Bernoulli **27**(1), 2021, 419–441
https://doi.org/10.3150/20-BEJ1244

Comparing a large number of multivariate distributions

ILMUN KIM

Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
E-mail: ilmunk@stat.cmu.edu

In this paper, we propose a test for the equality of multiple distributions based on kernel mean embeddings. Our framework provides a flexible way to handle multivariate data by virtue of kernel methods and allows the number of distributions to increase with the sample size. This is in contrast to previous studies that have been mostly restricted to classical univariate settings with a fixed number of distributions. By building on Cramér-type moderate deviation for degenerate two-sample V-statistics, we derive the limiting null distribution of the test statistic and show that it converges to a Gumbel distribution. The limiting distribution, however, depends on an infinite number of nuisance parameters, which makes it infeasible for use in practice. To address this issue, the proposed test is implemented via the permutation procedure and is shown to be minimax rate optimal against sparse alternatives. During our analysis, an exponential concentration inequality for the permuted test statistic is developed which may be of independent interest.

Keywords: Bobkov's inequality; K-sample test; maximum mean discrepancy; permutation test

1. Introduction

Let $P_1, \dots, P_K$ be probability distributions defined on a common measurable space $(\mathcal{X}, \mathcal{B})$ for $K \geq 2$. The K-sample problem is concerned with testing the null hypothesis $H_0 : P_1 = \cdots = P_K$ against the alternative hypothesis $H_1 : P_i \neq P_j$ for some $i, j \in \{1, \dots, K\}$. This fundamental problem of comparing multiple distributions is a classical topic in statistics with a wide range of applications (Thas [49], Chen and Pokojovy [15], for reviews). Despite its long history, previous approaches to the K-sample problem have several limitations. First, many methods are limited to dealing with univariate data. For instance, Kiefer [28] proposes the K-sample analogues of the Kolmogorov–Smirnov and Cramér–Von Mises tests. Scholz and Stephens [44] generalize the Anderson–Darling test (Anderson and Darling [3]) to the K-sample case. These approaches are based on empirical distribution functions and are not easily extendable to multivariate data. Some other references that are restricted to the univariate K-sample problem include Conover [16], Zhang and Wu [54], Wyłupek [52], Quessy and Éthier [42], Lemeshko and Veretelnikova [34]. Second, most research in this area has been carried out under classical asymptotic regimes where the sample size goes to infinity but the number of distributions is fixed (e.g., Burke [10], Bouzebda, Keziou and Zari [9], Hušková and Meintanis [25], Martínez-Camblor, De Uña-Álvarez and Corral [37], Jiang, Ye and Liu [27], Mukhopadhyay and Wang [41], Sosthene *et al.* [46]). Clearly this classical asymptotic analysis is not appropriate for a dataset with large K and it only provides a narrow picture of the behavior of a test. To the best of our knowledge, Zhan and Hart [53] is the only study in the literature that considers large K. However, their analysis is limited to univariate data with fixed sample size. Third, recent developments on the multivariate K-sample problem are largely built upon an average difference between distributions (Bouzebda, Keziou and Zari [9], Hušková and Meintanis [25], Rizzo and Székely [43], Zhan and Hart [53], Mukhopadhyay and Wang [41], Sosthene *et al.* [46]). It is well known that the test based on an average-type test statistic tends to be powerful against dense alternatives in which many of $P_1, \dots, P_K$ are different to each other. On the other

1350-7265 © 2021 ISI/BS

hand, it tends to suffer from low power against sparse alternatives where only a few of $P_1, \dots, P_K$ are different from the others. Recently, sparse alternatives have been motivated by numerous applications such as DNA microarray analysis and anomaly detection where there are a small number of treatments that can actually contribute response variables. These applications have led to recent developments of tests tailored to sparse alternatives in the context of testing a high-dimensional vector (Jeng, Cai and Li [26], Fan, Liao and Yao [18], Liu and Li [36]), two-sample mean or covariance testing (Cai, Liu and Xia [11,12], Cai and Xia [13]), analysis of variance (Arias-Castro, Candès and Plan [4], Cai and Xia [13]) and independence testing (Han, Chen and Liu [23]). To our knowledge, however, a multivariate K-sample test specifically designed for sparse alternatives is not available in the current literature.

In this study, we propose a new K-sample test that addresses the aforementioned limitations of the previous approaches. More specifically, we introduce a K-sample test based on the kernel mean embedding method that has been successfully applied to multivariate hypothesis testing. Our test statistic is defined as the maximum of pairwise maximum mean discrepancies (Gretton *et al.* [20,21]), which leads to a powerful test against sparse alternatives. Throughout this paper, we investigate statistical properties of the proposed test under the asymptotic regime where both the sample size and the number of distributions tend to infinity. Below, we summarize our main findings and contributions.

- *Limiting null distribution*: By building on Drton, Han and Shi [17], we develop Cramér-type moderate deviation for degenerate two-sample V-statistics. Based on this result, we study the limiting distribution of the proposed test statistic when the sample size and the number of distributions increase simultaneously. In particular, we show the test statistic converges to a Gumbel distribution under some appropriate conditions.
- *Concentration inequality under permutations*: We demonstrate the usefulness of Bobkov's inequality (Bobkov [7]) in studying a concentration inequality for the permuted test statistic. By applying his result, we derive an exponential concentration inequality for the proposed test statistic under permutations. In contrast to usual Hoeffding or Bernstein-type inequalities, the developed inequality relies solely on completely known and easily computable quantities without any moment assumption.
- *Uniform consistency of the permutation test*: Leveraging the developed concentration inequality for the permuted statistic, we prove the uniform consistency of the permutation test over the class of sparse alternatives. Under some regularity conditions, we also show that the power of the permutation test cannot be improved from a minimax point of view.
- *Empirical power comparison against sparse alternatives*: A simulation study is conducted to compare the performance of the proposed maximum-type test with the existing average-type tests in the literature. The simulation results show that the proposed test consistently outperforms the average-type tests against sparse alternatives based on isotropic Laplace and Gaussian distributions.

Outline. The paper is organized as follows. In Section 2, we briefly review the maximum mean discrepancy and introduce our test statistic. Section 3 studies the limiting distribution of the proposed test statistic when the sample size and the number of distributions tend to infinity simultaneously. Section 4 formally introduces permutation procedures. In Section 5, we provide an exponential concentration inequality for the proposed test statistic under permutations. Section 6 investigates the power of the proposed test and proves its optimality property against sparse alternatives. In Section 7, we demonstrate the finite-sample performance of the proposed approach via simulations. Finally, Section 8 concludes the paper and discusses future work. The proofs not presented in the main text can be found in the supplemental article (Kim [29]).

2. Test statistic

We start with a brief overview of the maximum mean discrepancy proposed by Gretton *et al.* [20,21]. Let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) on $\mathcal{X}$ with a reproducing kernel $h : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$. For two functions $f, g \in \mathcal{H}$, we write the inner product on $\mathcal{H}$ by $\langle f, g \rangle_{\mathcal{H}}$ and the associated norm by $\|f\|_{\mathcal{H}}$. Given a probability distribution P, the kernel mean embedding of P is given by $\mu_h(P) = \mathbb{E}_{X \sim P}[h(X, \cdot)]$. Using the feature map $\psi : \mathcal{X} \mapsto \mathcal{H}$, which satisfies $h(x, y) = \langle \psi(x), \psi(y) \rangle_{\mathcal{H}}$, the kernel mean embedding can also be written as $\mathbb{E}_{X \sim P}[\psi(X)]$ (see, e.g., Muandet *et al.* [40] for details). We now provide the definition of the maximum mean discrepancy (MMD) associated with kernel h.

Definition 2.1 (Maximum mean discrepancy). Given two probability distributions, say P_1 and P_2, such that $\mathbb{E}_{X_1 \sim P_1} \|\psi(X_1)\|_{\mathcal{H}} < \infty$ and $\mathbb{E}_{X_2 \sim P_2} \|\psi(X_2)\|_{\mathcal{H}} < \infty$, the maximum mean discrepancy is defined as the RKHS norm of the difference between $\mu_h(P_1)$ and $\mu_h(P_2)$, that is,

$$\mathcal{V}_h(P_1, P_2) = \big\| \mu_h(P_1) - \mu_h(P_2) \big\|_{\mathcal{H}}.$$

It has been shown that when kernel h is characteristic (see, e.g., Fukumizu *et al.* [19], Sriperumbudur, Fukumizu and Lanckriet [47]), the MMD becomes zero *if and only if* $P_1 = P_2$. Some examples of characteristic kernels include Gaussian and Laplace kernels on $\mathcal{X} = \mathbb{R}^d$. This characteristic property allows to have a consistent two-sample test against any fixed alternatives. For general K-sample cases, we consider the maximum of pairwise maximum mean discrepancies as our metric, that is,

$$\mathcal{V}_{h,\max}(P_1, \ldots, P_K) = \max_{1 \le k < l \le K} \big\| \mu_h(P_k) - \mu_h(P_l) \big\|_{\mathcal{H}}.$$

Hence as long as h is characteristic, it is clear to see that $\mathcal{V}_{h,\max}(P_1, \ldots, P_K)$ is zero *if and only if* $P_1 = \cdots = P_K$.

Suppose that we observe identically distributed samples $X_{1,k}, \ldots, X_{n_k,k} \sim P_k$ for each $k = 1, \ldots, K$ and assume that the samples are mutually independent. We propose our test statistic defined as a plug-in estimator of $\mathcal{V}_{h,\max}$:

$$\widehat{\mathcal{V}}_{h,\max} = \max_{1 \le k < l \le K} \left\| \frac{1}{n_k} \sum_{i_1=1}^{n_k} \psi(X_{i_1,k}) - \frac{1}{n_l} \sum_{i_2=1}^{n_l} \psi(X_{i_2,l}) \right\|_{\mathcal{H}}.$$

In practice, the test statistic can be computed in a straightforward manner based on the kernel trick (e.g., Lemma 6 of Gretton *et al.* [21]):

$$\widehat{\mathcal{V}}_{h,\max} = \max_{1 \le k < l \le K} \left\{ \frac{1}{n_k^2} \sum_{i_1,i_2=1}^{n_k} h(X_{i_1,k}, X_{i_2,k}) + \frac{1}{n_l^2} \sum_{i_1,i_2=1}^{n_l} h(X_{i_1,l}, X_{i_2,l}) \right.$$
$$\left. - \frac{2}{n_k n_l} \sum_{i_1=1}^{n_k} \sum_{i_2=1}^{n_l} h(X_{i_1,k}, X_{i_2,l}) \right\}^{1/2}.$$

Throughout this paper, we denote the pooled samples by $\{Z_1, \ldots, Z_N\} = \{X_{1,1}, \ldots, X_{n_K,K}\}$ where $N = \sum_{k=1}^K n_k$.

3. Limiting distribution

Given the test statistic, our next step is to determine a critical value of the test with correct size α and good power properties. A common way of calibrating the critical value is based on the limiting null distribution of the test statistic. In this asymptotic approach, the critical value is set to be the $1-\alpha$ quantile of the limiting null distribution and the null hypothesis is rejected when the test statistic exceeds the critical value. The purpose of this section is to demonstrate the difficulty of implementing this asymptotic-based test in our setting. In particular, we show that $\widehat{\mathcal{V}}_{h,\max}$ converges to a Gumbel distribution with a potentially infinite number of unknown parameters under certain conditions. Unfortunately, it is by no means trivial to consistently estimate these infinite nuisance parameters. Furthermore, it is well known that a maximum-type statistic converges slowly to its limiting distribution (e.g., Hall [22]), which also makes the asymptotic test less attractive in practice. These limitations motivate us to delve into the permutation approach later in Sections 4–6.

3.1. Cramér-type moderate deviation

In order to derive the limiting distribution of the maximum of pairwise MMD statistics, it is important to understand the tail behavior of the two-sample MMD statistic. The main tool to this end is Cramér-type moderate deviation for degenerate two-sample V-statistics that we will develop in this subsection. Our result largely builds upon Cramér-type moderate deviation for degenerate one-sample U-statistics recently presented by Drton, Han and Shi [17].

Let us start with some notation and assumptions. For notational convenience, we write the MMD statistic between P_1 and P_2 as

$$\widehat{\mathcal{V}}_{12}^2 = \frac{1}{n_1^2}\sum_{i_1,i_2=1}^{n_1} h(X_{i_1,1},X_{i_2,1}) + \frac{1}{n_2^2}\sum_{i_1,i_2=1}^{n_2} h(X_{i_1,2},X_{i_2,2}) - \frac{2}{n_1n_2}\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2} h(X_{i_1,1},X_{i_2,2}).$$

By defining $h^*(x_1,x_2;y_1,y_2) := h(x_1,x_2)+h(y_1,y_2)-h(x_1,y_1)/2-h(x_1,y_2)/2-h(x_2,y_1)/2-h(x_2,y_2)/2$, the MMD statistic can also be written in the form of a two-sample V-statistic

$$\widehat{\mathcal{V}}_{12}^2 = \frac{1}{n_1^2n_2^2}\sum_{i_1,i_2=1}^{n_1}\sum_{j_1,j_2=1}^{n_2} h^*(X_{i_1,1},X_{i_2,1};X_{j_1,2},X_{j_2,2}). \tag{3.1}$$

Under the null hypothesis, the considered V-statistic is *degenerate* meaning that the conditional expectation of $h^*(X_{i_1,1},X_{i_2,1};X_{j_1,2},X_{j_2,2})$ given any one of $X_{i_1,1}, X_{i_2,1}, X_{j_1,2}, X_{j_2,2}$ has zero variance whenever $i_1 \neq i_2$ and $j_1 \neq j_2$.

Let X_1, X_2 be independent random vectors from P_1. We then define the centered kernel

$$\overline{h}(x_1,x_2) := h(x_1,x_2) - \mathbb{E}\big[h(x_1,X_2)\big] - \mathbb{E}\big[h(X_1,x_2)\big] + \mathbb{E}\big[h(X_1,X_2)\big],$$

which satisfies $\mathbb{E}[\overline{h}(X_1,X_2)] = 0$ and $\mathbb{E}[\overline{h}(x_1,X_2)] = 0$ almost surely. Under the finite second moment condition of the centered kernel, i.e. $\mathbb{E}[\{\overline{h}(X_1,X_2)\}^2] < \infty$, we may write

$$\overline{h}(x_1,x_2) = \sum_{v=1}^{\infty} \lambda_v \varphi_v(x_1)\varphi_v(x_2), \tag{3.2}$$

where $\{\lambda_v\}_{v=1}^{\infty}$ and $\{\varphi_v(\cdot)\}_{v=1}^{\infty}$ are the eigenvalues and eigenfunctions of the integral equation $\mathbb{E}[\overline{h}(x_1, X_2)\varphi_v(X_2)] = \lambda_v\varphi_v(x_1)$ (see, e.g., page 80 of Lee [32]).

To facilitate the analysis, we make the following assumptions regarding the kernel function.

(A1) Assume that $\mathbb{E}[|\overline{h}(X_1, X_1)|] < \infty$.

(A2) Suppose that $\overline{h}(x_1, x_2)$ admits the decomposition in (3.2) with $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. For all $u, v \in \mathbb{S}^{T-1} := \{x \in \mathbb{R}^T : \|x\|_2 = 1\}$ where $\|\cdot\|_2$ is Euclidean norm in $\mathbb{R}^T$ and any positive integer T, assume that there exists a constant $\eta > 0$ independent of T such that

$$\mathbb{E}\big[\big|\{\varphi_{1\cdots T}(X_1)^\top u\}^2\{\varphi_{1\cdots T}(X_1)^\top v\}^{m-2}\big|\big] \leq \eta^m m^{m/2}, \tag{3.3}$$

where $\varphi_{1\cdots T}(X_1) := (\varphi_1(X_1), \dots, \varphi_T(X_1))^\top$ and $m = 3, 4, \dots$

It is worth noting that the given conditions are more general than those used in Drton, Han and Shi [17]. Specifically, Drton, Han and Shi [17] assume that the kernel h and its eigenfunctions are uniformly bounded. Clearly, (A1) and (A2) are fulfilled under their boundedness assumptions. We also note that $\overline{h}(x_1, x_2)$ is a valid positive definite kernel (Sejdinovic *et al.* [45]), which yields $\{\overline{h}(x_1, x_2)\}^2 \leq \overline{h}(x_1, x_1)\overline{h}(x_2, x_2)$. Hence, the second moment condition $\mathbb{E}[\{\overline{h}(X_1, X_2)\}^2] < \infty$ is also satisfied under (A1). Finally, the multivariate moment condition (3.3) implies that individual eigenfunctions are sub-Gaussian (e.g., Vershynin [51]).

Under the given conditions, we present Cramér-type moderate deviation for the two-sample degenerate V-statistic described in (3.1). The proof of the following theorem can be found in Appendix A.

Theorem 3.1 (Cramér-type moderate deviation). *Suppose that* (A1) *and* (A2) *are fulfilled. Assume that there exists a constant $C_1 \geq 1$ such that $C_1^{-1} \leq n_1/n_2 \leq C_1$ and n_1/N converges to a constant as $N := n_1 + n_2 \to \infty$. Then under the null hypothesis $P_1 = P_2$, we have*

$$\frac{\mathbb{P}(n_1 n_2 \widehat{\mathcal{V}}_{12}^2/N \geq x)}{\mathbb{P}(\sum_{v=1}^{\infty} \lambda_v \xi_v^2 \geq x)} = 1 + o(1), \tag{3.4}$$

uniformly over $x \in (0, o(N^\theta))$ where $\xi_1, \xi_2, \dots$ are independent and identically distributed as $N(0, 1)$. Here θ is a constant that satisfies

$$\theta < \sup\Bigg\{q \in [0, 1/3) : \sum_{v > \lfloor N^{(1-3q)/5}\rfloor} \lambda_v = O\big(N^{-q}\big)\Bigg\},$$

when there exist infinitely many non-zero eigenvalues and $\theta = 1/3$ otherwise.

Remark 3.1. Although we restrict our attention to the two-sample V-statistic with a second-order kernel $h^*(x_1, x_2; y_1, y_2)$, our result can be straightforwardly extended to higher-order kernels $h^*(x_1, \dots, x_r; y_1, \dots, y_r)$ for some $r \geq 3$. The key idea is to consider Hoeffding's decomposition of two-sample U-statistics (page 40 of Lee [32]) and properly control the remainder terms (see, Drton, Han and Shi [17] for one-sample case). Finally, using the relationship between U- and V-statistics (e.g., page 183 of Lee [32]), one can derive the desired result for the V-statistic with a higher-order kernel. We do not pursue this direction here since the second-order kernel is enough for our application.

3.2. Gumbel limiting distribution

With the aid of Theorem 3.1, we are now ready to describe the limiting distribution of the proposed statistic under large K and large N situations. The main ingredient is Chen–Stein method for Poisson approximations (Arratia, Goldstein and Gordon [5]) that has been successfully applied to approximate the distribution of a maximum-type test statistic to a Gumbel distribution (see, e.g., Han, Chen and Liu [23], Drton, Han and Shi [17]). For sake of completeness, we state Theorem 1 of Arratia, Goldstein and Gordon [5].

Lemma 3.1 (Theorem 1 of Arratia, Goldstein and Gordon [5]). *Let $\mathcal{I}$ be an arbitrary index set and for $i \in \mathcal{I}$, let Y_i be a Bernoulli random variable with $p_i = \mathbb{P}(Y_i = 1) > 0$. For each $i \in \mathcal{I}$, consider a subset of $\mathcal{I}$ such that $B_i \subset \mathcal{I}$ with $i \in B_i$. Let us define $W = \sum_{i \in \mathcal{I}} Y_i$ and $\lambda = \mathbb{E}(W) = \sum_{i \in \mathcal{I}} p_i$. Let V be a Poisson random variable with mean λ. Then we have that*

$$\big|\mathbb{P}(W = 0) - \mathbb{P}(V = 0)\big| \le \min\{1, \lambda^{-1}\}(b_1 + b_2 + b_3)$$

where

$$b_1 := \sum_{i \in \mathcal{I}} \sum_{j \in B_i} p_i p_j, \qquad b_2 := \sum_{i \in \mathcal{I}} \sum_{i \neq j \in B_i} \mathbb{E}(Y_i Y_j) \quad \textit{and}$$

$$b_3 := \sum_{i \in \mathcal{I}} \mathbb{E}\bigg|\mathbb{E}\bigg[Y_i - p_i \bigg| \sum_{j \in \mathcal{I} - B_i} Y_j\bigg]\bigg|.$$

Let us denote the two-sample MMD statistic between P_k and P_l by $\widehat{\mathcal{V}}_{kl}^2$, that is $\widehat{\mathcal{V}}_{kl}^2 = \|n_k^{-1} \times \sum_{i=1}^{n_k} \psi(X_{i,k}) - n_l^{-1} \sum_{j=1}^{n_l} \psi(X_{j,l})\|_{\mathcal{H}}^2$. Assume the sample sizes are the same as $n := n_1 = \cdots = n_K$ for simplicity. Then based on the following key observation

$$\mathbb{P}\big(n\widehat{\mathcal{V}}_{h,\max}^2/2 \le x\big) = \mathbb{P}\bigg\{\sum_{1 \le k < l \le K} \mathbb{1}\big(n\widehat{\mathcal{V}}_{kl}^2/2 > x\big) = 0\bigg\},$$

Lemma 3.1 can be applied in our context with $W = \sum_{1 \le k < l \le K} \mathbb{1}(n\widehat{\mathcal{V}}_{kl}^2/2 > x)$ and $\lambda = \sum_{1 \le k < l \le K} \mathbb{P}(n\widehat{\mathcal{V}}_{kl}^2/2 > x)$. Ultimately the proof boils down to showing that b_1, b_2, b_3 converge to zero under appropriate conditions. This has been established in Appendix A and the result is summarized as follows.

Theorem 3.2 (Gumbel limit). *Suppose that* (A1) *and* (A2) *are fulfilled. Consider a balanced sample case such that $n := n_1 = \cdots = n_K$. Let θ be a constant chosen as in Theorem 3.1 and assume that* $\log K = o(n^\theta)$. *Then under the null hypothesis $P_1 = \cdots = P_K$, for any $y \in \mathbb{R}$,*

$$\lim_{n,K \to \infty} \mathbb{P}\bigg(\frac{n}{2\lambda_1}\widehat{\mathcal{V}}_{h,\max}^2 - 4\log K - (\mu_1 - 2)\log\log K \le y\bigg) = \exp\bigg\{-\frac{2^{\mu_1/2-2}\kappa}{\Gamma(\mu_1/2)}\exp\bigg(-\frac{y}{2}\bigg)\bigg\},$$

where $\kappa = \prod_{v=\mu_1+1}^{\infty}(1 - \lambda_v/\lambda_1)^{-1/2}$ and μ_1 is the multiplicity of the largest eigenvalue among the sequence $\{\lambda_v\}_{v=1}^{\infty}$.

Remark 3.2. From Theorem 3.2, it is clear that we need to know or at least estimate a potentially infinite number of parameters $\{\lambda_v\}_{v=1}^{\infty}$ in order to implement the asymptotic test. Even if one has access to these eigenvalues, the asymptotic test might suffer from slow convergence. This means that the test can be too liberal or too conservative in finite sample size situations.

Remark 3.3. When the sample sizes are unbalanced, the limiting distribution of $\widehat{\mathcal{V}}^2_{h,\max}$ may not have an explicit expression as in Theorem 3.2. In particular, it depends on the limit values of $n_k/(n_k+n_l)$ for $1 \le k < l \le K$. To avoid this complication, we simply focus on the case of equal sample sizes and present the explicit formula for the limiting distribution. Nevertheless, if we instead use the weighted K-sample statistic:

$$\max_{1\le k<l\le K}\left(\frac{n_k n_l}{n_k+n_l}\widehat{\mathcal{V}}^2_{kl}\right),$$

we may obtain the same Gumbel limiting distribution as in Theorem 3.2 for general sample sizes.

3.3. Examples

In general, it is challenging to find closed-form expressions for $\{\lambda_v\}_{v=1}^{\infty}$ and $\{\varphi_v(\cdot)\}_{v=1}^{\infty}$ as they depend on the kernel as well as the underlying distribution. We end this section with two simple examples for which $\{\lambda_v\}_{v=1}^{\infty}$ and $\{\varphi_v(\cdot)\}_{v=1}^{\infty}$ are explicit. Based on these, we illustrate Theorem 3.2.

- *Linear kernel*: Suppose that $\{X_{1,1},\dots,X_{n,1},\dots X_{1,K},\dots,X_{n,K}\}$ are independent and identically distributed as a multivariate normal distribution with mean zero and covariance matrix Σ. Suppose further that Σ is a diagonal matrix whose diagonal entries are $\lambda_1 = \cdots = \lambda_{\mu_1} > \lambda_{\mu_1+1} \ge \cdots \ge \lambda_d > 0$ for some $\mu_1 \ge 1$. Let us consider the linear kernel given as $h(x_1,x_2) = x_1^\top x_2$. Then it is straightforward to see that the centered kernel in (3.2) has the eigenfunction decomposition as

$$\overline{h}(x_1,x_2) = \sum_{v=1}^{d}\lambda_v\varphi_v(x_1)\varphi_v(x_2) = \sum_{v=1}^{d}\lambda_v\big(x_1^{(v)}/\sqrt{\lambda_v}\big)\big(x_2^{(v)}/\sqrt{\lambda_v}\big)$$

  where $x_1^{(v)}$ is the vth component of x_1. Under the given setting, $\{\varphi_1(X_{1,1}),\dots,\varphi_d(X_{1,1})\}$ are independent and identically distributed as $N(0,1)$. It can be shown that the conditions in Theorem 3.2 are satisfied with $\theta = 1/3$ under the Gaussian assumption. Thus, the resulting test statistic converges to a Gumbel distribution as in Theorem 3.2.
- *Chi-square kernel*: Suppose that $\{X_{1,1},\dots,X_{n,1},\dots X_{1,K},\dots,X_{n,K}\}$ are independent and identically distributed on a discrete domain $\{1,\dots,m\}$ with fixed m. Let $p_v > 0$ be the probability of observing the value v among $\{1,\dots,m\}$ and consequently $\sum_{v=1}^m p_v = 1$. Consider the chi-square kernel defined as $h(x_1,x_2) = \sum_{v=1}^m p_v^{-1}\mathbb{1}(x_1 = v)\mathbb{1}(x_2 = v)$. Let A be a $(m-1)\times(m-1)$ matrix whose (v_1,v_2) entry is $a_{v_1,v_2} = p_{v_1}^{-1} + p_m^{-1}$ if $v_1 = v_2$ and $a_{v_1,v_2} = p_m^{-1}$ otherwise. Let us define the eigenfunction $\varphi_v(x)$ to be the vth row of $A^{1/2}\{\mathbb{1}(x=1)-p_1,\dots,\mathbb{1}(x=m-1)-p_{m-1}\}^\top$ for $v = 1,\dots,m-1$. Then, following the calculation in Theorem 14.3.1 of Lehmann and Romano [33],

$$\overline{h}(x_1,x_2) = \sum_{v=1}^{m-1}\lambda_v\varphi_v(x_1)\varphi_v(x_2) = \sum_{v=1}^{m}\frac{\{\mathbb{1}(x_1=v)-p_v\}\{\mathbb{1}(x_2=v)-p_v\}}{p_v},$$

  where $\lambda_1 = \cdots = \lambda_{m-1} = 1$ and $\lambda_v = 0$ for $v \ge m$ and the eigenfunctions are bounded. Thus, the conditions in Theorem 3.2 are satisfied with $\theta = 1/3$ and the resulting test statistic converges to a Gumbel distribution.

4. Permutation approach

So far we have investigated the limiting null distribution of the proposed test statistic and demonstrated the difficulty of implementing the resulting asymptotic test. To address the issue, we take an alternative approach based on permutations that does not require prior knowledge on unknown parameters. The key advantage of the permutation approach is that it yields a valid level α test (or a size α test via randomization) for any finite sample size and for any number of distributions. This attractive property is true for any type of underlying distributions, provided that $\{Z_1, \dots, Z_N\}$ are exchangeable under H_0. In the following, we briefly describe the original and randomized permutation procedures. The randomized procedure has a computational advantage over the original procedure by considering a random subset of all permutations.

- *Permutation approach*: Let $\mathcal{B}_N$ be the collection of all possible permutations of $\{1, \dots, N\}$. For $\boldsymbol{b} = (b_1, \dots, b_N) \in \mathcal{B}_N$, we denote by $\widehat{\mathcal{V}}_{h,\max}^{(\boldsymbol{b})}$ the test statistic computed based on the permuted dataset $\{Z_{b_1}, \dots, Z_{b_N}\}$. We then clearly have $\widehat{\mathcal{V}}_{h,\max}^{(\boldsymbol{b}_0)} = \widehat{\mathcal{V}}_{h,\max}$ for $\boldsymbol{b}_0 = (1, \dots, N)$. The permutation p-value is calculated by

$$p_{\mathrm{perm}} = \frac{1}{N!} \sum_{\boldsymbol{b} \in \mathcal{B}_N} \mathbb{1}\big(\widehat{\mathcal{V}}_{h,\max}^{(\boldsymbol{b})} \geq \widehat{\mathcal{V}}_{h,\max}\big). \tag{4.1}$$

  It is well known that $\mathbb{P}(p_{\mathrm{perm}} \leq t) \leq t$ for any $0 \leq t \leq 1$ under H_0 (e.g., Chapter 15 of Lehmann and Romano [33]). Consequently $\mathbb{1}(p_{\mathrm{perm}} \leq \alpha)$ is a valid level α test.
- *Randomized version*: For large N, it would be beneficial to consider a subset of $\mathcal{B}_N$ and compute the approximated permutation p-value. Suppose that $\boldsymbol{b}_1', \dots, \boldsymbol{b}_M'$ are sampled uniformly from $\mathcal{B}_N$ with replacement. We then define a Monte-Carlo version of the permutation p-value by

$$p_{\mathrm{MC}} = \frac{1}{M+1} \left\{ 1 + \sum_{i=1}^{M} \mathbb{1}\big(\widehat{\mathcal{V}}_{h,\max}^{(\boldsymbol{b}_i')} \geq \widehat{\mathcal{V}}_{h,\max}\big) \right\}. \tag{4.2}$$

  It can be shown that $\mathbb{P}(p_{\mathrm{MC}} \leq t) \leq t$ for any $0 \leq t \leq 1$ under H_0 (see, e.g., Chapter 15 of Lehmann and Romano [33]). Hence, $\mathbb{1}(p_{\mathrm{MC}} \leq \alpha)$ is a valid level α test as well.

Having motivated the permutation approach, we next analyze uniform consistency as well as minimax optimality of the resulting permutation test against sparse alternatives in Section 6, building on concentration inequalities developed in the following section.

5. Concentration inequalities under permutations

This section develops a concentration inequality for the permuted MMD statistic with an exponential tail bound. The result established here is especially useful for studying the type II error (or the power) of the proposed permutation test in Section 6. Our result can also be valuable in addressing the computational issue of the permutation test. The permutation approach suffers from high computational cost as the number of all possible permutations increases very quickly with the sample size. As a result, it is common in practice to use Monte-Carlo sampling of random permutations to approximate the p-value of a permutation test. However, in some application areas such as genetic where extremely small p-values are of interest, the Monte-Carlo approach still requires heavy computations (Knijnenburg *et al.* [30], He *et al.* [24]). Our concentration inequality has an exponential tail bound with completely known quantities. Based on this, one can find a sharp upper bound for the permutation p-value (or the permutation critical value) without any computational cost for permutations. We discuss this direction in more detail in Remark 5.2.

5.1. Bobkov's inequality

Before we state the main result of this section, we introduce Bobkov's inequality (Bobkov [7]), which is the key ingredient of our proof. To state his result, we need to prepare some notation in advance. Consider a discrete cube given by

$$\mathcal{G}_{N,m} = \big\{\boldsymbol{w} = (w_1, \dots, w_N) \in \{0,1\}^N : w_1 + \cdots w_N = m\big\}.$$

Note that for each $\boldsymbol{w} \in \mathcal{G}_{N,m}$, there are exactly $m(N-m)$ neighbors $\{s_{ij}\boldsymbol{w}\}_{i\in I(\boldsymbol{w}), j\in J(\boldsymbol{w})}$ where $I(\boldsymbol{w}) = \{i \le N : w_i = 1\}$ and $J(\boldsymbol{w}) = \{j \le N : w_j = 0\}$ such that $(s_{ij}\boldsymbol{w})_r = w_r$ for $r \neq i, j$ and $(s_{ij}\boldsymbol{w})_i = w_j$, $(s_{ij}\boldsymbol{w})_j = w_i$. Now for a function f defined on $\mathcal{G}_{N,m}$, the Euclidean length of discrete gradient $\nabla f(\boldsymbol{w})$ is given as

$$\big|\nabla f(\boldsymbol{w})\big|^2 = \sum_{i\in I(\boldsymbol{w})} \sum_{j\in J(\boldsymbol{w})} \big|f(\boldsymbol{w}) - f(s_{ij}\boldsymbol{w})\big|^2.$$

For more details, we refer to Bobkov [7]. Then Bobkov's inequality is stated as follows.

Lemma 5.1 (Theorem 2.1 of Bobkov [7]). *For every real-valued function f on $\mathcal{G}_{N,m}$ and $|\nabla f(\boldsymbol{w})| \le \Sigma$ for all $\boldsymbol{w}$,*

$$\mathbb{P}_{\boldsymbol{w}}\big[f(\boldsymbol{w}) - \mathbb{E}_{\boldsymbol{w}}\big\{f(\boldsymbol{w})\big\} \ge t\big] \le \exp\big\{-(N+2)t^2/\big(4\Sigma^2\big)\big\},$$

where $\mathbb{P}_{\boldsymbol{w}}(\cdot)$ represents a counting probability measure on $\mathcal{G}_{N,m}$ and $\mathbb{E}_{\boldsymbol{w}}(\cdot)$ is the expectation associated with $\mathbb{P}_{\boldsymbol{w}}(\cdot)$.

5.2. Two-sample case

We first focus on the two-sample case. When $K = 2$, it is clear that the proposed test statistic becomes the V-statistic in Gretton *et al.* [21] and

$$\widehat{\mathcal{V}}_{h,\max} = \frac{N}{n_2}\Bigg\|\frac{1}{n_1}\sum_{i_1=1}^{n_1}\psi(X_{i,1}) - \frac{1}{N}\sum_{j=1}^{N}\psi(Z_j)\Bigg\|_{\mathcal{H}} = \frac{N}{n_1 n_2}\Bigg\|\sum_{i_1=1}^{n_1}\overline{\psi}(Z_i)\Bigg\|_{\mathcal{H}}, \tag{5.1}$$

where $\overline{\psi}(Z_{i_1}) = \psi(Z_{i_1}) - \frac{1}{N}\sum_{j=1}^{N}\psi(Z_j)$. Recall that $\boldsymbol{b}$ is a N-dimensional random vector uniformly distributed over $\mathcal{B}_N$ in the permutation procedure. As before in Section 4, we denote the test statistic based on the permuted dataset $\{Z_{b_1}, \dots, Z_{b_N}\}$ by

$$\widehat{\mathcal{V}}_{h,\max}^{(\boldsymbol{b})} := \frac{N}{n_1 n_2}\Bigg\|\sum_{i_1=1}^{n_1}\overline{\psi}(Z_{b_{i_1}})\Bigg\|_{\mathcal{H}}.$$

We also denote the probability law under permutations (conditional on $Z_1, \dots, Z_N$) by $\mathbb{P}_{\boldsymbol{b}}(\cdot)$ and the expectation associated with $\mathbb{P}_{\boldsymbol{b}}(\cdot)$ by $\mathbb{E}_{\boldsymbol{b}}(\cdot)$.

It should be stressed that in the two-sample case, there exists $\boldsymbol{w} \in \mathcal{G}_{N,n_1}$ corresponding to each $\boldsymbol{b} \in \mathcal{B}_N$ such that

$$\widehat{\mathcal{V}}_{h,\max}^{(\boldsymbol{b})} = \widehat{\mathcal{V}}_{h,\max}^{[\boldsymbol{w}]} := \frac{N}{n_1 n_2} \left\| \sum_{i=1}^{N} w_i \overline{\psi}(Z_i) \right\|_{\mathcal{H}}.$$

More importantly, both $\widehat{\mathcal{V}}_{h,\max}^{(\boldsymbol{b})}$ and $\widehat{\mathcal{V}}_{h,\max}^{[\boldsymbol{w}]}$ have the same probability law when $\boldsymbol{b}$ and $\boldsymbol{w}$ are uniformly distributed over $\mathcal{B}_N$ and $\mathcal{G}_{N,n_1}$, respectively. In other words, we have

$$\mathbb{P}_{\boldsymbol{b}}\big\{\widehat{\mathcal{V}}_{h,\max}^{(\boldsymbol{b})} - \mathbb{E}_{\boldsymbol{b}}\big(\widehat{\mathcal{V}}_{h,\max}^{(\boldsymbol{b})}\big) \geq t\big\} = \mathbb{P}_{\boldsymbol{w}}\big\{\widehat{\mathcal{V}}_{h,\max}^{[\boldsymbol{w}]} - \mathbb{E}_{\boldsymbol{w}}\big(\widehat{\mathcal{V}}_{h,\max}^{[\boldsymbol{w}]}\big) \geq t\big\} \quad \text{for all } t \in \mathbb{R}.$$

This key observation allows us to apply Bobkov's inequality to obtain a concentration inequality for the permuted test statistic in the following theorem.

Theorem 5.1 (Concentration inequality for two-sample statistic). *For $K = 2$, let $\mathbb{P}_{\boldsymbol{b}}$ be the uniform probability measure over permutations conditional on $\{Z_1, \dots, Z_N\}$. Let us write $\gamma_{1,2} = n_1 n_2/(n_1 + n_2)^2$. Further denote $\widetilde{h}(Z_i, Z_j) = h(Z_i, Z_i) + h(Z_j, Z_j) - 2h(Z_i, Z_j) \geq 0$ and*

$$\widehat{\sigma}^2 = \frac{1}{N(N-1)} \sum_{i \neq j = 1}^{N} \widetilde{h}(Z_i, Z_j). \tag{5.2}$$

Then for all $t > 0$, we have

$$\mathbb{P}_{\boldsymbol{b}}\left(\widehat{\mathcal{V}}_{h,\max}^{(\boldsymbol{b})} \geq t + \sqrt{\frac{\widehat{\sigma}^2}{2N\gamma_{1,2}}}\right) \leq \exp\left(-\frac{N\gamma_{1,2}^2 t^2}{2\widehat{\sigma}^2}\right). \tag{5.3}$$

Proof. From the previous discussion, it suffices to investigate a concentration inequality for $f(\boldsymbol{w}) := \widehat{\mathcal{V}}_{h,\max}^{[\boldsymbol{w}]}$, which is uniformly distributed on $\mathcal{G}_{N,n_1}$. Since Bobkov's inequality holds for $f(\boldsymbol{w})$, all we need to do is to find meaningful bounds of the expected value of $f(\boldsymbol{w})$ and the Euclidean length of $\nabla f(\boldsymbol{w})$. We first bound the expected value of $f(\boldsymbol{w})$. Using the feature map representation of kernel h, it is straightforward to see that

$$\sum_{i=1}^{N} \|\overline{\psi}(Z_i)\|_{\mathcal{H}}^2 = -\sum_{i \neq j = 1}^{N} \langle \overline{\psi}(Z_i), \overline{\psi}(Z_j)\rangle_{\mathcal{H}} = \frac{1}{2N} \sum_{i \neq j = 1}^{N} \widetilde{h}(Z_i, Z_j). \tag{5.4}$$

Then using Jensen's inequality together with the above identities,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{w}}\left[\left\| \sum_{i=1}^{N} w_i \overline{\psi}(Z_i) \right\|_{\mathcal{H}}\right] &\leq \sqrt{\mathbb{E}_{\boldsymbol{w}}\left[\sum_{i=1}^{N} w_i^2 \|\overline{\psi}(Z_i)\|_{\mathcal{H}}^2 + \sum_{i \neq j = 1}^{N} w_i w_j \langle \overline{\psi}(Z_i), \overline{\psi}(Z_i)\rangle_{\mathcal{H}}\right]} \\
&= \sqrt{\frac{n_1}{N} \sum_{i=1}^{N} \|\overline{\psi}(Z_i)\|_{\mathcal{H}}^2 + \frac{n_1(n_1-1)}{N(N-1)} \sum_{i \neq j = 1}^{N} \langle \overline{\psi}(Z_i), \overline{\psi}(Z_j)\rangle_{\mathcal{H}}}
\end{aligned}$$

$$= \sqrt{\frac{n_1 n_2}{2N^2(N-1)} \sum_{i \neq j=1}^{N} \widetilde{h}(Z_i, Z_j)}.$$

By multiplying the scaling factor $N/(n_1 n_2)$ on both sides, we have $\mathbb{E}_{\boldsymbol{w}}[f(\boldsymbol{w})] \leq \sqrt{\widehat{\sigma}^2/(2N\gamma_{1,2})}$.

Next, we bound $|\nabla f(\boldsymbol{w})|$. Recall the definition of $s_{ij} w$ in Section 5.1. Using the triangle inequality, we see that

$$\left| \frac{N}{n_1 n_2} \left\| \sum_{l=1}^{N} w_l \overline{\psi}(Z_l) \right\|_{\mathcal{H}} - \frac{N}{n_1 n_2} \left\| \sum_{l=1}^{N} (s_{ij} w)_l \overline{\psi}(Z_l) \right\|_{\mathcal{H}} \right| \leq \frac{N}{n_1 n_2} \left\| \overline{\psi}(Z_i) - \overline{\psi}(Z_j) \right\|_{\mathcal{H}}.$$

Based on this observation, one can find Σ, which is independent of $\boldsymbol{w}$, as

$$\left|\nabla f(\boldsymbol{w})\right|^2 \leq \Sigma^2 := \frac{N^2}{n_1^2 n_2^2} \sum_{1 \leq i < j \leq N} \|\overline{\psi}(Z_i) - \overline{\psi}(Z_j)\|_{\mathcal{H}}^2 = \frac{N^2}{2n_1^2 n_2^2} \sum_{i \neq j=1}^{N} \widetilde{h}(Z_i, Z_j),$$

where the last equality uses the identities in (5.4). Now apply Bobkov's inequality with the above pieces to obtain the desired result. □

Remark 5.1. Before we move on, we make several comments on Theorem 5.1.

(a) The tail of the given concentration inequality relies solely on the variance term of the kernel. This is in sharp contrast to Hoeffding or Bernstein-type inequalities (Boucheron, Lugosi and Massart [8]) that usually depend on the (possibly unknown) range of random variables.

(b) The given concentration inequality requires no assumption on random variables such as boundedness or more generally sub-Gaussianity. Furthermore, it only depends on known and easily computable quantities in practice.

(c) For $0 < \alpha < 1$, consider a test function $\phi_2 : \{Z_1, \dots, Z_N\} \mapsto \{0, 1\}$ such that

$$\phi_2 = I\left\{\widehat{\mathcal{V}}_{h,\max} \geq \sqrt{\frac{2\widehat{\sigma}^2}{N\gamma_{1,2}^2} \log\left(\frac{1}{\alpha}\right)} + \sqrt{\frac{\widehat{\sigma}^2}{2N\gamma_{1,2}}}\right\}.$$

As a corollary of Theorem 5.1, it can be seen that ϕ_2 is a valid level α test whenever $\{Z_1, \dots, Z_N\}$ are exchangeable.

(d) We stress that our test statistic is a degenerate two-sample V-statistic. Therefore, the previous studies on concentration inequalities for the permuted simple sum (e.g., Chatterjee [14], Adamczak, Chafaï and Wolff [1], Albert [2]) cannot be applied in our context.

5.3. Numerical illustrations

We illustrate the usefulness of Theorem 5.1 via simulations. First of all, we can use Theorem 5.1 to compute an upper bound for the original permutation p-value. In detail, suppose that $n_1 = n_2$ with $N = n_1 + n_2$ for simplicity. Then it is straightforward to see that the permutation p-value is less than or equal to

$$p_{\mathrm{Bobkov}} := \begin{cases} \exp\left\{-\frac{N}{32\widehat{\sigma}^2}\left(\widehat{\mathcal{V}}_{h,\max} - \sqrt{\frac{2\widehat{\sigma}^2}{N}}\right)^2\right\}, & \text{if } \widehat{\mathcal{V}}_{h,\max} \geq \sqrt{\frac{2\widehat{\sigma}^2}{N}} \\ 1, & \text{else.} \end{cases}$$

By the nature of the permutation test, p_{Bobkov} is a valid p-value in any finite sample size, in a sense that $\mathbb{P}(p_{\mathrm{Bobkov}} \leq \alpha) \leq \alpha$ under H_0. Another way of obtaining a finite-sample valid p-value is to use an *unconditional* concentration inequality. For example, Gretton *et al.* [21] employ McDiarmid's inequality (McDiarmid [39]) to have an concentration inequality for the MMD V-statistic with a bounded kernel. Based on Theorem 7 of Gretton *et al.* [21] under the bounded kernel assumption $0 \leq h(x, y) \leq B$, another valid p-value can be obtained as

$$p_{\mathrm{McDiarmid}} := \begin{cases} \exp\left\{-\dfrac{N}{8B}\left(\widehat{\mathcal{V}}_{h,\max} - \sqrt{\dfrac{32B}{N}}\right)^2\right\}, & \text{if } \widehat{\mathcal{V}}_{h,\max} \geq \sqrt{\dfrac{32B}{N}} \\ 1, & \text{else.} \end{cases}$$

Both approaches provide exponentially decaying p-values in sample size but we should emphasize that p_{Bobkov} does not require any moment conditions on the kernel. Even if the kernel is bounded, p_{Bobkov} would be preferred to $p_{\mathrm{McDiarmid}}$ when $\widehat{\sigma}^2$ is much smaller than B. This point is illustrated under the following set-up.

Set-up. We consider two kernels: (1) energy distance kernel $h(x, y) = (\|x\|_2 + \|y\|_2 - \|x - y\|_2)/2$ and (2) linear kernel $h(x, y) = x^\top y$. Although these kernels are unbounded in general, they are bounded when the underlying distributions have compact support. For this purpose, we consider two truncated normal distributions with the different location parameters $\mu_1 = 1$ and $\mu_2 = -1$ and the same scale parameter $\sigma^2 = 1$. We let both distributions have the same support as $[-5, 5]$ so that we can calculate the bound B for each kernel. For each sample size N among $\{100, 200, \ldots, 900, 1000\}$, the experiments were repeated 200 times to estimate the expected values of the p-values.

Results. In Figure 1, we present the simulation results of the comparison between p_{Bobkov} and $p_{\mathrm{McDiarmid}}$ under the described scenario. The p-values are displayed in log-scale for better visual comparison. Under the given setting, we observe that $\widehat{\sigma}^2$ is much smaller than B for both kernels, which in turns leads to a smaller value of p_{Bobkov} compared to $p_{\mathrm{McDiarmid}}$. More specifically, we observe (1) $\widehat{\sigma}^2 \approx 1.61$ on average and $B = 10$ for the energy distance kernel and (2) $\widehat{\sigma}^2 \approx 4.01$ on average and $B = 100$ for the linear kernel. It is worth noting that the benefit of using p_{Bobkov} becomes more evident for unbounded random variables for which $p_{\mathrm{McDiarmid}}$ is not even applicable.

Remark 5.2. The test based on p_{Bobkov} may not be recommended when the sample size is small and the significance level α is of moderate size (e.g., $\alpha = 0.05$). In this case, the permutation test via Monte-Carlo simulations would be more satisfactory. However, when the sample size is large and the significance level is very small (e.g., $\alpha = 10^{-100}$), the Monte-Carlo approach would be computationally infeasible, requiring at least α^{-1} random permutations in order to reject H_0. In this large-sample and small α situation, the approach based on p_{Bobkov} would be practically valuable, which does not require any computational cost on permutations.

Remark 5.3. While we focused on the case where $\widehat{\sigma}^2 \ll B$ to highlight the advantage of p_{Bobkov}, it is definitely possible to observe that $p_{\mathrm{McDiarmid}}$ is smaller than p_{Bobkov}, especially when B is comparable to or smaller than $\widehat{\sigma}^2$.

5.4. K-Sample case

Next, we give a general result for arbitrary $K \geq 2$. Unfortunately, we cannot directly apply Bobkov's inequality when $K > 2$ since the inequality holds only for a function $f(\boldsymbol{w})$ defined on a binary discrete

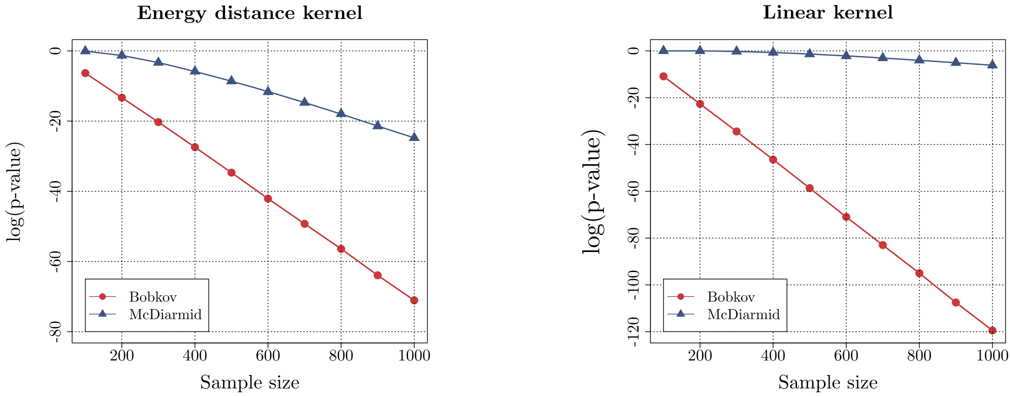


Figure 1. Comparisons between Bobkov's inequality and McDiarmid inequality in their application to p-value evaluation. In both energy distance kernel and linear kernel, Bobkov's inequality returns significantly smaller p-values than McDiarmid inequality. See Section 5.3 for details.

cube. Our strategy to overcome this problem is to first apply Bobkov's inequality to each pairwise MMD test statistic and then aggregate the results via the union bound. To start, we introduce $\widehat{\sigma}_K^2$ in Algorithm 1 that generalizes $\widehat{\sigma}^2$ to the K-sample case.

It can be seen that $\widehat{\sigma}_K^2$ is the same as $\widehat{\sigma}^2$ in (5.2) when $K = 2$ and can be computed in quadratic time for large K. Using $\widehat{\sigma}_K^2$, we extend Theorem 5.1 as follows.

Theorem 5.2 (Concentration inequality for K-sample statistic). *For $K \geq 2$, let $\mathbb{P}_{\boldsymbol{b}}$ be the uniform probability measure over permutations conditional on $\{Z_1, \ldots, Z_N\}$. For distinct $k, l \in \{1, \ldots, K\}$, let $\gamma_{k,l} = n_k n_l/(n_k + n_l)^2$ and consider $\widehat{\sigma}_K^2$ in Algorithm 1. Then for any $t \geq 0$,*

$$\begin{aligned} \mathbb{P}_{\boldsymbol{b}} \Biggl\{ \widehat{\mathcal{V}}_{h,\max}^{(\boldsymbol{b})} \geq t + \max_{1 \leq k < l \leq K} \sqrt{\frac{\widehat{\sigma}_K^2}{2(n_k + n_l)\gamma_{k,l}}} \Biggr\} \\ \leq \binom{K}{2} \exp\Biggl\{ - \min_{1 \leq k < l \leq K} \frac{(n_k + n_l)\gamma_{k,l}^2 t^2}{2\widehat{\sigma}_K^2} \Biggr\}. \end{aligned} \tag{5.5}$$

Algorithm 1: Calculation of $\widehat{\sigma}_K^2$

```
Require: the pooled samples $\{Z_1, \ldots, Z_N\}$, the number of samples $n_1, \ldots, n_K$.
    (1) Calculate $\widetilde{h}(Z_i, Z_j)$ for $1 \leq i \neq j \leq N$.
    (2) Sort and denote the previous outputs by $\widetilde{h}_{[1]} \geq \cdots \geq \widetilde{h}_{[N(N-1)]}$.
    (3) Compute $\widehat{\sigma}_K^2 := \max_{1 \leq k < l \leq K} \overline{\sigma}_{kl}^2$ where $\overline{\sigma}_{kl}^2$ is the sample average of $\widetilde{h}_{[1]}, \widetilde{h}_{[2]}, \ldots,$
            $\widetilde{h}_{[(n_k+n_l)(n_k+n_l-1)]}$.
    (4) Return $\widehat{\sigma}_K^2$.
```

Proof. For a given permutation $\boldsymbol{b} \in \mathcal{B}_N$, let us denote

$$\widehat{\mathcal{V}}_{kl}^{(\boldsymbol{b})} = \left\| \frac{1}{n_k}\sum_{i=1}^{n_k}\psi(Z_{b_{m_{k-1}+i}}) - \frac{1}{n_l}\sum_{j=1}^{n_l}\psi(Z_{b_{m_{l-1}+j}})\right\|_{\mathcal{H}},$$

where $m_{l-1} = \sum_{k=1}^{l-1} n_k$ and $m_0 = 0$ so that $\widehat{\mathcal{V}}_{h,\max}^{(\boldsymbol{b})} = \max_{1\leq k<l\leq K} \widehat{\mathcal{V}}_{kl}^{(\boldsymbol{b})}$. Based on the triangle inequality and the union bound, observe that

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{b}}\left\{\widehat{\mathcal{V}}_{h,\max}^{(\boldsymbol{b})} \geq t + \max_{1\leq k<l\leq K}\sqrt{\frac{\widehat{\sigma}_K^2}{2(n_k+n_l)\gamma_{k,l}}}\right\} \\
&\leq \mathbb{P}_{\boldsymbol{b}}\left[\max_{1\leq k<l\leq K}\left\{\widehat{\mathcal{V}}_{kl}^{(\boldsymbol{b})} - \sqrt{\frac{\widehat{\sigma}_K^2}{2(n_k+n_l)\gamma_{k,l}}}\right\} \geq t\right] \\
&\leq \sum_{1\leq k<l\leq K}\mathbb{P}_{\boldsymbol{b}}\left\{\widehat{\mathcal{V}}_{kl}^{(\boldsymbol{b})} \geq t + \sqrt{\frac{\widehat{\sigma}_K^2}{2(n_k+n_l)\gamma_{k,l}}}\right\}. \qquad (5.6)
\end{aligned}$$

Let $\widetilde{Z} = \{\widetilde{Z}_1, \ldots, \widetilde{Z}_{n_k+n_l}\}$ be the $n_k + n_l$ samples uniformly drawn from $\{Z_1, \ldots, Z_N\}$ without replacement. Write

$$\widehat{\mathcal{V}}_{kl}^{[\boldsymbol{w}]} = \frac{n_k+n_l}{n_k n_l}\left\|\sum_{i_1=1}^{n_k+n_l} w_{i_1}\left\{\psi(\widetilde{Z}_{i_1}) - \frac{1}{n_k+n_l}\sum_{i_2=1}^{n_k+n_l}\psi(\widetilde{Z}_{i_2})\right\}\right\|_{\mathcal{H}},$$

where $\boldsymbol{w} = \{w_1, \ldots, w_{n_k+n_l}\}$ is a set of Bernoulli random variables uniformly distributed on $\mathcal{G}_{n_k+n_l,n_k}$ as before. Then by the law of total expectation and a slight modification of the proof of Theorem 5.1, it can be seen that

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{b}}\left(\widehat{\mathcal{V}}_{kl}^{(\boldsymbol{b})} \geq t + \sqrt{\frac{\widehat{\sigma}_K^2}{2(n_k+n_l)\gamma_{k,l}}}\right) &= \mathbb{E}_{\widetilde{Z}}\left[\mathbb{P}_{\boldsymbol{w}}\left\{\widehat{\mathcal{V}}_{kl}^{[\boldsymbol{w}]} \geq t + \sqrt{\frac{\widehat{\sigma}_K^2}{2(n_k+n_l)\gamma_{k,l}}} | \widetilde{Z}\right\}\right] \\
&\leq \mathbb{E}_{\widetilde{Z}}\left[\exp\left\{-\frac{(n_k+n_l)\gamma_{k,l}^2 t^2}{2\widehat{\sigma}_K^2}\right\}\right] \\
&= \exp\left\{-\frac{(n_k+n_l)\gamma_{k,l}^2 t^2}{2\widehat{\sigma}_K^2}\right\},
\end{aligned}$$

where the last equality follows since $\widehat{\sigma}_K^2$ is invariant to the choice of $\widetilde{Z}$. By putting this result into the right-hand side of (5.6), the proof is complete. □

Remark 5.4. We provide some comments on Theorem 5.2.

(a) When $K = 2$, the concentration inequality given in (5.5) recovers the one in (5.3).

(b) One can replace $\widehat{\sigma}_K^2$ with $\max_{1\leq i<j\leq N}\widetilde{h}(Z_i, Z_j)$ in (5.5), which takes less time to compute, but at the expense of the loss of the tightness. Note, however, that the bound with $\max_{1\leq i<j\leq N}\widetilde{h}(Z_i, Z_j)$ is tight enough to prove minimax rate optimality of the proposed test. See the proof of Theorem 6.1 for details.

(c) As before in the two-sample case, the proposed K-sample concentration inequality is valid without any moment condition and it depends solely on known and easily computable quantities.

(d) Consider a test function $\phi_K : \{Z_1, \dots, Z_N\} \mapsto \{0, 1\}$ such that

$$\phi_K = I\Bigg[\widehat{\mathcal{V}}_{h,\max} \geq \max_{1 \leq k < l \leq K} \sqrt{\left\{\frac{2\widehat{\sigma}_K^2}{(n_k + n_l)\gamma_{k,l}^2}\right\} \log\left\{\frac{\binom{K}{2}}{\alpha}\right\}} + \max_{1 \leq k < l \leq K} \sqrt{\frac{\widehat{\sigma}_K^2}{2(n_k + n_l)\gamma_{k,l}}}\Bigg].$$

As a corollary of Theorem 5.2, it can be seen that ϕ_K is a valid level α test whenever $\{Z_1, \dots, Z_N\}$ are exchangeable under H_0.

6. Power analysis

In this section, we study the power of the permutation test based on the proposed test statistic and prove its minimax rate optimality against certain sparse alternatives. Throughout this section, we need the following assumptions:

(B1) Assume that kernel h is uniformly bounded by $0 \leq h(x, y) \leq B$ for all $x, y \in \mathcal{X}$.

(B2) There exists a fixed constant $c > 0$ such that $n_{\max}/n_{\min} \leq c$ for any sample sizes where $n_{\max}$ and $n_{\min}$ are the maximum and the minimum of $\{n_1, \dots, n_K\}$ respectively.

Note that the assumption (B1) is satisfied by some widely used kernels for example, Gaussian and Laplace kernels. It can also be satisfied by many other kernels when the underlying distributions have compact support. The second assumption (B2) states that $n_1, \dots, n_K$ are well-balanced. This assumption, for example, holds for the equal sample sizes with $c = 1$.

6.1. Power of the permutation test

Let $\mathcal{P}$ be the set of all distributions on $(\mathcal{X}, \mathcal{B})$. We characterize the difference between the null and the alternative in terms of $\max_{1 \leq k < l \leq K} \mathcal{V}_h(P_k, P_l)$, which is the population counterpart of the proposed test statistic $\widehat{\mathcal{V}}_{h,\max}$. In particular, for a given positive sequence ϵ_N and kernel h, let us define a class of alternatives:

$$\mathcal{F}_h(\epsilon_N) = \Big\{(P_1, \dots, P_K) \in \mathcal{P} : \max_{1 \leq k < l \leq K} \mathcal{V}_{kl} \geq \epsilon_N\Big\}, \tag{6.1}$$

where $\mathcal{V}_{kl} = \mathcal{V}_h(P_k, P_l)$ for simplicity. We call the collection of alternatives in $\mathcal{F}_h(\epsilon_N)$ as the sparse alternatives, in a sense that only a few of $\{\mathcal{V}_{kl}\}_{1 \leq k < l \leq K}$ are required to be greater than ϵ_N while the rest of them can be zero. Such sparse alternatives have been considered by many authors including Cai, Liu and Xia [11,12] and Han, Chen and Liu [23] in different contexts. The main goal of this subsection is to characterize the conditions under which the permutation test can be uniformly powerful over $\mathcal{F}_h(\epsilon_N)$. More specifically, we show that as long as the lower bound ϵ_N is sufficiently larger than

$$r_N^\star := \sqrt{\frac{\log K}{n_{\min}}},$$

then the proposed permutation test is uniformly consistent. Furthermore, in Section 6.2, we prove that this rate cannot be improved from a minimax perspective under some mild conditions on kernel h. In other words, the proposed test is minimax rate optimal against the sparse alternatives with the minimax rate $r_N^\star$.

We start by providing one lemma, which states that $\max_{1\leq k<l\leq K}|\widehat{\mathcal{V}}_{kl}-\mathcal{V}_{kl}|$ is bounded by $C\sqrt{\log K/n_{\min}}$ for some constant C with high probability.

Lemma 6.1. *Suppose that* (B1) *holds and recall that* $\widehat{\mathcal{V}}_{kl}=\|n_k^{-1}\sum_{i_1=1}^{n_k}\psi(X_{i_1,k})-n_l^{-1}\times\sum_{i_2=1}^{n_l}\psi(X_{i_2,l})\|_{\mathcal{H}}$. *Then with probability at least* $1-\beta$ *where* $0<\beta<1$, *we have*

$$\max_{1\leq k<l\leq K}|\widehat{\mathcal{V}}_{kl}-\mathcal{V}_{kl}|\leq 4\sqrt{\frac{B}{n_{\min}}}+2\sqrt{\frac{B}{n_{\min}}\log\left\{\frac{2}{\beta}\binom{K}{2}\right\}}.$$

Proof. Using Theorem 7 of Gretton *et al.* [21], one can obtain

$$\mathbb{P}\Big(|\widehat{\mathcal{V}}_{kl}-\mathcal{V}_{kl}|\geq 2\sqrt{n_k^{-1}B}+2\sqrt{n_l^{-1}B}+t\Big)\leq 2\exp\left\{-\frac{(n_k+n_l)\gamma_{k,l}t^2}{2B}\right\}.$$

Then the result follows by applying the union bound as in Theorem 5.2 and the following inequality

$$\min_{1\leq k<l\leq K}(n_k+n_l)\gamma_{k,l}\geq\frac{n_{\min}}{2}.$$

□

By building on Theorem 5.2 and Lemma 6.1, we prove the uniform consistency of the permutation test against $\mathcal{F}_h(\epsilon_N)$ when ϵ_N is much larger than $r_N^\star$. We provide the proof in Appendix A.

Theorem 6.1 (Uniform consistency of the original permutation test). *Assume that* (B1) *and* (B2) *are fulfilled. Denote the permutation test function by* $\phi_{K,\mathrm{perm}}=\mathbb{1}(p_{\mathrm{perm}}\leq\alpha)$ *where* p_{perm} *is given in* (4.1). *Then under* H_1,

$$\limsup_{n_{\min}\to\infty}\sup_{(P_1,\dots,P_K)\in\mathcal{F}_h(b_N r_N^\star)}\mathbb{P}(\phi_{K,\mathrm{perm}}=0)=0,$$

where b_N *is an arbitrary sequence that goes to infinity as* $n_{\min}\to\infty$.

Next by using Dvoretzky–Kiefer–Wolfowitz (DKW) inequality (e.g., Massart [38]), we extend the previous result to the randomized permutation test.

Corollary 6.1 (Uniform consistency of the randomized permutation test). *Assume that* (B1) *and* (B2) *are fulfilled. Denote the Monte-Carlo-based permutation test function by* $\phi_{K,\mathrm{MC}}=\mathbb{1}(p_{\mathrm{MC}}\leq\alpha)$ *where* p_{MC} *is given in* (4.2). *Then under* H_1,

$$\lim_{M\to\infty}\limsup_{n_{\min}\to\infty}\sup_{(P_1,\dots,P_K)\in\mathcal{F}_h(b_N r_N^\star)}\mathbb{P}(\phi_{K,\mathrm{MC}}=0)=0,$$

where b_N *is an arbitrary sequence that goes to infinity as* $n_{\min}\to\infty$.

Remark 6.1. It is worth pointing out that the results of both Theorem 6.1 and Corollary 6.1 hold regardless of whether K is fixed or increases with $n_{\min}$. However, we note that K cannot increase much faster than $e^{n_{\min}}$ as $\max_{1\leq k<l\leq K}\mathcal{V}_{kl}$ is upper bounded by a positive constant under (B1) and thereby $r_N^\star=\sqrt{\log K/n_{\min}}$ is also bounded.

6.2. Minimax rate optimality

Theorem 6.1 as well as Corollary 6.1 show that the original and randomized permutation tests can be uniformly powerful over $\mathcal{F}_h(b_N r_N^\star)$ when b_N is sufficiently large. In this subsection, we focus on the MMD associated with a translation invariant kernel defined on $\mathbb{R}^d$ and further show that the previous result cannot be improved from a minimax point of view. A kernel $h : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ is called *translation invariant* if there exists a symmetric positive definite function $\varphi : \mathbb{R}^d \mapsto \mathbb{R}$ such that $\varphi(x-y) = h(x,y)$ for all $x, y \in \mathbb{R}^d$ (Tolstikhin, Sriperumbudur and Muandet [50]). Then our result is stated as follows.

Theorem 6.2. *Let $0 < \alpha < 1$ and $0 < \zeta < 1-\alpha$. Suppose that $n_{\min} \to \infty$ and $K \to \infty$. Consider the class of sparse alternatives $\mathcal{F}_h(\epsilon_N)$ defined with a translation invariant kernel h on $\mathbb{R}^d$. Assume that there exists $z \in \mathbb{R}^d$ and $\kappa_1, \kappa_2 > 0$ such that $\varphi(0) - \varphi(z) \geq \kappa_1$ and $r_N^\star \leq \kappa_2$ for all $n_{\min}$. Further assume that* (B1) *and* (B2) *hold. Then under H_1, there exists a small constant $b > 0$ such that*

$$\liminf_{n_{\min}\to\infty} \inf_{\phi\in\Phi_N(\alpha)} \sup_{(P_1,\ldots,P_K)\in\mathcal{F}_h(br_N^\star)} \mathbb{P}(\phi = 0) \geq \zeta,$$

where $\Phi_N(\alpha)$ is the set of all level α test functions such that $\phi : \{Z_1, \ldots, Z_N\} \mapsto \{0, 1\}$.

Remark 6.2. The results in Theorem 6.1 and Theorem 6.2 imply that the proposed permutation test is not only consistent but also minimax rate optimal against the considered sparse alternatives. As far as we are aware, this is the first time that the power of the permutation test is theoretically analyzed under large N and large K situations.

Remark 6.3. In our problem setup, a distance between two distributions is measured in terms of the maximum mean discrepancy associated with kernel h. One can also study minimax optimality of the proposed test over a class of alternatives measured in terms of a more standard metric such as the L_2 distance. For this direction, the results of Li and Yuan [35] seem useful in which the authors explore minimax rate optimality of kernel mean embedding methods over a Sobolev space in the L_2 distance. We leave a detailed analysis of minimax optimality of the proposed test in other metrics to future work.

7. Simulations

In this section, we demonstrate the finite-sample performance of the proposed approach via simulations. We consider two characteristic kernels for our test statistic; (1) Gaussian kernel and (2) energy distance kernel. Gaussian kernel is given by $h(x,y) = \exp(-\|x-y\|_2^2/\sigma)$ for which we choose the tuning parameter σ by the median heuristic (Gretton *et al.* [21]). On the other hand, energy distance kernel is given by $h(x,y) = (\|x\|_2 + \|y\|_2 - \|x-y\|_2)/2$ as before. Note that the MMD statistic with energy distance kernel is equivalent to the energy statistic (Székely and Rizzo [48], Baringhaus and Franz [6]) in the two-sample case.

7.1. Other multivariate K-sample tests

We compare the performance of the proposed tests with two multivariate K-sample tests. The first one is the test based on DISCO statistic proposed by Rizzo and Székely [43]. Let $E_{kl,\alpha'}$ be the α'-energy statistic between P_k and P_l given by

$$E_{kl,\alpha'} = \frac{2}{n_k n_l} \sum_{i_1=1}^{n_k} \sum_{i_2=1}^{n_l} g_{\alpha'}(X_{i_1,k}, X_{i_2,l}) - \frac{1}{n_k^2} \sum_{i_1,i_2=1}^{n_k} g_{\alpha'}(X_{i_1,k}, X_{i_2,k}) \\ - \frac{1}{n_l^2} \sum_{i_1,i_2=1}^{n_l} g_{\alpha'}(X_{i_1,l}, X_{i_2,l}),$$

where $g_{\alpha'}(x, y) = \|x - y\|_2^{\alpha'}$. Let us write the between-sample and within-sample dispersions by $S_{\alpha'} = K^{-1} \sum_{1 \le k < l \le K} E_{kl,\alpha'}$ and $W_{\alpha'} = 2^{-1} \sum_{k=1}^{K} n_k^{-1} \sum_{i_1,i_2=1}^{n_k} g_{\alpha'}(X_{i_1,k}, X_{i_2,k})$. Then DISCO statistic is defined as ratio of the between-sample dispersion to the within-sample dispersion, that is

$$D_\gamma = \frac{S_{\alpha'}/(K-1)}{W_{\alpha'}/(N-K)}.$$

The second test, proposed by Hušková and Meintanis [25], is based on the empirical characteristic functions. For a given $\alpha'' \in \mathbb{R}$, Hušková and Meintanis [25] consider the weighted L_2 distance between empirical characteristic functions as their test statistic, that is

$$H_{\alpha''} = \sum_{k=1}^{K} \frac{N - n_k}{N n_k} \sum_{i_1,i_2=1}^{n_k} e^{-\|X_{i_1,k} - X_{i_2,k}\|_2^2/4\alpha''} - \frac{1}{N} \sum_{1 \le k \ne l \le K} \sum_{i_1=1}^{n_k} \sum_{i_2=1}^{n_l} e^{-\|X_{i_1,k} - X_{i_2,l}\|_2^2/4\alpha''}.$$

In their paper, Hušková and Meintanis [25] consider $\alpha'' = 1, 1.5, 2$ in their simulation study. Throughout our simulations, we choose $\alpha' = 1$ for $D_{\alpha'}$ and $\alpha'' = 1.5$ for $H_{\alpha''}$ and reject the null for large values of $D_{\alpha'}$ and $H_{\alpha''}$.

We also attempted to consider the graph-based K-sample test recently developed by Mukhopadhyay and Wang [41]. To implement their test, we used the R package provided by the same authors. Unfortunately, their method was not applicable when K is large due to numerical overflow in computing orthogonal polynomials. Hence, we focus on the first two methods described in this subsection and compare them with the proposed tests against sparse alternatives.

7.2. Set-up

Let us denote a multivariate normal distribution with mean vector μ and covariance matrix Σ by $N(\mu, \Sigma)$. Similarly we denote a multivariate Laplace distribution with mean vector μ and covariance matrix Σ by $L(\mu, \Sigma)$. We examine the performance of the considered tests under the following sparse alternatives:

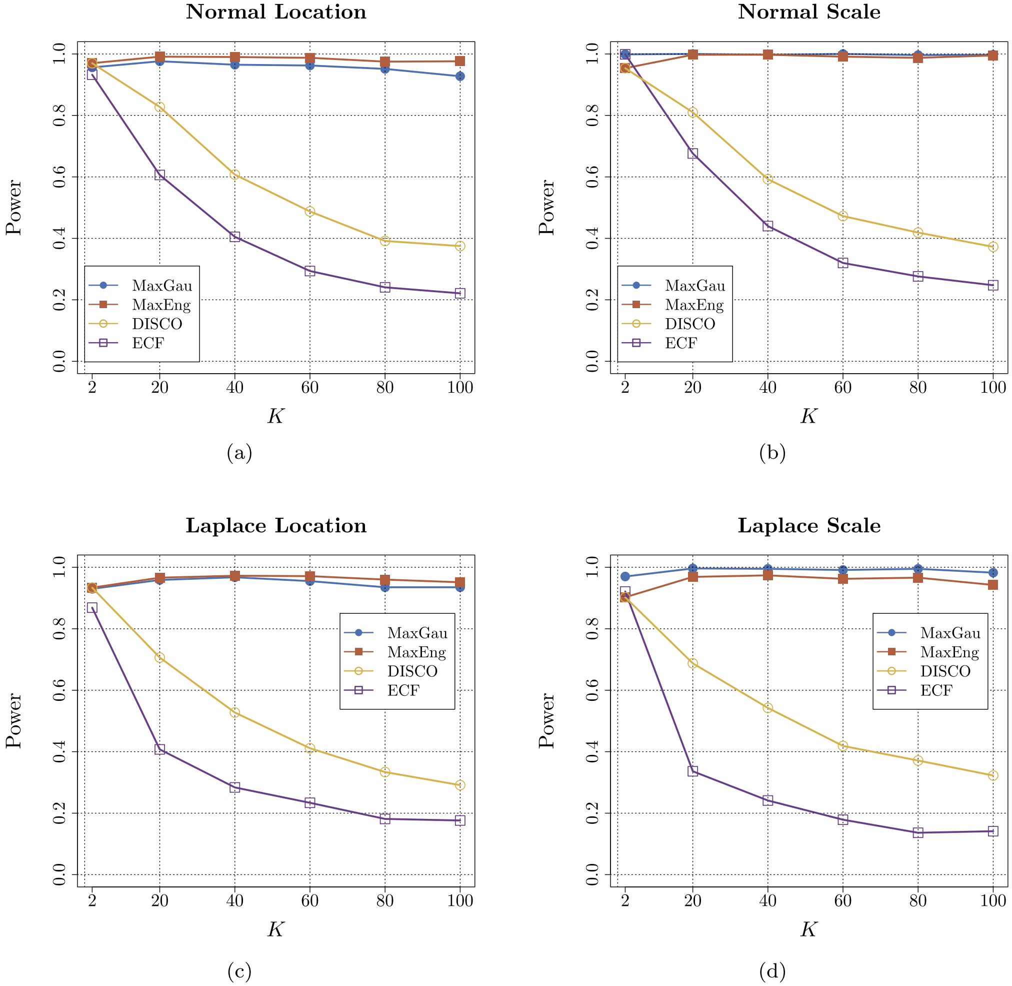
(a) *Normal Location*: $P_1 = N(\delta_1, I_d)$ and $P_2 = \cdots = P_K = N(\delta_0, I_d)$,
(b) *Normal Scale*: $P_1 = N(\delta_0, 3 \times I_d)$ and $P_2 = \cdots = P_K = N(\delta_0, I_d)$,
(c) *Laplace Location*: $P_1 = L(\delta_{1.2}, I_d)$ and $P_2 = \cdots = P_K = L(\delta_0, I_d)$,
(d) *Laplace Scale*: $P_1 = L(\delta_0, 3 \times I_d)$ and $P_2 = \cdots = P_K = L(\delta_0, I_d)$,

where $\delta_b = (b, \dots, b)^\top$ and I_d is the d-dimensional identity matrix. In words, we consider the sparse alternatives where only one of the distributions differs from the other $K - 1$ distributions. Consequently, the signal is getting sparser as K increases. Throughout our experiments, we fix sample sizes $n_1 = n_2 = \cdots = n_K = 10$ and dimension $d = 5$ while increasing the number of distributions

$K \in \{2, 20, 40, 60, 80, 100\}$. All tests were implemented via the randomized permutation procedure with $M = 200$ random permutations using the p-value in (4.2). As a result, they are all valid level α tests. Simulations were repeated 800 times to estimate the power at significance level $\alpha = 0.05$.

7.3. Results

From the results presented in Figure 2, we observe that the tests based on $D_{\alpha'}$ and $H_{\alpha''}$ have consistently decreasing power as K increases in all sparse scenarios. This can be explained by the fact that $D_{\alpha'}$ and $H_{\alpha''}$ are defined as an average between pairwise distances. Under the given sparse scenario, the average of pairwise distances, which is a signal to reject H_0, decreases as K increases. Hence, the resulting



(a)

(b)

(c)

(d)

Figure 2. Empirical power comparisons of the considered tests against (a) Normal location, (b) Normal scale, (c) Laplace location, (d) Laplace scale alternatives. We refer to the tests based on $\widehat{\mathcal{V}}_{h,\max}$ with Gaussian kernel and energy distance kernel as MaxGau and MaxEng, respectively. In addition, the tests based on $D_{\alpha'}$ and $H_{\alpha''}$ are referred to as DISCO and ECF, respectively. See Section 7 for details.

tests based on $D_{\alpha'}$ and $H_{\alpha''}$ suffer from low power in large K. On the other hand, the proposed tests show robust performance to the number of distributions K under the given setting. They in fact have power very close to one even when K is considerably large, which emphasizes the benefit of using the maximum-type statistic against sparse alternatives.

Despite their good performance over sparse alternatives, the proposed tests do not always perform better than the average-type tests based on $D_{\alpha'}$ and $H_{\alpha''}$. For example, these average-type tests may outperform the proposed maximum-type tests against dense alternatives where many of $P_1, \dots, P_K$ differ from each other. Given that prior knowledge on alternatives is not always available to users, developing a powerful test against both dense and sparse alternatives is an interesting direction for future work.

8. Conclusion

In this paper, we introduced a new nonparametric K-sample test based on the maximum mean discrepancy. The limiting distribution of the proposed test statistic was derived based on Cramér-type moderate deviation for degenerate two-sample V-statistics. Unfortunately, the limiting distribution relies on an infinite number of nuisance parameters, which are intractable in general. Due to this challenge, we considered the permutation approach to determine the cut-off value of the test. We provided a concentration inequality for the proposed test statistic with a sharp exponential tail bound under permutations. On the basis of this result, we studied the power of the permutation test in large K and large N situations and further proved its minimax rate optimality under some regularity conditions. From our simulation studies, the proposed test is shown to be powerful against sparse alternatives where the previous methods suffer from low power. These findings suggest that our method will be useful in application areas where only a small number of populations differ from the others.

The power analysis in Section 6 relies on the assumption that a kernel is uniformly bounded. Although some of the popular kernels satisfy this assumption, our result cannot be applied to unbounded cases. One possible way to address this issue is to impose appropriate moment conditions on a kernel and utilize a suitable concentration inequality (e.g., a modified McDiarmid's inequality in Kontorovich [31]) to obtain a similar result to Lemma 6.1. This topic is reserved for future work.

Acknowledgements

The author would like to thank the associate editor and the anonymous reviewers for their valuable comments. The author also thanks Sivaraman Balakrishnan and Larry Wasserman for their kind support and constructive feedback.

Supplementary Material

Supplement to "Comparing a large number of multivariate distributions" (DOI: 10.3150/20-BEJ1244SUPP; .pdf). This supplemental file includes the technical proofs omitted in the main text.

References

[1] Adamczak, R., Chafaï, D. and Wolff, P. (2016). Circular law for random matrices with exchangeable entries. *Random Structures Algorithms* **48** 454–479. MR3481269 https://doi.org/10.1002/rsa.20599

[2] Albert, M. (2019). Concentration inequalities for randomly permuted sums. In *High Dimensional Probability VIII* 341–383. Springer.

[3] Anderson, T.W. and Darling, D.A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Ann. Math. Stat.* **23** 193–212. MR0050238 https://doi.org/10.1214/aoms/1177729437

[4] Arias-Castro, E., Candès, E.J. and Plan, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *Ann. Statist.* **39** 2533–2556. MR2906877 https://doi.org/10.1214/11-AOS910

[5] Arratia, R., Goldstein, L. and Gordon, L. (1989). Two moments suffice for Poisson approximations: The Chen–Stein method. *Ann. Probab.* **17** 9–25. MR0972770

[6] Baringhaus, L. and Franz, C. (2004). On a new multivariate two-sample test. *J. Multivariate Anal.* **88** 190–206. MR2021870 https://doi.org/10.1016/S0047-259X(03)00079-4

[7] Bobkov, S.G. (2004). Concentration of normalized sums and a central limit theorem for noncorrelated random variables. *Ann. Probab.* **32** 2884–2907. MR2094433 https://doi.org/10.1214/009117904000000720

[8] Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities*: *A Nonasymptotic Theory of Independence*. Oxford: Oxford Univ. Press. With a foreword by Michel Ledoux. MR3185193 https://doi.org/10.1093/acprof:oso/9780199535255.001.0001

[9] Bouzebda, S., Keziou, A. and Zari, T. (2011). K-sample problem using strong approximations of empirical copula processes. *Math. Methods Statist.* **20** 14–29. MR2811029 https://doi.org/10.3103/S1066530711010029

[10] Burke, M.D. (1979). On the asymptotic power of some k-sample statistics based on the multivariate empirical process. *J. Multivariate Anal.* **9** 183–205. MR0538401 https://doi.org/10.1016/0047-259X(79)90078-2

[11] Cai, T., Liu, W. and Xia, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *J. Amer. Statist. Assoc.* **108** 265–277. MR3174618 https://doi.org/10.1080/01621459.2012.758041

[12] Cai, T.T., Liu, W. and Xia, Y. (2014). Two-sample test of high dimensional means under dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 349–372. MR3164870 https://doi.org/10.1111/rssb.12034

[13] Cai, T.T. and Xia, Y. (2014). High-dimensional sparse MANOVA. *J. Multivariate Anal.* **131** 174–196. MR3252643 https://doi.org/10.1016/j.jmva.2014.07.002

[14] Chatterjee, S. (2007). Stein's method for concentration inequalities. *Probab. Theory Related Fields* **138** 305–321. MR2288072 https://doi.org/10.1007/s00440-006-0029-y

[15] Chen, S. and Pokojovy, M. (2018). Modern and classical k-sample omnibus tests. *Wiley Interdiscip. Rev.*: *Comput. Stat.* **10** e1418, 12. MR3749550 https://doi.org/10.1002/wics.1418

[16] Conover, W.J. (1965). Several k-sample Kolmogorov–Smirnov tests. *Ann. Math. Stat.* **36** 1019–1026. MR0175230 https://doi.org/10.1214/aoms/1177700073

[17] Drton, M., Han, F. and Shi, H. (2018). High dimensional independence testing with maxima of rank correlations. arXiv preprint arXiv:1812.06189.

[18] Fan, J., Liao, Y. and Yao, J. (2015). Power enhancement in high-dimensional cross-sectional tests. *Econometrica* **83** 1497–1541. MR3384226 https://doi.org/10.3982/ECTA12749

[19] Fukumizu, K., Gretton, A., Sun, X. and Schölkopf, B. (2008). Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems* 489–496.

[20] Gretton, A., Borgwardt, K.M., Rasch, M., Schölkopf, B. and Smola, A.J. (2007). A kernel method for the two-sample problem. In *Advances in Neural Information Processing Systems* 513–520.

[21] Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B. and Smola, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.* **13** 723–773. MR2913716

[22] Hall, P. (1991). On convergence rates of suprema. *Probab. Theory Related Fields* **89** 447–455. MR1118558 https://doi.org/10.1007/BF01199788

[23] Han, F., Chen, S. and Liu, H. (2017). Distribution-free tests of independence in high dimensions. *Biometrika* **104** 813–828. MR3737306 https://doi.org/10.1093/biomet/asx050

[24] He, H.Y., Basu, K., Zhao, Q. and Owen, A.B. (2019). Permutation p-value approximation via generalized Stolarsky invariance. *Ann. Statist.* **47** 583–611. MR3909943 https://doi.org/10.1214/18-AOS1702

[25] Hušková, M. and Meintanis, S.G. (2008). Tests for the multivariate k-sample problem based on the empirical characteristic function. *J. Nonparametr. Stat.* **20** 263–277. MR2421770 https://doi.org/10.1080/10485250801948294

[26] Jeng, X.J., Cai, T.T. and Li, H. (2010). Optimal sparse segment identification with application in copy number variation analysis. *J. Amer. Statist. Assoc.* **105** 1156–1166. MR2752611 https://doi.org/10.1198/jasa.2010.tm10083

[27] Jiang, B., Ye, C. and Liu, J.S. (2015). Nonparametric K-sample tests via dynamic slicing. *J. Amer. Statist. Assoc.* **110** 642–653. MR3367254 https://doi.org/10.1080/01621459.2014.920257

[28] Kiefer, J. (1959). K-sample analogues of the Kolmogorov–Smirnov and Cramér–V. Mises tests. *Ann. Math. Stat.* **30** 420–447. MR0102882 https://doi.org/10.1214/aoms/1177706261

[29] Kim, I. (2020). Supplement to "Comparing a large number of multivariate distributions." https://doi.org/10.3150/20-BEJ1244SUPP

[30] Knijnenburg, T.A., Wessels, L.F., Reinders, M.J. and Shmulevich, I. (2009). Fewer permutations, more accurate P-values. *Bioinformatics* **25** i161–i168.

[31] Kontorovich, A. (2014). Concentration in unbounded metric spaces and algorithmic stability. In *International Conference on Machine Learning* 28–36.

[32] Lee, A.J. (1990). *U-Statistics: Theory and Practice. Statistics: Textbooks and Monographs* **110**. New York: Dekker. MR1075417

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

[34] Lemeshko, B.Y. and Veretelnikova, I.V. (2018). On some new K-samples tests for testing the homogeneity of distribution laws. In 2018 *XIV International Scientific-Technical Conference on Actual Problems of Electronics Instrument Engineering* (*APEIE*) 153–157. IEEE.

[35] Li, T. and Yuan, M. (2019). On the optimality of Gaussian kernel based nonparametric tests against smooth alternatives. arXiv preprint arXiv:1909.03302.

[36] Liu, W. and Li, Y.Q. (2020). Sign-based test for mean vector in high-dimensional and sparse settings. *Acta Math. Sin.* (*Engl. Ser.*) **36** 93–108. MR4049304 https://doi.org/10.1007/s10114-019-8290-z

[37] Martínez-Camblor, P., De Uña-Álvarez, J. and Corral, N. (2008). k-sample test based on the common area of kernel density estimators. *J. Statist. Plann. Inference* **138** 4006–4020. MR2455983 https://doi.org/10.1016/j.jspi.2008.02.008

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

[39] McDiarmid, C. (1989). On the method of bounded differences. In *Surveys in Combinatorics*, 1989 (*Norwich*, 1989). *London Mathematical Society Lecture Note Series* **141** 148–188. Cambridge: Cambridge Univ. Press. MR1036755

[40] Muandet, K., Fukumizu, K., Sriperumbudur, B. and Schölkopf, B. (2016). Kernel mean embedding of distributions: A review and beyonds. *Stat* **1050** 31.

[41] Mukhopadhyay, S. and Wang, K. (2018). Nonparametric high-dimensional K-sample comparison. arXiv preprint arXiv:1810.01724.

[42] Quessy, J.-F. and Éthier, F. (2012). Cramér–von Mises and characteristic function tests for the two and k-sample problems with dependent data. *Comput. Statist. Data Anal.* **56** 2097–2111. MR2892402 https://doi.org/10.1016/j.csda.2011.12.021

[43] Rizzo, M.L. and Székely, G.J. (2010). DISCO analysis: A nonparametric extension of analysis of variance. *Ann. Appl. Stat.* **4** 1034–1055. MR2758432 https://doi.org/10.1214/09-AOAS245

[44] Scholz, F.-W. and Stephens, M.A. (1987). k-sample Anderson–Darling tests. *J. Amer. Statist. Assoc.* **82** 918–924. MR0910001

[45] Sejdinovic, D., Sriperumbudur, B., Gretton, A. and Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Statist.* **41** 2263–2291. MR3127866 https://doi.org/10.1214/13-AOS1140

[46] Sosthene, A., Balogoun, K., Martial Nkiet, G. and Ogouyandjou, C. (2018). Kernel based method for the k-sample problem. arXiv preprint arXiv:1812.00100.

[47] Sriperumbudur, B.K., Fukumizu, K. and Lanckriet, G.R.G. (2011). Universality, characteristic kernels and RKHS embedding of measures. *J. Mach. Learn. Res.* **12** 2389–2410. MR2825431

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

[49] Thas, O. (2010). *Comparing Distributions*. *Springer Series in Statistics*. New York: Springer. MR2547894 https://doi.org/10.1007/b99044

[50] Tolstikhin, I., Sriperumbudur, B.K. and Muandet, K. (2017). Minimax estimation of kernel mean embeddings. *J. Mach. Learn. Res.* **18** Paper No. 86, 47. MR3714249

[51] Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. *Cambridge Series in Statistical and Probabilistic Mathematics* **47**. Cambridge: Cambridge Univ. Press. With a foreword by Sara van de Geer. MR3837109 https://doi.org/10.1017/9781108231596

[52] Wyłupek, G. (2010). Data-driven k-sample tests. *Technometrics* **52** 107–123. MR2654991 https://doi.org/10.1198/TECH.2009.08101

[53] Zhan, D. and Hart, J.D. (2014). Testing equality of a large number of densities. *Biometrika* **101** 449–464. MR3215359 https://doi.org/10.1093/biomet/asu002

[54] Zhang, J. and Wu, Y. (2007). k-sample tests based on the likelihood ratio. *Comput. Statist. Data Anal.* **51** 4682–4691. MR2364473 https://doi.org/10.1016/j.csda.2006.08.029

Received May 2019 and revised February 2020