Permutation Invariant Functions: Statistical Testing, Density Estimation, and Metric Entropy

Wee Chaimanowong
The Chinese University of Hong Kong

Ying Zhu
University of California San Diego

Abstract

Permutation invariance is among the most common symmetries that can be exploited to simplify complex problems in machine learning. There has been a tremendous surge of research activities in building permutation invariant machine learning architectures. However, less attention is given to: (1) how to statistically test for the assumption of permutation invariance of coordinates in a random vector where the dimension is allowed to grow with the sample size; (2) how to estimate permutation invariant density functions; (3) how much "smaller" is the class of smooth functions with permutation invariance compared to that without permutation invariance. In this paper, we take a step back and examine these fundamental questions. In particular, our testing method is based on a sorting trick, and our estimation method is based on an averaging trick. These tricks substantially simplify the exploitation of permutation invariance. We also analyze the metric entropy of permutation invariant function classes and compare them with their counterparts without imposing permutation invariance.[1][2]

1 INTRODUCTION

Many applications can benefit from exploiting a known symmetry in the data. One of the most basic symmetries is permutation invariance, where the function outputs are invariant to the order of the inputs. Permutation invariance plays a crucial role in machine learning applications such as set anomaly detection, text concept set retrieval, and point cloud classification. Building permutation invariant machine learning architectures and assessing their computational performance has been a popular topic in the field. Viewing neural network inputs and outputs as random variables, Bloem-Reddy et al. (2020) studies the structure of neural networks that are useful for modeling data that are invariant, and show a connection between functional and probabilistic symmetry.

The discussion of Bloem-Reddy et al. (2020) raises the fundamental concept, *exchangeability*, which is related to permutation invariance in probability distributions. In particular, a sequence of random variables $X_1, X_2, X_3, \cdots$ is called exchangeable if for any permutation σ, the permuted sequence $X_{\sigma(1)}, X_{\sigma(2)}, X_{\sigma(3)}, \cdots$ has the same joint distribution as the original sequence. Using this terminology, a CDF $F(t)$ associated with a sequence of d real–valued random variables $(X_1, ..., X_d) \in \mathbb{R}^d$ is permutation invariant if such a sequence of random variables is exchangeable. Suppose that n random vectors $t_1, t_2, \cdots, t_n$ are independently drawn from an arbitrary $d-$variate distribution $F(t)$, where $t_i = (t_i^1, t_i^2, \cdots, t_i^d) \in \mathbb{R}^d$ for every $i = 1, ..., n$. Note that $t_1, t_2, \cdots, t_n$ is exchangeable, but each $t_i \sim F$ is *not* exchangeable if F is not permutation invariant.

The example above raises the importance of distinguishing *permuting samples* from *permuting the coordinates of a random vector*. There is a vast literature on the former but a relatively scarce literature on the latter (as pointed out by (Kalina and Janáček, 2023, page 3143)). Our paper focuses on the latter. The former has been discussed in the contexts of conformal prediction, testing independence, and testing the equality of two distributions (for example, Kuchibhotla (2020)). There is a separate literature related to permutation tests (Anderson and Robinson, 2001; Koning, 2024; Koning and Hemerik, 2024; Ramdas et al., 2023) based on U-statistics such as (Van der

[1]Both authors contributed equally to this paper and are listed alphabetically.

[2]Code for the numerical studies can be found at `https://github.com/wchaimanowong/perm_inv`.

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

Vaart, 2000, Chapter 12–13). Permutation test statistics are U-statistics which permute samples that satisfy i.i.d. or weak dependence conditions. In contrast, our problems permute on the dimension and make *no* assumptions about the dependence among the coordinates of a random vector t_i.

This fact is crucial for numerous applications in health sciences, finance, and climatology where researchers are interested in whether features (instead of samples) are permutable. For example, one application tests whether the red and white blood cell counts and hemoglobin concentration are permutable in athletes, and measurements were sampled (Kalina and Janáček, 2023). A researcher may not want to impose any condition on the dependence between different types of measurements. Applications like this one have motivated statisticians to develop tests for the assumption of permutation invariance of coordinates in a random vector; see "Related work" in Section 2 for the details.

In this paper, we propose a statistical procedure that tests *directly* whether the coordinates of a random vector from an unknown multivariate distribution are permutable. The term "directly" sets our contribution apart from most of the literature discussed at the end of Section 2: the majority of these methods test weaker conditions of permutation invariance (instead of itself), and consequently, our proposed test has more power. Specifically, our test statistics take the form $T := \sup_{t\in[0,1]^d} \sqrt{n} \left|\widetilde{F}_n(t) - F_n(t)\right|$, where $F_n(t)$ is the empirical CDF at t, $\widetilde{F}_n(t) = F_n(\operatorname{sort} t)$, and n is the number of the random samples. We approximate the quantile of T with the multiplier bootstrap method and show that our test attains the pre-specified significance level asymptotically.

Suppose that our test cannot reject the null hypothesis of permutation invariance. Then one may be interested in estimating the underlying density function by exploiting the potential symmetry. We propose a kernel density estimator (KDE) that averages the standard KDE over a carefully constructed subset of permutations. When the true density is indeed permutation invariant, the averaged KDE yields the same bias as the standard KDE but reduces the variance of the standard KDE by a factor of order $(b^{-\bar{d}}) \wedge \bar{d}$, where $0 < b < 1$ depends on the separation of entries of the point the density is evaluated at and $\bar{d}$ is the number of unique entries in that point.

Fundamentally, a class of multivariate functions with permutation invariance is "smaller" than without imposing permutation invariance. This intuition can be formalized by the covering number. As a third contribution, we analyze the covering numbers of two permutation invariant function classes and compare them with their counterparts where permutation invariance is not imposed. We show that the *logarithm* of the covering number for the permutation invariant Hölder class with a boundary condition is reduced by a factor of $d!$. Similarly, for the permutation invariant ellipsoid class, the upper and lower bounds on the *logarithm* of the covering number reduce those of the couterpart without imposing permutation invariance by a factor of $d!$.

Notation. For two functions $f(n,\gamma)$ and $g(n,\gamma)$, let us write $f(n,\gamma) \gtrsim g(n,\gamma)$ if $f(n,\gamma) \geq cg(n,\gamma)$ for a universal constant $c \in (0,\infty)$; similarly, we write $f(n,\gamma) \lesssim g(n,\gamma)$ if $f(n,\gamma) \leq c'g(n,\gamma)$ for a universal constant $c' \in (0,\infty)$; and $f(n,\gamma) \asymp g(n,\gamma)$ if $f(n,\gamma) \gtrsim g(n,\gamma)$ and $f(n,\gamma) \lesssim g(n,\gamma)$.

Definition 1.1 A permutation invariant function is a function $f : \mathbb{R}^d \to \mathbb{R}$ such that $f(\sigma(t)) = f(t)$ for any permutation σ and $t \in \mathbb{R}^d$. We write S_d to denote the set of all permutations.

2 TESTING PERMUTATION INVARIANCE WITH SORTING

Let us consider i.i.d. random samples $\{t_i \in [0,1]^d\}_{i=1}^n$ drawn from an unknown distribution F over $[0,1]^d$, where potentially the dimension $d \to \infty$ as $n \to \infty$. In this section, we are interested in testing the hypothesis:

$$\begin{cases} H_0 : F \text{ is permutation invariant} \\ H_1 : F \text{ is not permutation invariant} \end{cases}.$$

Our test leverages the following proposition.

Proposition 2.1 *A function f on $[0,1]^d$ is permutation invariant if and only if $f(\operatorname{sort} t) = f(t)$ for all $t \in [0,1]^d$.*

Proof: Suppose f is permutation invariant. Then, for any $t \in [0,1]^d$, there exists $\sigma^* \in S_d$ such that $\operatorname{sort} t = \sigma^*(t)$. Consequently, $f(\operatorname{sort} t) = f(\sigma^*(t)) = f(t)$. Suppose that $f(\operatorname{sort} t) = f(t)$ for all $t \in [0,1]^d$. Then, for any $\sigma \in S_d$, we have $f(\sigma(t)) = f(\operatorname{sort} \sigma(t)) = f(\operatorname{sort} t) = f(t)$. □

2.1 The Multiplier Bootstrap Test with a Sorting Trick

We define the empirical CDF

$$F_n(t) := \frac{1}{n}\sum_{i=1}^{n} [t_i \leq t]$$

and the *sorted* empirical CDF

$$\widetilde{F}_n(t) := F_n(\text{sort}\, t) = \frac{1}{n}\sum_{i=1}^{n}[t_i \le \text{sort}\, t].$$

Given $t = (t^1, \cdots, t^d) \in [0,1]^d$, we define $\text{sort}\, t := (t^{\pi(1)}, \cdots, t^{\pi(d)})$ for some permutation $\pi \in S_d$ such that $0 \le t^{\pi(1)} \le t^{\pi(2)} \le \cdots \le t^{\pi(d)} \le 1$.[3]

We propose the following statistics

$$T := \sup_{t\in[0,1]^d} \sqrt{n}\left|\widetilde{F}_n(t) - F_n(t)\right| \tag{1}$$

and the multiplier bootstrap version

$$W := \sup_{t\in[0,1]^d} \frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n}\left([t_i \le \text{sort}\, t] - [t_i \le t]\right)e_i\right|, \quad e_i \sim_{iid} \mathcal{N}(0,1) \tag{2}$$

along with the corresponding bootstrap critical value

$$c_W(\alpha) := \inf\left\{t \in \mathbb{R} \,:\, \mathbb{P}_e[W \le t] \ge 1-\alpha\right\}.$$

The following result shows that our test attains the pre-specified significance level asymptotically.

Theorem 2.2 *Suppose that $d = o(n^{c_0})$ for some $c_0 \in (0, 1/7)$ and the CDF F is continuous. Under H_0, there exists some universal constants $c, C \in (0,\infty)$, such that*

$$\sup_{\alpha\in(0,1)}\left|\mathbb{P}\left[\sup_{t\in[0,1]^d}\sqrt{n}\left|\widetilde{F}_n(t) - F_n(t)\right| > c_W(\alpha)\right] - \alpha\right| < Cn^{-c} \to 0 \text{ as } n \to \infty.$$

Remark. Note that we make *no* assumptions about the dependence among the coordinates of a random vector $t_i = (t_i^1, ...t_i^d)$ for any $i = 1, ..., n$, even though the samples $\{t_i \in [0,1]^d\}_{i=1}^n$ are i.i.d.

Sketch of the proof. Suppose that we have a list of points $\{v_j\}_{j=1,\cdots,N}$ in $[0,1]^d$ which is sufficiently large and well chosen. We should be able to approximate $\sup_{t\in[0,1]^d}\sqrt{n}\left|\widetilde{F}_n(t) - F_n(t)\right|$ by the maximum of the coordinates of

$$\left(\sqrt{n}\left|\widetilde{F}_n(v_1) - F_n(v_1)\right|, \cdots, \sqrt{n}\left|\widetilde{F}_n(v_N) - F_n(v_N)\right|\right).$$

The above can be expressed using a sum of independent random vectors. From there, we apply the result of Chernozhukov et al. (2013). So, the key to the proof is to construct a desired list of points $\{v_j\}$. We let $\{v_j\}$ be the points on a $n^m \times \cdots \times n^m$ grid on $[0,1]^d$ for some $m \ge 4$ and argue: when the grid is fine enough, the probability that the supremum is reached at one of the n^{md} grid point approaches one sufficiently quickly.

The full proof can be found in the supplementary material.

Implementation. In practice, given the data $\{t_i\}_{i=1}^n$, we estimate the supremum

$$\sup_{t\in[0,1]^d}\sqrt{n}\left|\widetilde{F}_n(t) - F_n(t)\right|.$$

Similarly, to compute the supremum W for N_W draws of $\{e_i\}$, one can numerically estimate $c_W(\alpha)$. The test in Theorem 2.2 rejects H_0 at $\alpha \in (0,1)$ significance level (e.g., $\alpha = 0.05$ or 0.01) if $\sup_{t\in[0,1]^d}\sqrt{n}\left|\widetilde{F}_n(t) - F_n(t)\right| > c_W(\alpha)$. There are a number of ways the supremum of

$$Z_n(t; \{e_i\}_{i=1}^n) := \frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n}\left([t_i \le \text{sort}\, t] - [t_i \le t]\right)e_i\right|$$

(i.e., W) can be estimated. The problem of finding the supremum of $\left|\widetilde{F}_n(t) - F_n(t)\right|$ is the special case where $e_i = 1$ for all $i = 1, \cdots, n$, and we write $Z_n(t) := Z_n(t; \{e_i = 1\}_{i=1}^n)$ for convenience.

The supremum of $Z_n(t; \{e_i\}_{i=1}^n)$ can be reached at some unsorted point t^* of the form

$$t^* = \left(t_{i_1}^{a_1}, \cdots, t_{i_d}^{a_d}\right)$$

for some $a_1, \cdots, a_d \in \{1, \cdots, d\}$ and $i_1, \cdots, i_d \in \{1, \cdots, n\}$. In the $\{e_i = 1\}_{i=1}^n$ case, it can be seen that given $t, t' \in [0,1]^d$, $Z_n(t)$ and $Z_n(t')$ can differ from each other by no more than $\frac{1}{\sqrt{n}}$ multiple of the sum of the number of times each coordinate of t' needs to *cross* any of the nd numbers $\{t_i^a\}$ to get to t:

$$|Z_n(t) - Z_n(t')| \le \frac{1}{\sqrt{n}}\sum_{b=1}^{d}\left|\{t_i^a\}_{a=1,\cdots,d;i=1,\cdots,n} \cap [t^b, t'^b]\right|. \tag{3}$$

Therefore, the upper bound estimate for the number of points needed to be searched to find the supremum is $(nd)^d$. When n and d are large, this upper bound is unfavorable.

We propose a more practical solution by first defining a smoothened version of the CDF, given $\varepsilon > 0$:

$$F_{n,h}(t) := \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2^d}\prod_{a=1}^{d}\left(\tanh\left(\frac{t^a - t_i^a}{h}\right) + 1\right)$$

[3] In this paper, $t = (t^1, \cdots, t^d)$ denotes a point (a $d-$dimensional vector) a function is evaluated at, whereas $\{t_i \in [0,1]^d\}_{i=1}^n$ denote random samples drawn from some $d-$variate probability distribution.

and $\widetilde{F}_{n,h}(t) := F_{n,h}(\operatorname{sort} t)$. The parameter h controls smoothness, with $F_{n,h}$ and $\widetilde{F}_{n,h}$ converges in L^1 to their empirical counterpart. The added smoothness provides the needed gradient for a standard maximization algorithm to approach the supremum given a randomly chosen starting point. For our implementation, we use COBYLA, $h = 0.001$, and $n/2$ random starting points, which work well from our observation.

2.2 Related Literature and Discussion

The first test of permutation invariance was for bivariate distributions, proposed in Hollander (1971) and motivated by the question of whether a medical treatment improves patient conditions. A likelihood test for parametric distribution families (such as Gaussian) was discussed in Eaton (1989). Also, see Lyu and Belalia (2023); Yanagimoto and Sibuya (1976) for other proposals. More recently, tests of permutation invariance for distributions with more than two dimensions have been proposed in Bahraoui and Quessy (2022); Harder and Stadtmüller (2017); Kalina and Janáček (2023). The procedure proposed in Kalina and Janáček (2023) tests the null hypothesis of *pairwise* symmetry instead of permutation invariance per se. As acknowledged in Kalina and Janáček (2023), pairwise symmetry is a weaker condition of permutation invariance: the latter implies the former but *not* vice versa. We illustrate this point with a real-world data set in Section 5.3 and demonstrate that our test has more power. The papers Bahraoui and Quessy (2022); Harder and Stadtmüller (2017) propose tests of permutation invariance for multivariate copulas, which is also a weaker condition of permutation invariance of the distribution. This fact follows from Sklar's theorem that a multivariate distribution is permutation invariant if all its marginals are equal and its corresponding copula is permutation invariant.

Unlike the aforementioned literature, our procedure tests *directly* whether the coordinates of a random vector from an unknown multivariate probability distribution are permutable. This characteristic is shared by the Monte Carlo test proposed in Chiu and Bloem-Reddy (2023) and the concept of group-generating sets proposed in Soleymani et al. (2025). Both Chiu and Bloem-Reddy (2023) and Soleymani et al. (2025) consider the Maximum Mean Discrepancy (MMD) between the probability measures. In particular, the test proposed in Chiu and Bloem-Reddy (2023) samples random elements from the permutation group; Soleymani et al. (2025) shows that the notion of generating sets under MMD provides a sufficient and necessary condition for permutation invariance. This result of Soleymani et al. (2025) and the multiplier bootstrap trick can be integrated in a fashion similar to what was proposed in Section 2.1 to obtain an alternative version of Theorem 2.2, for which we refer the readers to (10)-(11) and Theorem B.1 in the supplementary materials. Like Theorem 2.2, Theorem B.1 attains the pre-specified significance level asymptotically.

As demonstrated in Theorem B.1, integrating the idea of generation sets under MMD in Soleymani et al. (2025) with the multiplier bootstrap trick allows for a faster growth rate of the dimension, e.g. $d = o(e^{n^{c_0}})$ for some $c_0 \in (0, 1/7)$, if the generating set is chosen to be $S := \{(1,2), (1,3), \cdots, (1,d)\}$. This much faster growth rate of d in Theorem B.1 may appear to be a substantial improvement at first glance. However, it should be noted that the choice of a generating set is not unique and, as shown in (Soleymani et al., 2025, Corollary 6.2), there is a trade-off between $|S|$ and the maximum length of the minimal representations of the group elements. In the context of Theorem B.1, this trade-off translates to one between the growth rate of d and the convergence rate Cn^{-c}. Importantly, the trade-off revealed in (Soleymani et al., 2025, Corollary 6.2) can also be crucial to the power of the test (10)-(11) in the supplementary materials. Comparing the power of the various procedures opens an important question for future research.

2.2.1 More on the Sorting Trick

Earlier versions of this paper in March and May 2024 (e.g., Chaimanowong and Zhu (2024)) also propose the sorting trick in kernel ridge regressions and kernel interpolations. Particularly, (Chaimanowong and Zhu, 2024, Theorem 5.2) bounds the error from approximating a permutation invariant function in a reproducing Kernel Hilbert space with a function constructed based on the sorted reproducing kernel and provides a computationally efficient embedding scheme. These results are no longer included in this paper due to the space limit and to have a better focus. To our best knowledge, this paper and (Chaimanowong and Zhu, 2024, Theorem 5.2) are the first to explore the sorting trick in testing and approximation, which showcase the usefulness of the sorting trick in various problems.

3 ESTIMATING PERMUTATION INVARIANT DENSITIES WITH AVERAGING

In this section, we focus on permutation invariant density functions and show a way to exploit permutation variance in the (local) kernel density estimation method. Kernel density estimation is among the foremost practical methods in statistics and machine learnning, and serves as a building block of numerous nonparametric and semiparametric estimators.

The averaging trick. Given $t = (t^1, \cdots, t^d) \in \mathbb{R}^d$, the *standard* kernel density estimator is given by

$$\hat{f}(t) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{t - t_i}{h}\right).$$

To exploit the permutation invariance structure, we propose the *averaged* kernel density estimator

$$\tilde{f}(t) = \frac{1}{\bar{d}} \sum_{\sigma \in S_d^*} \hat{f}(\sigma(t)),$$

where $\bar{d}$ and S_d^* are defined as follows.

Definition 3.1 Let $t \in \mathbb{R}^d$ be a vector with $\bar{d}$ distinct entries. The set S_d^* can be *any* set consisting of $\bar{d}$ permutations such that any $\sigma, \pi \in S_d^*$ take different values on exactly $\bar{d}$ positions. When $\bar{d} < d$, a set satisfying the aforementioned property is not unique.

For example, take $t = (1, 2, 3)$. Then $\bar{d} = 3$ and $S_d^* = \{(1,2,3), (3,1,2), (2,3,1)\}$. As another example, take $t = (1, 1, 2)$. Then $\bar{d} = 2$ and the choice of S_d^* can be any of the three sets: $\{(1,1,2), (2,1,1)\}$, $\{(1,1,2), (1,2,1)\}$, $\{(2,1,1), (1,2,1)\}$. Clearly, when all entries of t take the same value, $\bar{d} = 0$ and $\hat{f}(t) = \tilde{f}(t)$.

Our theoretical guarantees are based on the following assumption.

Assumption 3.2 The random samples $\{t_i \in [0,1]^d\}_{i=1}^n$ are independently drawn from a pdf f, which is permutation invariant and twice differentiable with bounded derivatives. The non-negative kernel K satisfies: (a) $\int_{-\infty}^{\infty} K(v)dv = 1$; (b) $K(v) = K(-v)$ for all v; (c) $\int_{-\infty}^{\infty} vv^T K(v)dv < \infty$; (d) $\int_{-\infty}^{\infty} K(v)^2 dv < \infty$.

In what follows, we compare the bias and variance of $\tilde{f}$ with those of $\hat{f}$. When writing *higher order terms*, we mean that these terms have a smaller order than the leading term(s).

Lemma 3.3 *Let Assumption 3.2(a-c) hold. Suppose $h \to 0$. Then,*

$$\mathbb{E}[\hat{f}(t)] - f(t) = \mathbb{E}[\tilde{f}(t)] - f(t)$$
$$= \frac{h^2}{2} tr\left(\frac{\partial^2 f(t)}{\partial t \partial t^T} \int vv^T K(v)dv\right) + \textit{higher order terms.}$$

This result shows that the biases of $\tilde{f}$ and $\hat{f}$ have the same leading term.

Lemma 3.4 *Let Assumption 3.2 hold. Suppose $n \to \infty$ and $h \to 0$ while $nh^d \to \infty$.*

(i) Then, the variances are computed as

$$\mathbb{V}(\hat{f}(t)) = \frac{f(t)}{nh^d} \int_{-\infty}^{\infty} K(v)^2 dv + \textit{higher order terms} \tag{4}$$

and

$$\mathbb{V}(\tilde{f}(t)) = \frac{f(t)}{(\bar{d})^2 nh^d} \sum_{\pi, \sigma \in S_d^* \, s.t. \, \pi \neq \sigma} K * K\Big(\frac{\sigma(t) - \pi(t)}{h}\Big)$$
$$+ \frac{1}{\bar{d}} \frac{f(t)}{nh^d} \int_{-\infty}^{\infty} K(v)^2 dv + \textit{ higher order terms} \tag{5}$$

where $*$ denotes the convolution.

(ii) Further, if we consider the product kernel $K(v) = k(v^1)k(v^2)...k(v^d)$ where $k(\cdot)$ is a non-negative univariate kernel satisfying conditions (a)-(d) in Assumption 3.2 and $k(v)$ decreases in $|v|$,[4] *then*

$$K * K\Big(\frac{\sigma(t) - \pi(t)}{h}\Big) \leq b^{\bar{d}} \int K(v)^2 dv \tag{6}$$

where $b \in (0, 1)$.

The proofs of Lemma 3.3 and Lemma 3.4 can be found in the supplementary material.

Remark. The form of our estimator $\tilde{f}(t)$ for $f(t)$ resembles a U-statistic at first glance, but it is *not* a U-statistic: in particular, U-statistics permute samples that satisfy i.i.d. or weak dependence conditions, while our estimator $\tilde{f}(t)$ permutes the coordinates of the *fixed* evaluation point t. The result in (4) is well known in the literature of kernel density estimation (see, e.g., Tsybakov (2009)). From (6) and the first two terms in (5), compared to the standard product kernel density estimator $\hat{f}(t)$, we see a reduction in the variance of the *averaged* kernel density estimator $\tilde{f}(t)$ when $\bar{d} \geq 1$. The larger $\bar{d}$ and $|\frac{\sigma_j(t) - \pi_j(t)}{h}|$ are (where $\sigma_j(t)$ denotes the jth coordinate of $\sigma(t)$ for $j = 1, ..., d$), the smaller $b^{\bar{d}}$ is, and hence the greater the reduction. Given that the point-wise mean square error $\text{MSE}(t) = \mathbb{V}(t) + (\text{Bias}(t))^2$ and Lemma 3.3, the reduction in the variance implies that $\tilde{f}(t)$ has a smaller MSE(t) than $\hat{f}(t)$ when $\bar{d} \geq 1$. Consequently, $\tilde{f}$ also has a smaller mean integrated squared error (MISE) than $\hat{f}$.

[4] Examples of kernels satisfying these assumptions include the triangular kernel, the Gaussian kernel, the cosine kernel, the Epanechnikov kernel, the quartic kernel, the triweight kernel, the tricube kernel, the logistic kernel, the sigmoid function and etc.

4 A FUNDAMENTAL PERSPECTIVE

Fundamentally, a class of multivariate functions with permutation invariance has a smaller "size" than without imposing permutation invariance. A measure of "size" is the metric entropy such as covering number (Kolmogorov and Tikhomirov, 1959). Metric entropy is foundational to important theoretical objects such as Rademacher and sub-Gaussian complexity. In the following, we compare the metric entropy of two permutation invariant function classes with the metric entropy of their counterparts where permutation invariance is not imposed.

Let $p = (p_j)_{j=1}^d$ and $P = \sum_{j=1}^d p_j$ where p_js are non-negative integers; $x = (x_j)_{j=1}^d$ and $x^p = \prod_{j=1}^d x_j^{p_j}$. Write $D^p f(x) = \partial^P f / \partial x_1^{p_1} \cdots \partial x_d^{p_d}$.[5]

Definition 4.1 [Hölder classes with a boundary condition] For a non-negative integer γ, let the permutation invariant Hölder class $\mathcal{U}^{perm}$ be the class of functions such that any function $f \in \mathcal{U}^{perm}$ satisfies: (1) f is continuous and permutation invariant on $[0,1]^d$, and all partial derivatives of f exist for all p with $P := \sum_{k=1}^d p_k \leq \gamma$; (2) $|D^p f(x)| \leq C$ for all p with $P = k$ $(k = 0, ..., \gamma)$ and $x \in [0,1]^d$ such that $D^p f(0) = 0$ (the boundary condition), where $D^0 f(x) = f(x)$; (3) $\left|D^p f(x) - D^p f(x')\right| \leq C \left|x - x'\right|_\infty$ for all p with $P = \gamma$ and $x, x' \in [0,1]^d$. When permutation invariance is not imposed, we denote the Hölder class by $\mathcal{U}$.

Theorem 4.2 *We have*

$$\log N_2(\delta, \mathcal{U}^{perm}) \asymp \log N_\infty(\delta, \mathcal{U}^{perm}) \asymp \frac{1}{d!} \log N_\infty(\delta, \mathcal{U}) \asymp \frac{1}{d!} b_{d,\gamma}^d \delta^{\frac{-d}{\gamma+1}}$$

where $b_{d,\gamma}$ is a function of (d,γ) only and independent of δ, $N_2(\delta, \mathcal{U}^{perm})$ denotes the $\delta-$covering number of $\mathcal{U}^{perm}$ with respect to the L^2-norm and $N_\infty(\delta, \mathcal{U}^{perm})$ ($N_\infty(\delta, \mathcal{U})$) denotes the $\delta-$covering number of $\mathcal{U}^{perm}$ (respectively, $\mathcal{U}$) with respect to the sup norm.

The proof of Theorem 4.2 can be found in the supplementary material.

Definition 4.3 [Ellipsoid classes] Given a sequence of non–negative real numbers $\{\mu_k\}_{k \in \mathbb{Z}_{\geq 0}^d}$ such that $\sum_{k \in \mathbb{Z}_{\geq 0}^d} \mu_k < \infty$, we define the ellipsoid

$$\mathcal{E} := \left\{ (\beta_k)_{k \in \mathbb{Z}_{\geq 0}^d} \mid \sum_{k \in \mathbb{Z}_{\geq 0}^d} \frac{\beta_k^2}{\mu_k} \leq 1 \right\},$$

and its permutation invariant subset:

$$\mathcal{E}^{perm} := \left\{ (\beta_k)_{k \in \mathbb{Z}_{\geq 0}^d} \in \mathcal{E} \mid \beta_{\text{sort}\, k} = \beta_k \right\}.$$

Consider a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ of functions over $[0,1]^d$ with a Mercer's kernel $\mathcal{K}$, whose associated eigenfunctions $\{\phi_k\}_{k \in \mathbb{Z}_{\geq 0}^d}$ satisfy $\phi_k(\sigma t) = \phi_{\sigma^{-1} k}(t)$ for all $\sigma \in S_d$, and we denote the associated eigenvalues by $\{\mu_k\}_{k \in \mathbb{Z}_{\geq 0}^d}$. By definition, $\{\phi_k\}_{k \in \mathbb{Z}_{\geq 0}^d}$ gives an orthonormal basis for $L^2([0,1]^d, \mathbb{P})$. Since the kernel is continuous on the compact domain $[0,1]^d$, we have the convergence $\sum_{k \in \mathbb{Z}_{\geq 0}^d} \mu_k = \int_{[0,1]^d} \mathcal{K}(t,t) d\mathbb{P}(t) < \infty$. It is well known (Wainwright, 2019, Corollary 12.26) that $\mathcal{H}$ can be identified with the ellipsoid $\mathcal{E}$ where any $f \in \mathcal{H}$ can be written in the form $f = \sum_{k \in \mathbb{Z}_{\geq 0}^d} \beta_k \phi_k$. Now, consider the subspace of permutation invariant functions

$$\mathcal{H}^{perm} := \{ f \in \mathcal{H} \mid f(\text{sort}\, t) = f(t), \ \forall t \} \subset \mathcal{H}.$$

It follows that if $f = \sum_{k \in \mathbb{Z}_{\geq 0}^d} \beta_k \phi_k \in \mathcal{H}^{perm}$, then

$$\begin{aligned} \sum_{k \in \mathbb{Z}_{\geq 0}^d} \beta_k \phi_k(t) &= \sum_{k \in \mathbb{Z}_{\geq 0}^d} \beta_k \phi_k(\sigma t) \\ &= \sum_{k \in \mathbb{Z}_{\geq 0}^d} \beta_{\sigma k} \phi_{\sigma k}(\sigma t) = \sum_{k \in \mathbb{Z}_{\geq 0}^d} \beta_{\sigma k} \phi_k(t) \end{aligned}$$

for all t and σ. Hence, $\beta_{\text{sort}\, k} = \beta_k$ and $\mathcal{H}^{perm}$ can be identified with $\mathcal{E}^{perm}$.

Define the norms $\|.\|_{l^2}$ and $\|.\|_{l^2}^{perm}$ on $\mathcal{E}$ and $\mathcal{E}^{perm}$,

$$\|\beta - \beta'\|_{l^2} := \sqrt{\sum_{k \in \mathbb{Z}_{\geq 0}^d} (\beta_k - \beta_k')^2},$$

$$\|\beta - \beta'\|_{l^2}^{perm} := \sqrt{\sum_{k \in \text{sort}\, \mathbb{Z}_{\geq 0}^d} (\beta_k - \beta_k')^2},$$

for any $\beta, \beta' \in l^2(\mathbb{Z}_{\geq 0}^d)$. The following result shows the reduction in metric entropy from imposing permutation invariance.

Theorem 4.4 *There exists $\underline{g}, \bar{g} : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ such that*

$$\underline{g}(\delta) \lesssim \log N(\delta, \mathcal{E}, \|.\|_{l^2}) \lesssim \bar{g}(\delta)$$

[5] We use t in most of the results in this paper but use x here to avoid confusion in the notation.

and

$$\frac{1}{d!}\underline{g}(\delta) \lesssim \log N(\delta, \mathcal{E}^{perm}, \|.\|_{l^2}^{perm}) \lesssim \frac{1}{d!}\bar{g}(\delta)$$

where $N(\delta, \mathcal{E}, \|.\|_{l^2})$ denotes the $\delta-$covering number of $\mathcal{E}$ with respect to the $\|.\|_{l^2}-$norm and $N(\delta, \mathcal{E}^{perm}, \|.\|_{l^2}^{perm})$ denotes the $\delta-$covering number of $\mathcal{E}^{perm}$ with respect to the $\|.\|_{l^2}^{perm}-$norm.

The proof of Theorem 4.4 can be found in the supplementary material.

Like $\mathcal{U}^{perm}$, both the lower and upper bounds on the logarithm of the covering number for $\mathcal{E}^{perm}$ are reduced by a factor of $d!$ when permutation invariance is imposed.

An example. Let $\mathcal{H} = W_{per}^{s,2}([0,1]^d)$ be the periodic Sobolev space: the subspace of the Sobolev space $W^{s,2}([0,1]^d)$ for functions f such that $f(t+k) = f(t)$ for any $k \in \mathbb{Z}^d$. This space is an RKHS when $s > d/2$. In the supplementary material, we derive the reproducing kernel and the eigenvalues of $W_{per}^{s,2}([0,1]^d)$. From these results, we can see that $\bar{g}(\delta) = \underline{g}(\delta) := \left(\frac{1}{\delta}\right)^{d/s}$ in Theorem 4.4. Therefore, we have the sharp result: $\log N(\delta, \mathcal{E}^{perm}, \|, \|_{l^2}^{perm}) \asymp \frac{1}{d!} \log N(\delta, \mathcal{E}, \|, \|_{l^2})$.

Related work. Other metric entropy calculations in the literature of invariant learning show up in Chen et al. (2023); Sokolic et al. (2017). However, these results concern a different type of symmetry that requires the assumption $\|\sigma(t) - \sigma'(t)\|_2 > 2\delta$ for all the underlying symmetry transformations $\sigma \neq \sigma'$ in Chen et al. (2023); Sokolic et al. (2017). This assumption is neither desirable nor needed for our results; for example, for all $t \in [0,1]^d$ such that all entries are the same, this assumption would be too restrictive (to our setup) as it does not allow $\sigma(t) = t$ for some non-identity permutation σ.

5 NUMERICAL STUDIES

5.1 Testing Permutation Invariance

Throughout we use $N_W = 1000$ to numerically estimate $c_W(\alpha)$, and $N = 1000$ Monte–Carlo replications. To estimate the suprema, T and W, we use the COBYLA maximization algorithm on the smoothened empirical CDF described in Section 2.1. For key performance indicators, we denote by "Pow" the power of the test (the probability of rejecting H_0 when H_0 is false) and "Cov" the coverage of the test (the probability of not rejecting H_0 when H_0 is true). For the dimension $d = 2$, we demonstrate the performance of our test given a various number of samples over $[0,1]^d$ from the normal distribution

$$\mathcal{N}\left((\mu_1, 0.5), \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & 0.01 \end{pmatrix}\right)$$

modulo $\mathbb{Z}^d$.[6] We adjust (μ_1, σ_1^2) to create various setups for the experiment, with the $(\mu_1, \sigma_1^2) = (0.5, 0.01)$ being the control case where the distribution is permutation invariant. The results are presented in Table 1.

Next, we study the performance of our test with 100 samples drawn from the normal distribution:

$$\mathcal{N}\left(\underbrace{(\mu_1, 0.5, \cdots, 0.5)}_{d}, 0.01 \cdot I_d\right)$$

modulo $\mathbb{Z}^d$, where we consider various dimensions $d = 3, 4, 5$. The results are presented in Table 2.

Generally, for a fixed d, we can see the improvement in performance with higher n. However, with a fixed n, it becomes increasingly challenging with higher d for the optimization algorithm to estimate the suprema T and W, where a higher n would also be needed.

	(μ_1, σ_1^2)	(0.5, 0.01)		(0.4, 0.01)		(0.5, 0.05)	
n	α	95%	99%	95%	99%	95%	99%
100	Cov	**0.96**	**0.99**	< 0.01	< 0.01	0.09	0.29
	Pow	0.04	0.01	$>$ **0.99**	$>$ **0.99**	**0.91**	**0.71**
200	Cov	**0.95**	**0.99**	< 0.01	< 0.01	< 0.01	0.01
	Pow	0.05	0.01	$>$ **0.99**	$>$ **0.99**	$>$ **0.99**	**0.99**
300	Cov	**0.95**	**0.99**	< 0.01	< 0.01	< 0.01	< 0.01
	Pow	0.05	0.01	$>$ **0.99**	$>$ **0.99**	$>$ **0.99**	$>$ **0.99**

Table 1: Simulation results with $d = 2$.

	μ_1	0.5		0.4	
d	α	95%	99%	95%	99%
3	Cov	**0.96**	**0.99**	< 0.01	0.01
	Pow	0.04	0.01	$>$ **0.99**	**0.99**
4	Cov	**0.95**	$>$ **0.99**	< 0.01	0.01
	Pow	0.05	< 0.01	$>$ **0.99**	**0.99**
5	Cov	**0.96**	**0.99**	0.03	0.11
	Pow	0.04	0.01	**0.97**	**0.89**

Table 2: Simulation results with $n = 100$.

5.2 Estimating Permutation Invariant Densities

We use the product of univariate triangular kernels for illustration, although other choices that satisfy the

[6] For example, if a vector $(1.4, 0.7)$ is drawn from $\mathcal{N}\left((\mu_1, 0.5), \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & 0.01 \end{pmatrix}\right)$, then after $\mathbb{Z}^d$ modulo, we produce the sample point $(0.4, 0.7)$ on $[0,1]^d$.

assumptions in Lemma 3.4(ii) would also work. For the results in Table 3, we draw 10000 samples from $N(0, I_d)$ and compare $\hat{f}(t)$ with $\tilde{f}(t)$ at different values of t when dimensions range from 3 to 5. The simulations in Table 3 have $h = 3\left(\frac{1}{n}\right)^{\frac{1}{d+4}}$, and are repeated 1000 times. Figure 1 focuses on $t = (0, 0.25, 0.5, 0.75)$ and exhibits the biases and variances of the standard kernel density estimator $\hat{f}(t)$ and the averaged kernel density estimator $\tilde{f}(t)$ with n increasing from 1000 to 10000. Compared to $\hat{f}(t)$, $\tilde{f}(t)$ exhibits much lower variances and similar biases.

t	$\hat{f}(t)$ bias	$\tilde{f}(t)$ bias	$\hat{f}(t)$ variance	$\tilde{f}(t)$ variance
	(all numbers below $\times 10^{-3}$)			
(0,0.5,1)	-0.1202	-0.3107	0.0523	0.0170
(0, 0.25, 0.5)	-1.0118	-0.9651	0.0802	0.0282
(0, 0.25, 0.5, 0.75)	-0.3838	-0.2504	0.0328	0.0082
(0, 0.25, 0.5, 0.75, 0.75)	-0.0913	-0.1394	0.0114	0.0026
(0, 0.25, 0.5, 0.75, 1)	-0.0332	-0.1423	0.0087	0.0017

Table 3: Bias and variance comparisons with $n = 10000$, $h = 3\left(\frac{1}{n}\right)^{\frac{1}{d+4}}$ and 1000 replications.

5.3 Real-world Data Examples

We demonstrate our proposed multiplier bootstrap test with two real-world data sets. We first consider the Australian Athletes (AA) data set from Maindonald et al. (2015), which consists of $n = 202$ observations. We test the permutation invariance hypothesis (H_0) of the $d = 3$ features: red blood cell count, white blood cell count, and hemoglobin concentration. A quick visual inspection reveals that some variables appear to have a larger range than others. This observation motivates us to normalize each variable to $[0, 1]$. Table 4 shows that our test rejects H_0.

As another example, we study the Wine Quality data set from Paulo et al. (2009). The data set is used for quality classifier training based on the given wine's physicochemical composition. We investigate the permutation invariance (H_0) of the red wine's $d = 3$ acidity features: fixed acidity, volatile acidity, and citric acid. The full data set consists of over 1000 observations. For our purpose, we randomly subsample $n = 202$ observations to facilitate the comparison with our test on the AA data set. We also normalize each variable to $[0, 1]$. Table 4 shows that our test rejects H_0.

To explore how sensitive our test is to detect the lack of permutation invariance in the presence of pairwise symmetry, we look at the pairwise correlations for each data set. One can see that the scatter plots in the left panel for the (normalized) AA data set appears more symmetric along the diagonal than the plots in the right panel for the (normalized) Wine Quality data set. However, it is important to note that pairwise

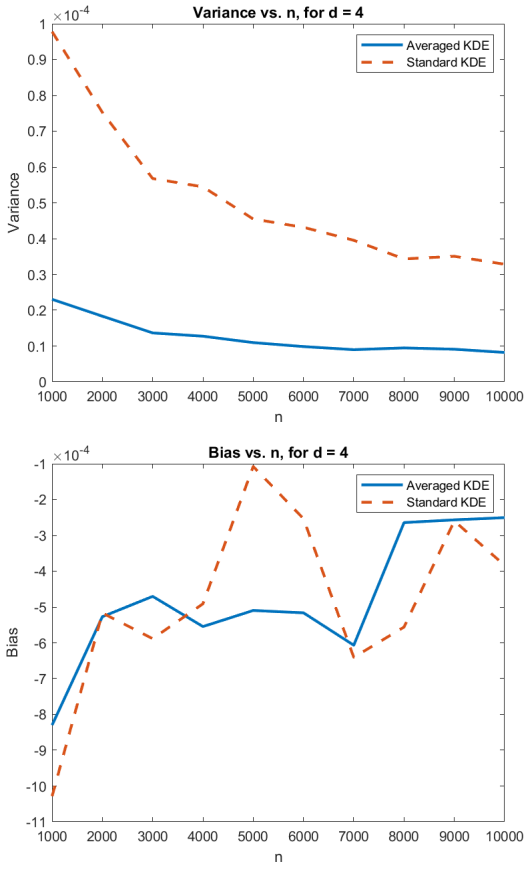

Figure 1: Variance and bias at $t = (0, 0.25, 0.5, 0.75)$ as n increases from 1000 to 10000 using the bandwidth of $h = 3\left(\frac{1}{n}\right)^{\frac{1}{d+4}}$ and 1000 replications. The variance of the averaged kernel density estimator is consistently smaller than that of the standard kernel density estimator throughout.

Data set	α	T	$c_W(\alpha)$	Reject H_0
Australian Athletes	95%	2.814	1.858	Yes
	99%		2.171	Yes
Wine Quality	95%	3.658	2.117	Yes
	99%		2.556	Yes

Table 4: Results from our proposed bootstrap test for two real-world data sets with $n = 202, d = 3$.

Data set	α	p_{12}	p_{13}	p_{23}	Reject H_0
Australian Athletes	95%	0.014	0.073	0.049	Yes
	99%				No
Wine Quality	95%	0.015	< 0.001	0.009	Yes
	99%				Yes

Table 5: Results from multiple pairwise Hollander Bivariate Symmetry tests with $n = 202, d = 3$.

symmetry is implied by permutation invariance but *not* vice versa. To illustrate this fact, we run the multiple comparison test of $D = d(d-1)/2$ pairwise bivariate symmetries, as proposed in Kalina and Janáček (2023). We use the R package implementation NSM3 (Schneider et al., 2023) of Hollander Bivariate Symmetry test (Hollander, 1971). We denote by p_{ij}, the p-value for the test on dimension-(i,j)-pair. To ensure the family-wise error rate (FWER) $< \alpha$, we reject H_0 if $p_{ij} < \alpha/D$ for some (i,j) pair, according to the Bonferroni method. Table 5 shows that the multiple pairwise test is able to reject H_0 for the Wine Quality data set but fails to do so for the AA data set at $\alpha = 0.01$ significance level. Despite that the scatter plots for AA indicate pairwise symmetry, our test still rejects the null hypothesis of permutation invariance at both $\alpha = 0.05$ and $\alpha = 0.01$ levels. This fact shows that our test has more power.

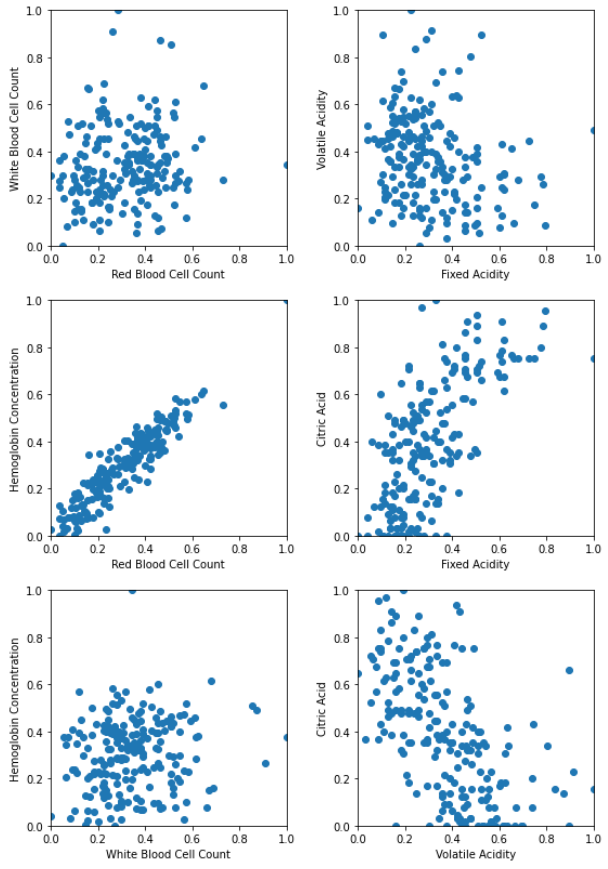

Figure 2: Scatter plots ($n = 202$) for all pairs of dimension of each data set. The variables have been normalized to $[0,1]$.

Acknowledgments

Ying Zhu is grateful to the Society of Hellman Fellows at University of California, and thanks Connor Goldstick's research assistance, which was funded by Zhu's Hellman Fellowship Award.

References

Anderson, M. J. and Robinson, J. (2001). Permutation tests for linear models. *Australian & New Zealand Journal of Statistics*, 43(1):75–88.

Bahraoui, T. and Quessy, J.-F. (2022). Tests of multivariate copula exchangeability based on lévy measures. *Scandinavian Journal of Statistics*, 49(3):1215–1243.

Bloem-Reddy, B., Whye, Y., et al. (2020). Probabilistic symmetries and invariant neural networks. *Journal of Machine Learning Research*, 21(90):1–61.

Chaimanowong, W. and Zhu, Y. (2024). Permutation invariant functions: statistical tests, density estimation, and computationally efficient embedding. *arXiv preprint arXiv:2403.01671v3*.

Chen, Z., Katsoulakis, M., Rey-Bellet, L., and Zhu, W. (2023). Sample complexity of probability divergences under group symmetry. In *International Conference on Machine Learning*, pages 4713–4734. PMLR.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786 – 2819.

Chiu, K. and Bloem-Reddy, B. (2023). Hypothesis tests for distributional group symmetry with applications to particle physics. In *NeurIPS 2023 AI for Science Workshop*.

Eaton, M. L. (1989). Group invariance applications in statistics. IMS.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.

Harder, M. and Stadtmüller, U. (2017). Testing exchangeability of copulas in arbitrary dimension. *Journal of Nonparametric Statistics*, 29(1):40–60.

Hollander, M. (1971). A nonparametric test for bivariate symmetry. *Biometrika*, 58(1):203–212.

Kalina, J. and Janáček, P. (2023). Testing exchangeability of multivariate distributions. *Journal of Applied Statistics*, 50(15):3142–3156.

Kolmogorov, A. N. and Tikhomirov, V. M. (1959). ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk*, 14(2):3–86.

Koning, N. W. (2024). More power by using fewer permutations. *Biometrika*, 111(4):1405–1412.

Koning, N. W. and Hemerik, J. (2024). More efficient exact group invariance testing: using a representative subgroup. *Biometrika*, 111(2):441–458.

Kuchibhotla, A. K. (2020). Exchangeability, conformal prediction, and rank tests. *arXiv preprint arXiv:2005.06095*.

Lyu, G. and Belalia, M. (2023). Testing symmetry for bivariate copulas using bernstein polynomials. *Statistics and Computing*, 33(6):128.

Maindonald, J. H., Braun, W. J., and Braun, M. W. J. (2015). Package 'daag'. *Data Analysis and Graphics Data and Functions*.

Novak, E., Ullrich, M., Woźniakowski, H., and Zhang, S. (2018). Reproducing kernels of sobolev spaces on $\mathbb{R}^d$ and applications to embedding constants and tractability. *Analysis and Applications*, 16(05):693–715.

Paulo, C., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009). Wine Quality. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C56S3T.

Ramdas, A., Barber, R. F., Candès, E. J., and Tibshirani, R. J. (2023). Permutation tests using arbitrary permutation distributions. *Sankhya A*, 85(2):1156–1177.

Schneider, G., Chicken, E., Becvarik, R., and Schneider, M. G. (2023). Package 'nsm3'.

Sokolic, J., Giryes, R., Sapiro, G., and Rodrigues, M. (2017). Generalization error of invariant classifiers. In *Artificial Intelligence and Statistics*, pages 1094–1103. PMLR.

Soleymani, A., Tahmasebi, B., Jegelka, S., and Jaillet, P. (2025). A robust kernel statistical test of invariance: Detecting subtle asymmetries. MIT, TUM.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, NY, 1 edition.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.

Yanagimoto, T. and Sibuya, M. (1976). Test of symmetry of a bivariate distribution. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 105–115.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **Yes**.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **Yes**.
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. **Yes**.
   (b) Complete proofs of all theoretical results. **Yes**.
   (c) Clear explanations of any assumptions. **Yes**.
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). **Yes**. A GitHub URL to the code for the numerical studies is provided.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **Yes**. Details are given in §5.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). **Yes**. Details are given in §5.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). **Not Applicable**. No special hardware requirements are needed.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. **Yes**.
   (b) The license information of the assets, if applicable. **Yes**. Data we used for the numerical studies are available on the UCI Machine Learning Repository.
   (c) New assets either in the supplemental material or as a URL, if applicable. **Yes**.
   (d) Information about consent from data providers/curators. **Not Applicable**. Please refer to UCI Machine Learning Repository user agreement.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. **Not Applicable**.
5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. **Not Applicable**.

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. **Not Applicable**.

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. **Not Applicable**.

Permutation Invariant Functions: Statistical Testing, Density Estimation, and Metric Entropy (Supplementary Materials)

A PROOFS

A.1 Proof of Theorem 2.2

Proof: Consider i.i.d. samples $\{t_i\}_{i=1,\cdots,n}$ drawn from a distribution with a continuous CDF F. Let us choose $\{z_k \in [0,1]\}_{k=1\cdots,n^m}$ for some $m \geq 4$ such that each slab

$$B_k^a := \{t \in [0,1]^d \mid z_{k-1} \leq t^a \leq z_k\}$$

contains equal probability for any coordinate $a = 1, \cdots, d$ and $k = 1, \cdots, n^m$: i.e.

$$\int_{B_k^a} 1 \cdot dF(t) = \frac{1}{n^m},$$

where we take $z_0 = 0$ by convention, and let us define: $\{v_{k_1,\cdots,k_d} := (z_{k_1}, \cdots, z_{k_d}) \in [0,1]^d\}_{k_1,\cdots,k_d=1,\cdots,n^m}$. Fix a set $A \subset [0,1]^d$ bounded away from the boundaries and the diagonal of $[0,1]^d$ such that

$$\inf_{v \in A} \int_{[0,1]^d} \left([t \leq \text{sort}\, v] - [t \leq v]\right)^2 dF(t) > 0$$

and let $\{\widetilde{v}_j \subset A\}_{j=1,\cdots,n^{md}}$ be any subset of n^{md} points in A. For convenience, let us relabel the elements of $\{v_{k_1,\cdots,k_d}\} \cup \{\widetilde{v}_j\}$ in some ways as $\{v_j\}_{j=1,\cdots,2n^{md}}$ where $v_j \in \{v_{i_1,\cdots,i_d}\}$ for $j = 1, \cdots, n^{md}$, and $v_j \in \{\widetilde{v}_j\}_{j=1,\cdots,n^{md}}$ for $j = n^{md}+1, \cdots, 2n^{md}$. For $j = 1, \cdots, n^{md}$, define the box

$$B_j := \{t \in [0,1]^d \mid t \leq v_j \text{ and } t \not\leq v_{j'} \text{ for any } v_{j'} < v_j\},$$

and let us assume that $\int_{B_j} 1 \cdot dF(t) \leq \frac{1}{n^{(m-1)d}}$ [7] Alternatively, we can write $B_j = B_{k_1,\cdots,k_d} := \bigcap_{a=1}^d B_{k_a}^a$ if $v_j = v_{k_1,\cdots,k_d}$. We apply the results of Chernozhukov et al. (2013) with

$$x_{ij} := \begin{cases} [t_i \leq \text{sort}\, v_j] - [t_i \leq v_j], & i = 1, \cdots, \lceil n/2 \rceil - 1; j = 1, \cdots, 2n^{md} \\ [t_i \leq \text{sort}\, v_{j+n^{md}}] - [t_i \leq v_{j+n^{md}}], & i = \lceil n/2 \rceil, \cdots, n; j = 1, \cdots, n^{md} \\ [t_i \leq \text{sort}\, v_{j-n^{md}}] - [t_i \leq v_{j-n^{md}}], & i = \lceil n/2 \rceil, \cdots, n; j = n^{md}+1, \cdots, 2n^{md} \\ -x_{i,j-2n^{md}}, & \forall i; j = 2n^{md}+1, \cdots, 4n^{md} \end{cases}.$$

[7] If necessary, we may start the construction from finding $\{\widetilde{z}_k^a\}_{k=1,\cdots,2n^{m-1};a=1,\cdots,d}$, $\widetilde{z}_{i_1,\cdots,i_d} := (\widetilde{z}_{i_1}^1, \cdots, \widetilde{z}_{i_d}^d) \in [0,1]^d$, such that each $\widetilde{B}_{i_1,\cdots,i_d} := \{t \in [0,1]^d \mid \widetilde{z}_{i_1-1,\cdots,i_d-1} \leq t \leq \widetilde{z}_{i_1,\cdots,i_d}\}$ contains equal probability of $\frac{1}{2^d n^{(m-1)d}}$. Hence each slab $\widetilde{B}_k^a := \{t \in [0,1]^d \mid \widetilde{z}_{k-1}^a \leq t^a \leq \widetilde{z}_k^a\}$ contains equal probability of $\frac{1}{2n^{m-1}}$. The needed symmetric grid $\{z_k\}_{k=1,\cdots,n^m}$ must be 'smaller' than each $\{\widetilde{z}_k^a\}$ in a sense that for any $a = 1, \cdots, p$ we have $[z_{k-1}, z_k] \subset [\widetilde{z}_{k'-1}^a, \widetilde{z}_{k'+1}^a]$ for some $k' \in \{1, \cdots, 2n^{m-1}-1\}$, which implies $\mathbb{P}[B_j] \leq 2^d \cdot \frac{1}{2^d n^{(m-1)d}} = \frac{1}{n^{(m-1)d}}$.

Then $x_i = (x_{i1}, \cdots, x_{i,2n^{md}}) \in \mathbb{R}^{2n^{md}}$ for $i = 1, \cdots, n$ are independent random vectors. Moreover, it follows from the permutation invariance of the CDF F that each x_i is centered, i.e. $\mathbb{E}[x_{ij}] = 0$. For convenience, we introduce the following notation

$$Z_n(t; \{c_i\}) := \frac{1}{\sqrt{n}} \left| \sum_{i=1}^{n} \left([t_i \leq \operatorname{sort} v_j] - [t_i \leq v_j] \right) c_i \right|$$

and

$$T_0 := \max_{1 \leq j \leq 4n^{md}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{ij} = \max_{1 \leq j \leq 2n^{md}} Z_n(v_j; \{c_i = 1\})$$

$$W_0 := \max_{1 \leq j \leq 4n^{md}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{ij} e_i = \max_{1 \leq j \leq 2n^{md}} Z_n(v_j; \{c_i = e_i\}), \quad e_i \sim \mathcal{N}(0,1).$$

When there is no ambiguity, we will adopt the notation $Z_n(t) := Z_n(t; \{c_i = 1\})$. We would like to compare T_0 and W_0 to

$$T := \sup_{t \in [0,1]^d} Z_n(t), \quad W := \sup_{t \in [0,1]^d} Z_n(t; \{e_i\}).$$

More precisely, as required in Chernozhukov et al. (2013), we will show that there exists $\zeta_1, \zeta_2 \geq 0$ both depending on n such that $\zeta_1 \sqrt{\log 4n^{md}} + \zeta_2 \leq C_2 n^{-c_2} \to 0$ as $n \to \infty$ for some constants $C_2, c_2 > 0$, and that

$$\mathbb{P}\left[|T - T_0| > \zeta_1\right] < \zeta_2, \quad \mathbb{P}\left[\mathbb{P}_e\left[|W - W_0| > \zeta_1\right] > \zeta_2\right] < \zeta_2. \tag{7}$$

For any fixed drawn $\{e_i\}_{i=1,\cdots,n}$, the supremum of $Z_n(t; \{e_i\})$ can be reached at some unsorted point t^* of the form

$$t^* = \left(t_{i_1}^{a_1}, \cdots, t_{i_d}^{a_d}\right)$$

for some $a_1, \cdots, a_d \in \{1, \cdots, d\}$ and $i_1, \cdots, i_d \in \{1, \cdots, n\}$. Suppose that $t^* \in B_{j^*} = \bigcap_{a=1}^{d} B_{i_a^*}^a$ for some $j^* \in \{1, \cdots, n^{md}\}$. If $\operatorname{sort} v_{j^*} =: \pi^*(v_{j^*}) \neq v_{j^*}$ i.e. t^* is not too close to the diagonal, and $\bigcup_{a=1}^{d} B_{i_a^*}^a \cup \bigcup_{a=1}^{d} B_{i_a^*}^{\pi^*(a)}$ contains no other sampled points apart from $t_{i_1}, \cdots, t_{i_d}$, then

$$Z_n(x_{j^*}; \{e_i\}) = Z_n(t^*; \{e_i\}) = \sup_{t \in [0,1]^d} Z_n(t; \{e_i\}).$$

A sufficient condition for the above requirements to be satisfied is that each interval $[z_{k-1}, z_k]$, $k = 1, \cdots, n^m$ contains at most $d-1$ coordinates of at most one sampled vectors $\{t_{i'}\}_{i'=1,\cdots,n}$, i.e. $\forall i = 1, \cdots, n^m$:

$$\begin{aligned} d - 1 \geq \left|[z_{i-1}, z_i] \cap \{t_{i'}^1, \cdots, t_{i'}^d\}\right| \geq 0, \quad \forall i' = 1, \cdots, n \\ \left|[z_{i-1}, z_i] \cap \{t_{i'}^1, \cdots, t_{i'}^d\}\right| > 0, \quad \text{for at most one } i' = 1, \cdots, n. \end{aligned}$$

If this is satisfied, we say that there is no *coordinate collision*. We compute the upper bound for the collision probability as follows. Let $I \subset \{1, \cdots, n^m\}$ be any fixed subset of size $(n-1)d$. Then,

$$\begin{aligned} \mathbb{P}\left[\text{Coordinate collision}\right] &\leq \mathbb{P}\left[\bigcup_{i=1}^{n} \left\{ t_i \in \bigcup_{j \in \{1, \cdots, n^m\}} B_{j,\cdots,j} \cup \bigcup_{(j,a) \in I \times \{1,\cdots,d\}} B_j^a \right\}\right] \\ &\leq n \cdot n^m \cdot \frac{1}{n^{(m-1)d}} + n \cdot d \cdot n(d-1) \cdot \frac{1}{n^m} \sim \frac{1}{n^{md-d-m-1}} + \frac{d^2}{n^{m-2}} \to 0, \end{aligned}$$

as $n \to \infty$, since $md - d - m - 1 > 0$ for all $m \geq 4, d \geq 2$ and $d^2/n^{m-2} \lesssim 1/n^{m-2-2c_0} \to 0$ given that $d = o(n^{c_0})$ for a $c_0 \in (0, 1/7)$. In the second inequality, the first term came from counting the number of diagonal boxes, each of them has probability bounded above by $\frac{1}{n^{(m-1)d}}$ by our construction. The second term counts the upper bound for the probability that at least one of the coordinates of any t_i shares the interval with one of the coordinates of one of the other $n-1$ drawn vectors. Since each of the $n(d-1)$ slabs contains an equal probability of $\frac{1}{n^m}$, the exact choice of I does not matter, as long as $|I| = (n-1)d$. It follows that

$$\mathbb{P}\left[|T - T_0| > 0\right] < \frac{1}{n^{md-d-m-1}} + \frac{d^2}{n^{m-2}}$$

and

$$\mathbb{P}\left[\mathbb{P}_e\left[|W-W_0|>0\right]>\frac{1}{n^{md-d-m-1}}+\frac{d^2}{n^{m-2}}\right]\leq\mathbb{P}\left[\mathbb{P}_e\left[|W-W_0|>0\right]>0\right]<\frac{1}{n^{md-d-m-1}}+\frac{d^2}{n^{m-2}}.$$

Therefore, (7) holds with $\zeta_1:=0,\zeta_2:=\frac{1}{n^{md-d-m-1}}+\frac{d^2}{n^{m-2}}$.

In the language of Chernozhukov et al. (2013), we also have for $j=1,\cdots,4n^{mp}$:

$$\begin{aligned}\bar{E}[x_{ij}^2]:=\frac{1}{n}\sum_{i=1}^{n}E[x_{ij}^2]&\geq\frac{1}{n}\left\lfloor\frac{n}{2}\right\rfloor\int_{[0,1]^d}\left([t\leq\operatorname{sort}\widetilde{v}_{j'}]-[t\leq\widetilde{v}_{j'}]\right)^2dF(t)\\&\quad+\frac{1}{n}\left\lfloor\frac{n}{2}\right\rfloor\int_{[0,1]^d}\left([t\leq\operatorname{sort}v_{i_1,\cdots,i_d}]-[t\leq v_{i_1,\cdots,i_d}]\right)^2dF(t)\\&\geq\frac{1}{3}\inf_{v\in A}\int_{[0,1]^d}\left([t\leq\operatorname{sort}v]-[t\leq v]\right)^2dF(t)=:c_1>0\end{aligned}$$

where the first inequality holds for some $j'\in\{1,\cdots,n^{md}\}$ and $(i_1,\cdots,i_d)\in\{1,\cdots,n^m\}^d$ depending on j. On the other hand, it is clear that $\bar{E}[x_{ij}^2]\leq C_1:=1$. We can also choose $B_n=1$ to satisfy the condition:

$$\max_{k=1,2}\bar{E}\left[|x_{ij}|^{2+k}/B_n^k\right]+E[\exp\left(|x_{ij}|/B_n\right)]\leq 1+e\leq 4.$$

It follows from Corollary 3.1 of Chernozhukov et al. (2013) and condition E.1. with

$$\frac{B_n^2(\log(2n^{md}\cdot n))^7}{n}\leq\frac{\left[(md+1)\log n+\log 2\right]^7}{n}\leq C_2n^{-c_2}\tag{8}$$

for some $C_2,c_2>0$, since we have assumed $d=o(n^{c_0})$ for a $c_0\in(0,1/7)$, that there exists $c>0,C>0$ depending only on C_1,c_1,C_2,c_2 such that

$$\sup_{\alpha\in(0,1)}\left|\mathbb{P}\left[T>c_W(\alpha)\right]-\alpha\right|\leq Cn^{-c}.$$

Unpacking the definition of T, we find that this is the statement of the theorem. □

A.2 Proof of Lemma 3.3

Proof: The expected value of a standard kernel estimator is computed as

$$\mathbb{E}[\hat{f}(t)]=f(t)+\frac{h^2}{2}tr\left(\frac{\partial^2f(t)}{\partial t\partial t^T}\int vv^TK(v)dv\right)+\text{higher order terms.}$$

For a permutation σ, we have that

$$\mathbb{E}[\hat{f}(\sigma(t))]=f(\sigma(t))+\frac{h^2}{2}tr\left(\frac{\partial^2f(\sigma(t)}{\partial t\partial t^T}\int vv^TK(v)dv\right)\text{higher order terms.}$$

Because f is permutation invariant, $f(\sigma(t))=f(t)$ and $\frac{\partial^2f(\sigma(t)}{\partial t\partial t^T}=\frac{\partial^2f(t)}{\partial t\partial t^T}$. □

A.3 Proof of Lemma 3.4

Proof: The first variance is a known result for multivariate kernel density estimators. To obtain the second variance, we have

$$\begin{aligned}\mathbb{E}(\tilde{f}(t)^2)&=\frac{1}{(\bar{d})^2}\mathbb{E}\Big(\sum_{\sigma\in S_d^*}\hat{f}(\sigma(t))\Big)^2\\&=\frac{1}{(\bar{d})^2}\mathbb{E}\Bigg(\sum_{\sigma\in S_d^*}\hat{f}(\sigma(t))^2+\sum_{\sigma\neq\pi}\hat{f}(\sigma(t))\hat{f}(\pi(t))\Bigg)\\&=\frac{1}{(\bar{d})^2}\sum_{\sigma\in S_d^*}\mathbb{E}\Big(\hat{f}(\sigma(t))^2\Big)+\frac{1}{(\bar{d})^2}\sum_{\sigma\neq\pi}\mathbb{E}\Big(\hat{f}(\sigma(t))\hat{f}(\pi(t))\Big).\end{aligned}$$

In terms of the first term, we have

$$\mathbb{E}\left(\hat{f}(\sigma(t))^2\right) = \frac{1}{nh^d} f(\sigma(t)) \int K(v)^2 dv + \text{higher order terms.}$$

Since there are $\bar{d}$ permutations in S_d^*, we have

$$\mathbb{E}[\tilde{f}(t)^2] = \frac{1}{\bar{d}} \frac{1}{nh^d} f(t) \int K(v)^2 dv + \frac{1}{(\bar{d})^2} \sum_{\sigma \neq \pi} \mathbb{E}\Big(\hat{f}(\sigma(t))\hat{f}(\pi(t))\Big) + \text{higher order terms.}$$

We now turn to the cross-product terms.

$$\begin{aligned}
\mathbb{E}\Big(\hat{f}(\sigma(t))\hat{f}(\pi(t))\Big) &= \mathbb{E}\Big(\frac{1}{nh^d}\Big(\sum_{i=1}^n K\Big(\frac{\sigma(t)-t^i}{h}\Big)\Big)\frac{1}{nh^d}\Big(\sum_{i=1}^n K\Big(\frac{\pi(t)-t^i}{h}\Big)\Big)\Big) \\
&= \frac{1}{n^2h^{2d}}\mathbb{E}\left(\sum_{i=1}^n K\Big(\frac{\sigma(t)-t^i}{h}\Big)K\Big(\frac{\pi(t)-t^i}{h}\Big) + \sum_{i\neq j} K\Big(\frac{\sigma(t)-t^i}{h}\Big)K\Big(\frac{\pi(t)-t^j}{h}\Big)\right) \\
&= \frac{1}{n^2h^{2d}}\sum_{i=1}^n \mathbb{E}\left(K\Big(\frac{\sigma(t)-t^i}{h}\Big)K\Big(\frac{\pi(t)-t^i}{h}\Big)\right) + \frac{1}{n^2h^{2d}}\sum_{i\neq j} \mathbb{E}\left(K\Big(\frac{\sigma(t)-t^i}{h}\Big)K\Big(\frac{\pi(t)-t^j}{h}\Big)\right) \\
&= \frac{1}{nh^{2d}}\underbrace{\mathbb{E}\left(K\Big(\frac{\sigma(t)-t^i}{h}\Big)K\Big(\frac{\pi(t)-t^i}{h}\Big)\right)}_{A} + \frac{n-1}{nh^{2d}}\underbrace{\mathbb{E}\left(K\Big(\frac{\sigma(t)-t^i}{h}\Big)K\Big(\frac{\pi(t)-t^j}{h}\Big)\right)}_{B}.
\end{aligned}$$

We start with the second term, B:

$$\begin{aligned}
\mathbb{E}\left(K\Big(\frac{\sigma(t)-t^i}{h}\Big)K\Big(\frac{\pi(t)-t^j}{h}\Big)\right) &= \int K\Big(\frac{\sigma(t)-t^i}{h}\Big)K\Big(\frac{\pi(t)-t^j}{h}\Big) f(t^i)f(t^j)dt^i dt^j \\
&= \left(\int K\Big(\frac{\sigma(t)-t^i}{h}\Big) f(t^i)dt^i\right)\left(\int K\Big(\frac{\pi(t)-t^i}{h}\Big) f(t^i)dt^i\right).
\end{aligned}$$

These are the same integrals except for the difference of σ and π. Without loss of generality, we work with the first one:

$$\begin{aligned}
\int K\Big(\frac{\sigma(t)-t^i}{h}\Big) f(t^i)dt^i &= h^d \int K(u) f(hu+\sigma(t))du \\
&\approx h^d \int \Big[K(u)f(\sigma(t)) + h\underbrace{K(u)u^T}_{\text{Integrates to 0}}\nabla f(\sigma(t))\Big]du \\
&= h^d f(\sigma(t)).
\end{aligned}$$

There are $n^2 - n$ terms where $i \neq j$ which means that

$$\frac{1}{n^2h^{2d}}\mathbb{E}\left(\sum_{i\neq j} K\Big(\frac{\sigma(t)-t^i}{h}\Big)K\Big(\frac{\pi(t)-t^j}{h}\Big)\right) \approx \frac{n-1}{n} f(\sigma(t))f(\pi(t)) = f(t)^2 - \frac{1}{n}f(t)^2.$$

We now turn to the first term, A:

$$\mathbb{E}\left(K\Big(\frac{\sigma(t)-t^i}{h}\Big)K\Big(\frac{\pi(t)-t^i}{h}\Big)\right) = \int K\Big(\frac{\sigma(t)-\pi(t)}{h} + \frac{\pi(t)-t^i}{h}\Big)K\Big(\frac{\pi(t)-t^i}{h}\Big) f(t^i)dt^i.$$

Letting $u = \frac{t^i - \pi(t)}{h}$, we have

$$\begin{aligned}
\mathbb{E}\Big(K\Big(\frac{\sigma(t)-t^i}{h}\Big)K\Big(\frac{\pi(t)-t^i}{h}\Big)\Big) &= h^d \int K\Big(\frac{\sigma(t)-\pi(t)}{h} - u\Big)K\Big(u\Big)f(hu+\pi(t))du \\
&\approx h^d \int K\Big(\frac{\sigma(t)-\pi(t)}{h} - u\Big)K\Big(u\Big)f(\pi(t))du \\
&= h^d f(\pi(t)) \int K\Big(\frac{\sigma(t)-\pi(t)}{h} - u\Big)K\Big(u\Big)du \\
&= h^d f(t)(K*K)\Big(\frac{\sigma(t)-\pi(t)}{h}\Big).
\end{aligned}$$

Putting everything together yields the second variance.

For the results in part (ii), note that we have

$$K*K\Big(\frac{\sigma(t)-\pi(t)}{h}\Big) = \prod_{j=1}^{d}(k*k)\Big(\frac{\sigma_j(t)-\pi_j(t)}{h}\Big),$$

where $\sigma_j(t)$ denotes the jth coordinate of $\sigma(t)$. The following observations can be made about the term above:

1. given the conditions on $k(\cdot)$ in part (ii) of Lemma 3.4, $(k*k)\Big(\frac{\sigma_j(t)-\pi_j(t)}{h}\Big)$ decreases in $|\frac{\sigma_j(t)-\pi_j(t)}{h}|$;
2. given the fact above and that $\sigma, \pi \in S_d^*$, by the construction of S_d^*, we have

$$\prod_{j=1}^{d} k*k\Big(\frac{\sigma_j(t)-\pi_j(t)}{h}\Big) \leq b^{\bar{d}} \int K(v)^2 dv$$

where $b \in (0,1)$.

□

A.4 Proof of Theorem 4.2

The following technical result will be needed.

Lemma A.1 *Let $\mathcal{P}_d^b = \big\{x \in [0,b]^d : 0 \leq x_1 \leq x_2 \leq ... \leq x_d \leq b\big\}$. Then the volume of $\mathcal{P}_d^b$, Vol$(\mathcal{P}_d^b) = \frac{b^d}{d!}$.*

Proof: We show this by induction.

Base case: If $d = 1$, $\mathcal{P}_d^b = [0,b]$. Then, Vol$(\mathcal{P}_1^b) = \frac{b}{1!}$.

Inductive step: Suppose that Vol$(\mathcal{P}_d^b) = \frac{b^d}{d!}$. Then,

$$\text{Vol}(\mathcal{P}_{d+1}^b) = \int_{x \in \mathcal{P}_{d+1}^b} dx = \int_0^b dx_{d+1} \underbrace{\int_0^{x_{d+1}} dx_d \cdots \int_0^{x_2} dx_1}_{\text{Vol}(\mathcal{P}_d^{x_{d+1}})} = \int_0^b \frac{x_{d+1}^d}{d!} dx_{d+1} = \frac{b^{d+1}}{(d+1)!}.$$

□

Proof: (of Theorem 4.2) We use the argument in Kolmogorov and Tikhomirov (1959) and Lemma A.1. When permutation invariance is absent, to derive an upper bound on $\log N_\infty(\delta, \mathcal{U})$, we consider a $b_{d,\gamma}^{-1}\delta^{\frac{1}{\gamma+1}}-$grid of points (where $b_{d,\gamma}$ is a function of (d,γ) only and independent of δ) in each dimension of $[0,1]^d$:

$$x_{0,j} = 0 < x_{1,j} < \cdots < x_{s-1,j} < x_{s,j}, \quad j \in \{1, ..., d\}$$

with $s \lesssim b_{d,\gamma}\delta^{\frac{-1}{\gamma+1}}$, and show that bounding $N_\infty(\delta, \mathcal{U})$ can be reduced to bounding the cardinality of

$$\Lambda = \left\{ \left(\left\lfloor \frac{D^p f(x_{i_1,1}, ..., x_{i_d,d})}{\delta_k} \right\rfloor, 0 \le i_1, ..., i_d \le s,\ 0 \le k \le \gamma \right) : f \in \mathcal{U} \right\}$$

with $\lfloor x \rfloor$ denoting the largest integer smaller than or equal to x. Then, using the fact that $D^p f(0) = 0$ for all p with $P = k$ $(k = 0, ..., \gamma)$, the argument in Kolmogorov and Tikhomirov (1959) implies that $|\Lambda| \le c^{s^d}$, where $c \in (0, \infty)$ is a constant independent of δ and (d, γ). Now with permutation invariance, by Lemma A.1, the number of points we need to consider scales as $\frac{1}{d!}s^d$. This fact is also applied along with the construction of the class of functions in Kolmogorov and Tikhomirov (1959) and the relationship between covering numbers and packing numbers to yield the lower bound. In addition, $\log N_2(\delta, \mathcal{U}) \lesssim \log N_\infty(\delta, \mathcal{U})$. Standard argument in the literature using the Vasharmov-Gilbert Lemma and the relationship between covering numbers and packing numbers further give $\log N_2(\delta, \mathcal{U}) \gtrsim \log N_\infty(\delta, \mathcal{U})$. In sum, $\log N_2(\delta, \mathcal{U}) \asymp \log N_\infty(\delta, \mathcal{U}) \asymp b_{d,\gamma}^d \delta^{\frac{-d}{\gamma+1}}$ and $\log N_2(\delta, \mathcal{U}^{perm}) \asymp \log N_\infty(\delta, \mathcal{U}^{perm}) \asymp \frac{1}{d!} b_{d,\gamma}^d \delta^{\frac{-d}{\gamma+1}}$. □

A.5 Proof of Theorem 4.4

Proof: From definition, we have $\mu_k \to 0$ as $\max_{a=1,\cdots,d} k^a \to \infty$, therefore we can find $\mathbb{K}_\delta := \left\{ k \in \mathbb{Z}^d_{\ge 0} \mid \max_{a=1,\cdots,d} k^a \le \bar{k} \right\}$ for some $\bar{k} > 0$ such that $\sum_{k \in \mathbb{Z}^d_{\ge 0} \setminus \mathbb{K}_\delta} \beta_k^2 \le \delta^2$ for all $\beta \in \mathcal{E}$. Then any δ–cover $\{\beta^1, \cdots, \beta^N\}$ of the $D_\delta := |\mathbb{K}_\delta|$–dimensional truncated ellipsoid $\widetilde{\mathcal{E}} := \{\beta \in \mathcal{E} \mid \beta_k = 0, \forall k \notin \mathbb{K}_\delta\}$ is a $\sqrt{2}\delta$–cover of $\mathcal{E}$, since

$$\min_{j=1,\cdots,N} \|\beta - \beta^j\|_{l^2}^2 = \min_{j=1,\cdots,N} \sum_{k \in \mathbb{K}_\delta} (\beta_k - \beta_k^j)^2 + \sum_{k \in \mathbb{Z}^d_{\ge 0} \setminus \mathbb{K}_\delta} (\beta_k)^2 \le 2\delta^2.$$

for any $\beta \in \mathcal{E}$. It follows from Lemma 5.7 of Wainwright (2019) that

$$\left(\frac{\sqrt{2}}{\delta}\right)^{D_\delta} \frac{vol\left(\widetilde{\mathcal{E}}\right)}{vol\left(\mathbb{B}_2^{D_\delta}(1)\right)} \le N(\delta, \mathcal{E}, \|.\|_{l_2}) \le \left(\frac{2\sqrt{2}}{\delta}\right)^{D_\delta} \frac{vol\left(\widetilde{\mathcal{E}} + \mathbb{B}_2^{D_\delta}(\delta/2)\right)}{vol\left(\mathbb{B}_2^{D_\delta}(1)\right)}.$$

Let $\underline{\mu}_\delta := \min_{k \in \mathbb{K}_\delta} \mu_k$ and $\bar{\mu}_\delta := \max_{k \in \mathbb{K}_\delta} \mu_k$, then it follows from $\mathbb{B}_2^{D_\delta}(\sqrt{\underline{\mu}_\delta}) \subset \widetilde{\mathcal{E}} \subset \mathbb{B}_2^{D_\delta}(\sqrt{\bar{\mu}_\delta})$ that:

$$\left(\frac{\sqrt{2\underline{\mu}_\delta}}{\delta}\right)^{D_\delta} \le N(\delta, \mathcal{E}, \|.\|_{l_2}) \le \left(\frac{2\sqrt{2\bar{\mu}_\delta}}{\delta} + 1\right)^{D_\delta}.$$

On the other hand, let $\mathbb{K}_\delta$ be as chosen previously, and let $\mathbb{K}_\delta^{perm} := \operatorname{sort} \mathbb{K}_\delta$ then $D_\delta^{perm} := |\mathbb{K}_\delta^{perm}| \asymp \frac{D_\delta}{d!}$ by the construction that $\mathbb{K}_\delta$ contains the entire S_d–orbit of all its elements. Then $\sum_{k \in \operatorname{sort} \mathbb{Z}^d_{\ge 0} \setminus \mathbb{K}_\delta^{perm}} \beta_k^2 \le \sum_{k \in \mathbb{Z}^d_{\ge 0} \setminus \mathbb{K}_\delta} \beta_k^2 \le \delta^2$, and therefore any δ–cover $\{\beta'^1, \cdots, \beta'^{N'}\}$ of the D_δ^{perm}–dimensional truncated permutation invariant ellipsoid $\widetilde{\mathcal{E}}^{perm} := \{\beta \in \mathcal{E}^{perm} \mid \beta_k = 0, \forall k \notin \mathbb{K}_\delta^{perm}\}$ is a $\sqrt{2}\delta$–cover of $\mathcal{E}^{perm}$, since

$$\min_{j=1,\cdots,N'} \|\beta - \beta'^j\|_{l^2}^{perm2} = \min_{j=1,\cdots,N'} \sum_{k \in \mathbb{K}_\delta^{perm}} (\beta_k - \beta_k'^j)^2 + \sum_{k \in \operatorname{sort} \mathbb{Z}^d_{\ge 0} \setminus \mathbb{K}_\delta^{perm}} (\beta_k)^2 \le 2\delta^2$$

for any $\beta \in \mathcal{E}^{perm}$. Following the same analysis as before, we obtain

$$\left(\frac{\sqrt{2\underline{\mu}_\delta}}{\delta}\right)^{D_\delta^{perm}} \le N(\delta, \mathcal{E}^{perm}, \|.\|_{l_2}^{perm}) \le \left(\frac{2\sqrt{2\bar{\mu}_\delta}}{\delta} + 1\right)^{D_\delta^{perm}}.$$

The result follows by identifying $\underline{g}(\delta) := D_\delta \log\left(\frac{\sqrt{2\underline{\mu}_\delta}}{\delta}\right)$ and $\bar{g}(\delta) := D_\delta \log\left(\frac{2\sqrt{2\bar{\mu}_\delta}}{\delta} + 1\right)$. □

B MULTIPLIER BOOTSTRAP TEST WITH MMD METRIC

In the following, we let $\mathcal{X} := [0,1]^d$. Given an RKHS $\mathcal{H}$ with a *characteristic* kernel $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we can define the maximum mean discrepancy (MMD) metric on the space of probability distribution $\mathcal{P}(\mathcal{X})$ as follows. Given $\mu, \nu \in \mathcal{P}(\mathcal{X})$, we denote by $\mu_{\mathcal{H}}, \nu_{\mathcal{H}}$ the element in $\mathcal{H}$ (unique due to the characteristic $\mathcal{K}$) such that $\langle f, \mu_{\mathcal{H}} \rangle = \mathbb{E}_{X\sim\mu}[f(X)]$ for any $f \in \mathcal{H}$, and similarly for $\nu_{\mathcal{H}}$. We define $MMD(\mu,\nu) := \|\mu_{\mathcal{H}} - \nu_{\mathcal{H}}\|_{\mathcal{H}}$.

From Theorem 5 of Soleymani et al. (2025) we have that μ is permutation invariant iff $\max_{g\in S} \|(g\mu)_{\mathcal{H}} - \mu_{\mathcal{H}}\|_{\mathcal{H}}^2 = 0$, where $S \subset S_d$ is a generating set of the permutation group. From Proposition 2 of Soleymani et al. (2025) (or Gretton et al. (2012), Lemma 6), we have the identity:

$$\|(g\mu)_{\mathcal{H}} - \mu_{\mathcal{H}}\|_{\mathcal{H}}^2 = 2\mathbb{E}_{(X,X')\sim\mu\times\mu}[\mathcal{K}(X,X')] - 2\mathbb{E}_{(X,X')\sim\mu\times\mu}[\mathcal{K}(X,gX')].$$

In particular,

$$2\mathbb{E}_{(X,X')\sim\mu\times\mu}[\mathcal{K}(X,X')] - 2\mathbb{E}_{(X,X')\sim\mu\times\mu}[\mathcal{K}(X,gX')] = 0 \tag{9}$$

for all $g \in S$ if μ is permutation invariant. We can take $S = \{(1,2),(1,3),\cdots,(1,d)\}$, so $|S| = d-1$. Also, let us use the notation $g_k := (1, k+1)$, $k = 1,\cdots,d-1$ to represent the permutation which switches the 1 and $k+1$ coordinates while leaving the rest unchanged. Let us consider the following statistics

$$T := \max_{k=1,\cdots,d-1} \frac{2}{\sqrt{n}} \sum_{i=1}^{n} \left(\mathcal{K}(t_i, t_{i+n}) - \mathcal{K}(t_i, g_k t_{i+n})\right) \tag{10}$$

and the multiplier bootstrap version

$$W := \max_{k=1,\cdots,d-1} \frac{2}{\sqrt{n}} \sum_{i=1}^{n} \left(\mathcal{K}(t_i, t_{i+1}) - \mathcal{K}(t_i, g_k t_{i+n})\right) e_i, \quad e_i \sim_{iid} \mathcal{N}(0,1) \tag{11}$$

along with the corresponding bootstrap critical value $c_W(\alpha) := \inf\{t \in \mathbb{R} : \mathbb{P}_e[W \le t] \ge 1-\alpha\}$. The following result shows that the alternative test above attains the pre-specified significance level asymptotically.

Theorem B.1 *Suppose that $d = o\left(e^{n^{c_0}}\right)$ for some $c_0 \in (0, 1/7)$. Under H_0: μ is permutation invariant, there exists some universal constants $c, C \in (0,\infty)$ such that*

$$\sup_{\alpha\in(0,1)} \left| \mathbb{P}\left[\max_{k=1,\cdots,d-1} \frac{2}{\sqrt{n}} \sum_{i=1}^{n} \left(\mathcal{K}(t_i, t_{i+n}) - \mathcal{K}(t_i, g_k t_{i+n})\right) > c_W(\alpha) \right] - \alpha \right| < Cn^{-c} \to 0 \quad as \quad n \to \infty.$$

Proof: Consider i.i.d. samples $\{t_i\}_{i=1,\cdots,2n}$ drawn from the distribution μ. Then $\{(t_i, t_{i+n})\}_{i=1,\cdots,n}$ are i.i.d. samples on $\mathcal{X}\times\mathcal{X}$ drawn from the distribution $\mu\times\mu$. Define

$$x_{ik} := 2\mathcal{K}(t_i, t_{i+n}) - 2\mathcal{K}(t_i, g_k t_{i+n}).$$

Then, $x_i := (x_{i1}, \cdots, x_{i,d-1}) \in \mathbb{R}^{d-1}$ for $i = 1,\cdots,n$ are independent random vectors, and $\mathbb{E}[x_{ik}] = 0$ from (9). Moreover, we have

$$\begin{aligned}
\mathbb{E}[x_{ik}^2] &= 4\mathbb{E}_{(X,X')\sim\mu\times\mu}\left[\left(\mathcal{K}(X,X') - \mathcal{K}(X,g_kX')\right)^2\right] \\
&= 4\mathbb{E}_{(X,X')\sim\mu\times\mu}\left[\mathcal{K}(X,X')^2\right] + 4\mathbb{E}_{(X,X')\sim\mu\times\mu}\left[\mathcal{K}(X,g_kX')^2\right] - 8\mathbb{E}_{(X,X')\sim\mu\times\mu}\left[\mathcal{K}(X,X')\mathcal{K}(X,g_kX')\right] \\
&= 8\mathbb{E}_{(X,X')\sim\mu\times\mu}\left[\mathcal{K}(X,X')^2\right] - 8\mathbb{E}_{(X,X')\sim\mu\times\mu}\left[\mathcal{K}(X,X')\mathcal{K}(X,g_kX')\right] > 0
\end{aligned}$$

for all k. We let $c_1 := \min_{k=1,\cdots,d-1} \mathbb{E}[x_{ik}^2]$. On the other hand, it is clear that $\mathbb{E}[x_{ik}^2] \le C_1 := \max_{k=1,\cdots,d-1} \mathbb{E}[x_{ik}^2]$. We can also choose a sufficiently large constant B_n to satisfy the condition

$$\max_{k'=1,2} \bar{E}\left[|x_{ik}|^{2+k'}/B_n^{k'}\right] + E[\exp\left(|x_{ik}|/B_n\right)] \le 4.$$

Define

$$T_0 := \max_{k=1,\cdots,d-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{ik}$$

$$W_0 := \max_{k=1,\cdots,d-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{ik} e_i, \qquad e_i \sim \mathcal{N}(0,1).$$

Let $T := T_0, W := W_0$. It follows from (Chernozhukov et al., 2013, Corollary 3.1) and condition E.1. with

$$\frac{B_n^2 \left[\log(nd-n)\right]^7}{n} \lesssim n^{7c_0-1} \leq C_2 n^{-c_2},$$

since $d = o\left(e^{n^{c_0}}\right)$ for a $c_0 \in (0, 1/7)$, for some C_2 and $c_2 \in (0, 1-7c_0)$, that there exists $c, C > 0$ depending only on C_1, c_1, C_2, c_2 such that

$$\sup_{\alpha \in (0,1)} \left|\mathbb{P}\left[T > c_W(\alpha)\right] - \alpha\right| \leq C n^{-c}.$$

Unpacking the definition of T, we find that this is the statement of the theorem. □

C PERIODIC SOBOLEV REPRODUCING KERNEL

Here, we derive the reproducing kernel of the space of periodic Sobolev functions $W^{s,2}_{per}([0,1]^d)$, i.e. the subspace of the Sobolev space $W^{s,2}([0,1]^d)$ for functions f such that $f(t+k) = f(t)$ for any $k \in \mathbb{Z}^d$. We only consider the case when $s > d/2$ so that we have an embedding $W^{s,2}([0,1]^d) \subset C^{0,s-d/2}([0,1]^d)$ to guarantee that $W^{s,2}_{per}([0,1]^d)$ is an RKHS. For reasons that are not obvious to us, results on reproducing kernels for *multivariate* Sobolev space appear sparse in the literature despite its importance. One result we found is for the reproducing kernel of $W^{s,2}(\mathbb{R}^d)$ and derived in Novak et al. (2018). Our derivation supplements their result for the compact domain case. The periodicity is a natural condition that is needed to make sense of the specified smoothness at the boundary of $[0,1]^d$. Equivalently, we may consider Sobolev functions on a d–dimensional torus $\mathbb{T}^d$.

Lemma C.1 *The reproducing kernel $\mathcal{K} : [0,1]^d \times [0,1]^d \to \mathbb{R}$ for the periodic Sobolev space $W^{s,2}_{per}([0,1]^d)$ with $s > d/2$ is*

$$\mathcal{K}_{d,s}(t,t') = \sum_{k \in \mathbb{Z}^d_{\geq 0}} \frac{2\cos 2\pi k \cdot (t-t')}{v_{d,s}[k]^2} - 1,$$

where $v_{d,s}[k] := \left[\sum_{|\alpha| \leq s} \prod_{j=1}^{d} (2\pi k_j)^{2\alpha_j}\right]^{1/2}$ and the corresponding eigenvalues $\{\mu_k = 1/v_{d,s}[k]^2\}_{k \in \mathbb{Z}^d}$.

By restricting to real–valued functions, the eigenvectors of the corresponding Hilbert-Schmidt operator are $\{e_k^+ := t \mapsto \cos 2\pi k \cdot t, e_k^- := t \mapsto \sin 2\pi k \cdot t\}_{k \in \mathbb{Z}^d_{\geq 0}}$ with eigenvalues $\{\mu_k = 1/v_{d,s}[k]^2\}_{k \in \mathbb{Z}^d_{\geq 0}}$, and the RKHS can be written as an ellipsoid:

$$W^{s,2}_{per}([0,1]^d) \cong \left\{ (\beta_k)_{k \in \mathbb{Z}^d_{\geq 0}} \mid \sum_{k \in \mathbb{Z}^d_{\geq 0}} v_{d,s}[k]^2 \beta_k^2 < \infty \right\},$$

where $\langle f, g \rangle_{W^{s,2}_{per}} = \sum_{k \in \mathbb{Z}^d_{\geq 0}} v_{d,s}[k]^2 \beta_{f,k} \beta_{g,k}$. For the space of permutation invariant periodic Sobolev functions $W^{s,2}_{per,perm}([0,1]^d) \subset W^{s,2}_{per}([0,1]^d)$, we simply restrict the above ellipsoid to $\beta_{\text{sort } k} = \beta_k$.

Proof: For any $f \in W^{s,2}_{per}([0,1]^d) \subset L^2([0,1]^d)$, the corresponding Fourier series is given by

$$f(t) = \sum_{k_1,\cdots,k_d=-\infty}^{+\infty} \mathcal{F}f[k] e^{i2\pi k \cdot t}$$

where

$$\mathcal{F}f[k] := \int_{[0,1]^p} f(t) e^{-i2\pi k \cdot t} dt.$$

The inner products are given by:

$$\langle f, g\rangle_{L^2} := \int_{[0,1]^p} f(t)^* g(t) dt, \quad \langle f, g\rangle_{W^{s,2}} := \sum_{|\alpha|\le s} \langle D^\alpha f, D^\alpha g\rangle_{L^2}.$$

The periodic properties help us easily compute the Fourier coefficients of any derivatives:

$$\mathcal{F}(D^\alpha f)[k] = \prod_{j=1}^d (i2\pi k_j)^{\alpha_j} \mathcal{F}f[k].$$

For convenience, we define $v_{d,s}[k] := \left[\sum_{|\alpha|\le s} \prod_{j=1}^d (2\pi k_j)^{2\alpha_j}\right]^{1/2}$. From Parseval's Theorem we also know that

$$\langle f, g\rangle_{L^2} = \sum_{k_1,\cdots,k_d=0}^{\infty} (\mathcal{F}f[k])^* (\mathcal{F}g[k]) =: \langle \mathcal{F}f, \mathcal{F}g\rangle_{l^2}.$$

Therefore,

$$\begin{aligned}\langle f, g\rangle_{W^{s,2}_{per}} = \sum_{|\alpha|\le s} \langle D^\alpha f, D^\alpha g\rangle_{L^2} &= \sum_{|\alpha|\le s} \langle \mathcal{F}D^\alpha f, \mathcal{F}D^\alpha g\rangle_{l^2} \\ &= \sum_{|\alpha|\le s} \sum_{k_1,\cdots,k_d=-\infty}^{+\infty} \prod_{j=1}^p (2\pi k_j)^{2\alpha_j} \mathcal{F}f[k]^* \mathcal{F}g[k] = \langle v_{d,s}\mathcal{F}f, v_{d,s}\mathcal{F}g\rangle_{l^2}.\end{aligned}$$

Let $K_{p,s}$ be the reproducing kernel. We have $K_{p,s}(.,t) \in W^{s,2}_{per}([0,1]^d)$ and for any $f \in W^{s,2}_{per}([0,1]^d)$:

$$f(t) = \langle f, K_{d,s}(.,t)\rangle_{W^{s,2}_{per}} = \langle v_{d,s}\mathcal{F}f, v_{d,s}\mathcal{F}K_{d,s}(.,t)\rangle_{l^2}.$$

On the other hand,

$$\begin{aligned}f(t) = \mathcal{F}^{-1}\mathcal{F}f(t) &= \sum_{k_1,\cdots,k_p=-\infty}^{+\infty} \mathcal{F}f[k] e^{-i2\pi k\cdot t} \\ &= \sum_{k_1,\cdots,k_d=-\infty}^{+\infty} v_{d,s}[k]\mathcal{F}f[k]^* \cdot v_{d,s}[k]\frac{e^{-i2\pi k\cdot t}}{v_{d,s}[k]^2} = \langle v_{d,s}\mathcal{F}f, v_{d,s}\frac{e^{-i2\pi k\cdot t}}{v_{d,s}^2}\rangle_{l^2}.\end{aligned}$$

Comparing the two above, since f is arbitrary, we must have $\mathcal{F}K_{d,s}(.,t)[k] = \frac{e^{-i2\pi k\cdot t}}{v_{d,s}[k]^2}$, which means:

$$K_{d,s}(t,t') = \sum_{k_1,\cdots,k_d=-\infty}^{+\infty} \frac{e^{i2\pi k\cdot(t-t')}}{v_{d,s}[k]^2} = \sum_{k\in\mathbb{Z}^d_{\ge 0}} \frac{2\cos 2\pi k\cdot(t-t')}{v_{d,s}[k]^2} - 1.$$

The corresponding Hilbert-Schmidt operator is $T_K : L^2([0,1]^d) \to L^2([0,1]^d)$:

$$T_K[f](t) := \int_{[0,1]^d} K(t,t') f(t') dt'$$

with eigenvectors $\{e_k := t \mapsto e^{i2\pi k\cdot t}\}_{k\in\mathbb{Z}^d}$ with the corresponding eigenvalues $\{\mu_k := 1/v_{d,s}[k]^2\}_{k\in\mathbb{Z}^d}$. □