 760–766.

Alon, U., N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. U.S.A. 96*(12), 6745–6750.

Aoshima, M. and K. Yata (2018). Two-sample tests for high-dimension, strongly spiked eigenvalue models. *Statistica Sinica 28*(1), 43–62.

Aslan, B. and G. Zech (2005). New test for the multivariate two-sample problem based on the concept of minimum energy. *J. Stat. Comput. Simul. 75*(2), 109–119.

Bai, Z. and H. Saranadasa (1996). Effect of high dimension: by an example of a two sample problem. *Statist. Sinica 6*(2), 311–329.

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

Bhattacharya, B. B. (2019). A general asymptotic framework for distribution-free graph-based

two-sample tests. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 81*(3), 575–602.

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

Biswas, M., M. Mukhopadhyay, and A. K. Ghosh (2014). A distribution-free two-sample run test applicable to high-dimensional data. *Biometrika 101*(4), 913–926.

Cai, T., W. Liu, and Y. Xia (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *J. Amer. Statist. Assoc. 108*(501), 265–277.

Chen, S. X. and Y.-L. Qin (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist. 38*(2), 808–835.

Friedman, J. H. and L. C. Rafsky (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *Ann. Statist. 7*(4), 697–717.

Ghosh, A. K. and M. Biswas (2016). Distribution-free high-dimensional two-sample tests based on discriminating hyperplanes. *TEST 25*(3), 525–547.

Gibbons, J. D. and S. Chakraborti (2011). *Nonparametric Statistical Inference*. CRC Press, Boca Raton, FL.

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

Hall, P., J. S. Marron, and A. Neeman (2005). Geometric representation of high dimension, low sample size data. *J. R. Stat. Soc. Ser. B Stat. Methodol. 67*(3), 427–444.

# REFERENCES

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

Hollander, M., D. A. Wolfe, and E. Chicken (2014). *Nonparametric Statistical Methods*. John Wiley & Sons, Inc., Hoboken, NJ.

Jung, S. and J. S. Marron (2009). PCA consistency in high dimension, low sample size context. *Ann. Statist. 37*(6B), 4104–4130.

Kim, I., S. Balakrishnan, and L. Wasserman (2020). Robust multivariate nonparametric tests via projection averaging. *Ann. Statist. 48*(6), 3417–3441.

Li, J. and S. X. Chen (2012). Two sample tests for high-dimensional covariance matrices. *Ann. Statist. 40*(2), 908–940.

Liu, Z. and R. Modarres (2011). A triangle test for equality of distribution functions in high dimensions. *J. Nonparametr. Stat. 23*(3), 605–615.

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

Pan, W., Y. Tian, X. Wang, and H. Zhang (2018). Ball divergence: nonparametric two sample test. *Ann. Statist. 46*(3), 1109–1137.

Pan, W., X. Wang, H. Zhang, H. Zhu, and J. Zhu (2020). Ball covariance: a generic measure of dependence in Banach space. *J. Amer. Statist. Assoc. 115*(529), 307–317.

Rosenbaum, P. R. (2005). An exact distribution-free test comparing two multivariate

distributions based on adjacency. *J. R. Stat. Soc. Ser. B Stat. Methodol. 67*(4), 515–530.

Sarkar, S., R. Biswas, and A. K. Ghosh (2020). On some graph-based two-sample tests for high dimension, low sample size data. *Mach. Learn. 109*(2), 279–306.

Sarkar, S. and A. K. Ghosh (2018). On some high-dimensional two-sample tests based on averages of inter-point distances. *Stat 7*, e187, 16.

Schilling, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *J. Amer. Statist. Assoc. 81*(395), 799–806.

Srivastava, R., P. Li, and D. Ruppert (2016). RAPTT: an exact two-sample test in high dimensions using random projections. *J. Comput. Graph. Statist. 25*(3), 954–970.

Tsukada, S.-i. (2019). High dimensional two-sample test based on the inter-point distance. *Comput. Statist. 34*(2), 599–615.

Wei, S., C. Lee, L. Wichers, and J. S. Marron (2016). Direction-projection-permutation for high-dimensional hypothesis tests. *J. Comput. Graph. Statist. 25*(2), 549–569.

Yata, K. and M. Aoshima (2012). Effective pca for high-dimension, low-sample-size data with noise reduction via geometric representations. *J. Multivariate Anal. 105*(1), 193–215.

Yata, K. and M. Aoshima (2020). Geometric consistency of principal component scores for high-dimensional mixture models and its application. *Scand. J. of Statist. 47*(3), 899–921.