

ORIGINAL ARTICLE

Consistency of permutation tests of independence using distance covariance, HSIC and dHSIC

David Rindt  | Dino Sejdinovic | David Steinsaltz

Department of Statistics, University of Oxford,
24-29 St Giles', Oxford, OX1 3LB, UK

Correspondence

David Rindt, Department of Statistics,
University of Oxford, 24-29 St Giles', Oxford,
OX1 3LB, UK.
Email: david.rindt@stats.ox.ac.uk

Funding information

UK Engineering and Physical Sciences
Research Council (EPSRC)

The Hilbert–Schmidt independence criterion (HSIC) and its d -variable extension dHSIC are measures of (joint) dependence between random variables. While combining these statistics with a permutation test has become a popular method of testing the null hypothesis of (joint) independence, it had thus far not been proved that this results in a consistent test. In this work, we provide a simple proof that the permutation test with the test statistic HSIC or dHSIC is indeed consistent when using characteristic kernels. That is, we prove that under each alternative hypothesis, the power of these permutation tests indeed converges to 1 as the sample size converges to infinity. Since the test is consistent for each number of permutations, we further give a brief discussion of how the number of permutations relates to the power of the test and how the number of permutations may be selected in practice.

KEYWORDS

mathematical statistics, non-parametric methods, resampling methods, statistical computing, statistical inference

1 | INTRODUCTION

During the last decade, many new tests of independence of random variables have been developed. Among these are tests based on distance covariance (Szekély & Rizzo, 2009) and the Hilbert–Schmidt independence criterion (HSIC) (Gretton et al. 2008). In fact, under a particular choice of kernels, HSIC has been shown to be equivalent to distance covariance (Sejdinovic et al. 2012). In Pfister et al. (2018), a d -variable extension of HSIC, named dHSIC, was proposed, which can be used to test the hypothesis of joint independence of d random variables.

These statistics have several desirable properties. For appropriate choices of kernels, the population value of HSIC equals zero if and only if the two variables are independent (Gretton et al. 2008). Similarly, the population value of dHSIC is zero if and only if the variables are indeed jointly independent (Pfister et al. 2018). One thus does not need to make assumptions about the form of the relationship among the variables. Furthermore, under mild conditions, the test statistic converges in probability to the population value. Additionally, these tests may be applied to multidimensional random variables, and even to variables that do not take values in the Euclidean domains, such as graphs or text (Gretton et al. 2008).

In practice, one does not have access to the true null distribution of these statistics. As discussed in Gretton et al. (2008) and Pfister et al. (2018), the three main methods to approximate a null distribution are using the Gamma distribution, a bootstrap procedure, and a permutation test.

To our knowledge, Pfister et al. (2018) was the first work to study the consistency properties of these three kernel-based testing procedures. A test is called consistent if the rejection rate of the test converges to 1 as the sample size increases for each distribution for which the alternative hypothesis holds. That work proved in particular that the bootstrap procedure results in a consistent test whereas the Gamma approximation does not guarantee a consistent test. Consistency of the permutation procedure remained an open question and is the topic of this work.

A permutation test is a test in which the statistic on the original data is compared to the statistic on a number of randomly permuted datasets. We define this in Definition 6. An important property of such permutation tests is that they are well known to have a valid level for each sample size and for each number of permutations (see, e.g., Hoeffding, 1952).

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2021 The Authors. Stat published by John Wiley & Sons Ltd.

There are several relevant recent works on the power of permutation tests. The question if a permutation test with the statistic dHSIC or HSIC results in a consistent test was posed in Pfister et al. (2018): see in particular table 1 and section 3.2.1 of that work. In Kim et al. (2020), uniform power (rather than power for a fixed alternative) of permutation tests has been studied against classes of alternative hypotheses meeting certain conditions on the densities of the joint distribution. In Berrett et al. (2020), uniform power of permutation tests has also been studied under certain conditions on the density, although not in the specific context of HSIC. In contrast to these last two papers, we study the power for each fixed alternative hypothesis, and we do not make any assumptions on the alternative hypothesis.

The main contribution of this work is to provide a simple proof of the consistency of the permutation test for each fixed alternative hypothesis. We do not follow the proof strategy proposed in remark 2 of Pfister et al. (2018) but use much more elementary techniques that can be traced back at least to Hoeffding (1952): as we discuss in Section 2, the test statistic dHSIC, with an appropriate choice of kernel, converges to a positive constant for each fixed alternative hypothesis. The main observation from which consistency will follow in Section 3 is that the statistic on the randomly permuted datasets converges to zero in probability (Lemma 1).

While only elementary techniques are used in proving consistency, we do believe the proved result is of importance. The main motivation for using kernel methods for independence testing is the ability to detect every possible type of dependence, and Theorem 3 shows that this is indeed the case for the permutation test using dHSIC. The analogous question of consistency of the bootstrap procedure was already answered in Pfister et al. (2018), where it was also shown that the Gamma-approximation procedure did not have a consistency guarantee and the question of consistency of the permutation test was posed. This result on the permutation test thus aims to fill this gap in the literature.

Since the permutation test is both consistent and has correct type 1 error for each number of permutations, it is natural to ask how to select the number of permutations. One may ask, for example, how the number of permutations influences the power of the test. Section 4 briefly discusses that these questions provide some guidelines on how one may select the number of permutations in practice.

2 | BACKGROUND

2.1 | Reproducing kernel Hilbert spaces

This section reviews some relevant information about reproducing kernel Hilbert spaces (RKHSs).

Definition 1. (RKHS, Schölkopf & Smola, 2001) Let \mathcal{X} be a non-empty set and \mathcal{H} a Hilbert space of functions $f : \mathcal{X} \rightarrow \mathbb{R}$ endowed with dot product $\langle \cdot, \cdot \rangle$. Then \mathcal{H} is called an RKHS if there exists a function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ with the following properties.

1. k has the reproducing property

$$\langle f, k(x, \cdot) \rangle = f(x) \text{ for all } f \in \mathcal{H}, x \in \mathcal{X}.$$

2. k spans \mathcal{H} ; that is, $\mathcal{H} = \overline{\text{span} \{k(x, \cdot) | x \in \mathcal{X}\}}$ where the bar denotes the completion of the space.

Let \mathcal{X} be a measurable space and \mathcal{H}_k be an RKHS on \mathcal{X} with kernel k . Let \mathbb{P} be a probability measure on \mathcal{X} . If $\mathbb{E}_{\mathbb{P}} \sqrt{k(X, X)} < \infty$, then there exists an element $\mu_{\mathbb{P}} \in \mathcal{H}_k$ such that $\mathbb{E}_{\mathbb{P}} f(X) = \langle f, \mu_{\mathbb{P}} \rangle$ for all $f \in \mathcal{H}_k$ (Gretton et al. 2012), where we use the notation $\mathbb{E}_{\mathbb{P}} f(X) := \int f(x) \mathbb{P}(dx)$. The element $\mu_{\mathbb{P}}$ is called the mean embedding of \mathbb{P} in \mathcal{H}_k . Given a sample $\{x_i\}_{i=1}^n$ and the corresponding empirical distribution, $\sum_{i=1}^n \delta_{x_i}/n$, the corresponding mean embedding is given by $\sum_{i=1}^n k(x_i, \cdot)/n$. Given a second distribution \mathbb{Q} on \mathcal{X} , of which a mean embedding exists, we can measure the dissimilarity of \mathbb{P} and \mathbb{Q} by the distance between their mean embeddings in \mathcal{H}_k . That is,

$$\text{MMD}(\mathbb{P}, \mathbb{Q}) := \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k}.$$

This is also called the maximum mean discrepancy (MMD). The name comes from the following equality (Gretton et al. 2012),

$$\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k} = \sup_{f \in \mathcal{H}_k} |\mathbb{E}_{\mathbb{P}} f(X) - \mathbb{E}_{\mathbb{Q}} f(X)|,$$

showing that MMD is an integral probability metric. Lastly, we give the definitions of a characteristic kernel and a c_0 -universal kernel.

Definition 2. (Characteristic kernel, Sriperumbudur et al. 2011) The kernel k is said to be characteristic when $\text{MMD}(\mathbb{P}, \mathbb{Q}) = 0$ if and only if $\mathbb{P} = \mathbb{Q}$.

Definition 3. (c_0 -Universal kernel, Sriperumbudur et al. 2011) Let \mathcal{X} be a locally compact Hausdorff space and let $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a reproducing kernel on \mathcal{X} . Then k is said to be c_0 -universal if it is continuous and its associated RKHS \mathcal{H}_k is dense in $C_0(\mathcal{X})$, the set of continuous bounded functions, with respect to the infinity (also called uniform) norm.

If a kernel is c_0 -universal on a domain \mathcal{X} , then the kernel is also characteristic on that domain (Sriperumbudur et al. 2011). We now give some examples of c_0 -universal (and thus also characteristic) kernels on \mathbb{R}^d .

1. Gaussian kernel: $k_\sigma(x, y) = \exp(-||x - y||^2 / \sigma^2)$ with $\sigma > 0$ for $x, y \in \mathbb{R}^d$. This kernel is both characteristic and c_0 -universal.
2. Matérn kernel: $k_{\nu, \rho}(x, y) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{||x-y||}{\rho} \right)^\nu K_\nu \left(\sqrt{2\nu} \frac{||x-y||}{\rho} \right)$, where Γ is the gamma function, K_ν the modified Bessel function of the second kind, and $\rho, \nu \in \mathbb{R}_{>0}$. This kernel is both characteristic and c_0 -universal.
3. Rational quadratic kernel: $k_{\alpha, \ell}(x, y) = \left(1 + \frac{||x-y||^2}{2\alpha\ell} \right)^{-\alpha}$ with $\alpha, \ell \in \mathbb{R}_{>0}$. This kernel is both characteristic and c_0 -universal.
4. Inverse multiquadric kernel: $k_c(x, y) = 1/\sqrt{||x-y||^2 + c}$ for $c \in \mathbb{R}_{>0}$. This kernel is both characteristic and c_0 -universal.
5. Covariance kernel of Brownian motion: $k(x, y) = (||x|| + ||y|| + ||x - y||)/2$. This kernel is not c_0 -universal on all of \mathbb{R}^d . However, when we view this as a kernel on a compact domain, rather than on all of \mathbb{R}^d , then the kernel is characteristic and c_0 -universal.

2.2 | dHSIC and HSIC

In Pfister et al. (2018), a measure of joint independence is proposed. Consider the following setting.

Proof. For $j = 1, \dots, d$, let \mathcal{X}^j be a locally compact metric space equipped with the Borel sigma algebra and let the Cartesian product $\mathcal{X} = \mathcal{X}^1 \times \dots \times \mathcal{X}^d$ be equipped with the product sigma algebra. Let $X^j : \Omega \rightarrow \mathcal{X}^j$ be random variables on the shared probability space $(\Omega, \mathbb{P}, \mathcal{F})$. In this section, the superscript on X^j always indexes X and never denotes a power of the variable X . Let $k^j(\cdot, \cdot) : \mathcal{X}^j \times \mathcal{X}^j \rightarrow \mathbb{R}$ be kernels on \mathcal{X}^j . Finally, let $k := k^1 \otimes \dots \otimes k^d$ be the tensor product of the d kernels. We let $\mathcal{H}_k, \mathcal{H}_{k^j}$ be the corresponding RKHSs. Assume that the kernel k is characteristic and bounded. \square

In practice, one may construct a bounded characteristic product kernel k , by choosing each k^j for $j = 1, \dots, d$ to be c_0 -universal and bounded: in Szabó and Sriperumbudur (2017), it is found that the product of c_0 -universal kernels is c_0 -universal and thus also characteristic. Kernels 1–4 of the examples below Definition 3 are all c_0 -universal and bounded. Kernel 5 is c_0 -universal and bounded when we restrict it to be a kernel on a compact subset of \mathbb{R}^d .

By definition (X^1, \dots, X^d) are said to be jointly independent if $\mathbb{P}_{X^1, \dots, X^d} = \mathbb{P}_{X^1} \times \dots \times \mathbb{P}_{X^d}$. The main topic of this work is the test of the hypothesis $H_0 : \mathbb{P}_{X^1, \dots, X^d} = \mathbb{P}_{X^1} \times \dots \times \mathbb{P}_{X^d}$ versus $H_1 : \mathbb{P}_{X^1, \dots, X^d} \neq \mathbb{P}_{X^1} \times \dots \times \mathbb{P}_{X^d}$. We now define dHSIC:

Definition 4. (dHSIC, Pfister et al. 2018) Assume Setting 1. Then dHSIC is defined as

$$\text{dHSIC}(X^1, \dots, X^d) := ||\mu_{\mathbb{P}_{X^1, \dots, X^d}} - \mu_{\mathbb{P}_{X^1} \times \dots \times \mathbb{P}_{X^d}}||_{\mathcal{H}_k}^2,$$

where we recall that for any distribution \mathbb{P} , the element $\mu_{\mathbb{P}} \in \mathcal{H}_k$ denotes the mean embedding of \mathbb{P} .

Note that, because in Setting 1 we have assumed k to be characteristic, $\text{dHSIC}(X^1, \dots, X^d) = 0$ if and only if $\mathbb{P}_{X^1, \dots, X^d} = \mathbb{P}_{X^1} \times \dots \times \mathbb{P}_{X^d}$.

As we typically do not have access to the full distribution \mathbb{P}_X , but only to a sample $D := (x_i)_{i=1}^n \in \mathcal{X}^n$, we study the estimator proposed in Pfister et al. (2018)

$$\begin{aligned} \widehat{\text{dHSIC}}(x_1, \dots, x_n) &:= \frac{1}{n^2} \sum_{M_2(n)} \prod_{j=1}^d k^j(x_{i_1}^j, x_{i_2}^j) + \frac{1}{n^{2d}} \sum_{M_{2d}(n)} \prod_{j=1}^d k^j(x_{i_{2j-1}}^j, x_{i_{2j}}^j) \\ &\quad - \frac{2}{n^{d+1}} \sum_{M_{d+1}(n)} \prod_{j=1}^d k^j(x_{i_1}^j, x_{i_{j+1}}^j). \end{aligned}$$

Here, $M_q(n) = \{1, \dots, n\}^q$. Note that this equals the RKHS distance between the mean embedding of the empirical distribution and that of the product of the marginal empirical distributions. The following theorem shows that $\widehat{\text{dHSIC}}(D)$ converges in probability to 0 if and only if the null hypothesis is true.

Theorem 1. (Convergence of dHSIC to population value, Pfister et al. 2018.) Assume Setting 1. Then

$$\widehat{\text{dHSIC}}(D) \rightarrow \text{dHSIC}(X^1, \dots, X^d)$$

as $n \rightarrow \infty$ in probability. Furthermore, under these assumptions, the limit $\text{dHSIC}(X^1, \dots, X^d) = 0$ if and only if the null hypothesis, $\mathbb{P}_{X^1, \dots, X^d} = \mathbb{P}_{X^1} \times \dots \times \mathbb{P}_{X^d}$ is true.

In the case where $d = 2$, dHSIC is better known as HSIC which is defined below.

Definition 5. (HSIC, Gretton et al. 2008) The HSIC of random variables $X^1 \in \mathcal{X}^1$ and $X^2 \in \mathcal{X}^2$ is defined as

$$\text{HSIC}(X^1, X^2) := ||\mu_{\mathbb{P}_{X^1, X^2}} - \mu_{\mathbb{P}_{X^1} \times \mathbb{P}_{X^2}}||_{\mathcal{H}_k}^2.$$

In Sejdinovic et al. (2012), HSIC was shown to be equivalent, under a certain choice of kernel, to a generalization of distance covariance (Lyons, 2013), a statistic proposed in Székely and Rizzo (2009). We will mainly prove statements for dHSIC, which then readily carry over to HSIC, as well as to distance covariance, by way of this equivalence.

2.3 | Permutation testing

We follow the notation of Pfister et al. (2018) for the permutation test using dHSIC. We denote by S_n the group of permutations on n elements, which consists of $n!$ distinct permutations. Let $\psi^j \in S_n$ for $j = 2, \dots, d$, and define the permutation vector $\psi = (\psi^2, \dots, \psi^d)$. Then ψ maps \mathcal{X}^n to itself by

$$\psi : \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \begin{pmatrix} x_1^1 & x_1^2 & \dots & x_1^d \\ \vdots & \vdots & & \vdots \\ x_n^1 & x_n^2 & \dots & x_n^d \end{pmatrix} \rightarrow \begin{pmatrix} x_1^\psi \\ \vdots \\ x_n^\psi \end{pmatrix} \begin{pmatrix} x_1^1 & x_{\psi_2(1)}^2 & \dots & x_{\psi_d(1)}^d \\ \vdots & \vdots & & \vdots \\ x_n^1 & x_{\psi_2(n)}^2 & \dots & x_{\psi_d(n)}^d \end{pmatrix}.$$

Note that we keep the first coordinate fixed and permute the remaining $d - 1$ coordinates. Hence, there are $(n!)^{d-1}$ different vectors of such permutations in total. The reason for keeping the first coordinate fixed is that the order in which the n observations appear does not affect HSIC, and thus for a fixed dataset D , the distribution of $\widehat{\text{HSIC}}(\psi D)$ is the same regardless of whether all coordinates are permuted or only the last $d - 1$. We are now ready to define the permutation test.

Definition 6. (Permutation test dHSIC sampling $B \times (d - 1)$ permutations) Let $\alpha \in (0, 1)$. For $b = 1, \dots, B$, let $\psi_b = (\psi_b^1, \dots, \psi_b^{d-1})$ be i.i.d. vectors of $d - 1$ permutations sampled i.i.d. uniformly from S_n . Then let R be the rank of the first entry of the vector

$$\left(\widehat{\text{dHSIC}}(D), \widehat{\text{dHSIC}}(\psi_1 D), \dots, \widehat{\text{dHSIC}}(\psi_B D) \right)$$

when breaking ties at random and where $R = 1$ denotes the rank of the largest elements and $R = B + 1$ of the smallest element. We reject H_0 if $p_B := R/(B + 1) \leq \alpha$. The quantity p_B denotes the p -value of the permutation test enumerating B permutations, and we call α the level of the test.

A well-known property of permutation tests is that they have a valid level for every value of B .

Definition 7. (Valid level) Let $\phi(X_1, \dots, X_n) \in \{0, 1\}$ be a hypothesis test of level α returning 1 if it rejects H_0 , and 0 otherwise. The test ϕ of level α is said to have a valid level if for all distributions $\mathbb{P}_X \in H_0$ (i.e., all distributions \mathbb{P}_X such that $\mathbb{P}_X = \mathbb{P}_{X^1} \times \dots \times \mathbb{P}_{X^d}$), it holds that $\mathbb{P}(\phi(X_1, \dots, X_n) = 1) \leq \alpha$.

Theorem 2. (Valid level permutation test, e.g., Hoeffding, 1952; Pfister et al. 2018) Assume H_0 is true, that is, the $x_1, \dots, x_n \in \mathcal{X}$ are sampled i.i.d. from a distribution $\mathbb{P}_{X^1} \times \dots \times \mathbb{P}_{X^d} \in H_0$. Then for the permutation test of Definition 6, it holds that

$$\mathbb{P}(p_B \leq \alpha) = \frac{\lfloor \alpha(B + 1) \rfloor}{B + 1} \leq \alpha.$$

3 | CONSISTENCY OF THE PERMUTATION TEST FOR HSIC AND DHSIC

We now turn to the main topic of this paper, which is proving that the permutation tests of independence using HSIC and dHSIC are consistent against each fixed alternative hypothesis. We first formally define consistency of a test. Let $\phi(X_1, \dots, X_n) = 1$ if the null hypothesis is rejected and $\phi(X_1, \dots, X_n) = 0$ otherwise.

Definition 8. (Consistency for each alternative) The test ϕ is called consistent for each alternative if for all distributions $\mathbb{P}_X \in H_1$ (i.e., for all distributions \mathbb{P}_X such that $\mathbb{P}_X \neq \mathbb{P}_{X^1} \times \dots \times \mathbb{P}_{X^d}$), it holds that

$$\lim_{n \rightarrow \infty} \mathbb{P}_X(\phi(X_1, \dots, X_n) = 1) = 1.$$

That is, a test is consistent for each alternative if for any fixed alternative hypothesis, the rejection rate converges to 1 as the sample size grows to infinity.

We begin by proving the empirical dHSIC of a randomly permuted sample converges to zero in probability.

Lemma 1. Let $\psi = (\psi^1, \dots, \psi^{d-1})$ be a vector of i.i.d. random permutations. Then

$$\widehat{\text{dHSIC}}(\psi D) \rightarrow 0$$

in probability.

Proof. We note that since $\widehat{\text{dHSIC}}$ is nonnegative it suffices to show that

$$\lim_{n \rightarrow \infty} \mathbb{E} \left[\widehat{\text{dHSIC}}(\psi D) \right] = 0.$$

For the context of this proof only, we slightly change the notation of the permutation vector. Recall that, since we only permute the coordinates $j = 2, \dots, d$, the permutation vector equals $\psi = (\psi^2, \dots, \psi^d)$. For convenience, we add to the notation that we apply the permutation $\psi^1 = id$ to the first coordinate. So we now define $\psi = (\psi^1, \dots, \psi^d)$, where $\psi^1 = id$ and ψ^2, \dots, ψ^d are sampled i.i.d from all permutations on n -elements. The statistic $\widehat{\text{dHSIC}}(\psi D)$ can then be seen to be

$$\begin{aligned} \widehat{\text{dHSIC}}(\psi D) = & \frac{1}{n^2} \sum_{M_2(n)} \prod_{j=1}^d k^j(X_{\psi^j(i_1)}^j, X_{\psi^j(i_2)}^j) + \frac{1}{n^{2d}} \sum_{M_{2d}(n)} \prod_{j=1}^d k^j(X_{\psi^j(i_{2j-1})}^j, X_{\psi^j(i_{2j})}^j) \\ & - \frac{2}{n^{d+1}} \sum_{M_{d+1}(n)} \prod_{j=1}^d k^j(X_{\psi^j(i_1)}^j, X_{\psi^j(i_{j+1})}^j), \end{aligned}$$

which we abbreviate as $A_n + B_n - 2C_n$. We now aim to show that $\lim_n \mathbb{E}A_n = \lim_n \mathbb{E}B_n = \lim_n \mathbb{E}C_n = \zeta$ where $\zeta = \prod_{j=1}^d \mathbb{E}(k^j(X^j, \tilde{X}^j))$ with X^j and \tilde{X}^j denoting independent copies of the same random variable with law \mathbb{P}_{X^j} for $j = 1, \dots, d$.

We begin proving that $\mathbb{E}A_n \rightarrow \zeta$. First define the following random (since they depend on ψ) subsets of $M_2(n) = \{1, \dots, n\}^2$:

$$U(2, \psi, n) := \{(i_1, i_2) \in M_2(n) : (\psi^j(i_1), \psi^j(i_2)) : j = 1, \dots, d \text{ are } 2 \times d \text{ distinct elements}\},$$

and $R(2, \psi, n) := M_2(n) \setminus U(2, \psi, n)$. We now show that, as the sample size increases, the expected value of the portion of tuples of $M_2(n)$ that is in $U(2, \psi, n)$ tends to 1. Note first that $n(n-1)$ is the number of pairs $(i_1, i_2) \in M_2(n)$ so that $i_1 \neq i_2$. Next assume that $i_1 \neq i_2$ and that $(\psi^j(i_1), \psi^j(i_2)) : j = 1, \dots, J-1$ are $2(J-1)$ distinct elements. Under that assumption, the fraction $\binom{n-2(J-1)}{2} / \binom{n}{2}$ corresponds with the probability that $\psi^J(i_1), \psi^J(i_2) \notin \{1, \dots, n\} \setminus \{\psi^j(i_1), \psi^j(i_2) : j = 1, \dots, J-1\}$. Repeating that argument, we obtain

$$\begin{aligned} \frac{\mathbb{E}|U(2, \psi, n)|}{n^2} &= \frac{n(n-1)}{n^2} \times \mathbb{P}((\psi^j(1), \psi^j(2)) : j = 1, \dots, d \text{ are } 2 \times d \text{ distinct elements}) \\ &= \frac{n(n-1)}{n^2} \frac{\binom{n-2}{2} \binom{n-4}{2} \dots \binom{n-2(d-1)}{2}}{\binom{n}{2} \binom{n}{2} \dots \binom{n}{2}} \\ &\rightarrow 1. \end{aligned}$$

As a result, it follows that $\mathbb{E}|R(2, \psi, n)|/n^2 \rightarrow 0$. Conditioning on ψ and using the tower property, we find

$$\begin{aligned} \mathbb{E}(A_n) &= \mathbb{E}(\mathbb{E}(A_n | \psi)) \\ &= \mathbb{E}\left(\frac{1}{n^2} \sum_{U(2, \psi, n)} \mathbb{E}\left(\prod_{j=1}^d k^j(X_{\psi^j(i_1)}^j, X_{\psi^j(i_2)}^j) \middle| \psi\right)\right) \\ &\quad + \mathbb{E}\left(\frac{1}{n^2} \sum_{R(2, \psi, n)} \mathbb{E}\left(\prod_{j=1}^d k^j(X_{\psi^j(i_1)}^j, X_{\psi^j(i_2)}^j) \middle| \psi\right)\right) \\ &= \mathbb{E}\left(\frac{|U(2, \psi, n)|}{n^2}\right) \prod_{j=1}^d \mathbb{E}(k^j(X^j, \tilde{X}^j)) \\ &\quad + \mathbb{E}\left(\frac{|R(2, \psi, n)|}{n^2}\right) \mathcal{O}(1) \\ &\rightarrow \zeta. \end{aligned}$$

Note that in the last equality, we use that in the first sum all indices in the product are distinct and the expectation factorizes, and in the second sum, the estimate $\mathcal{O}(1)$ follows from the fact that in Setting 1 the kernels are bounded.

Repeating this argument for B_n , we define

$$U(2d, \psi, n) := \left\{ (i_1, \dots, i_{2d}) \in M_{2d}(n) : (\psi^j(i_{2j-1}), \psi^j(i_{2j})) : j = 1, \dots, d \text{ are } 2 \times d \text{ distinct elements} \right\}$$

and let $R(2d, \psi, n) := M_{2d}(n) \setminus U(2d, \psi, n)$. Note that there are $n^d(n-1)^d$ index vectors $(i_1, \dots, i_{2d}) \in M_{2d}(n)$ so that $i_{2j-1} \neq i_{2j}$ for $j = 1, \dots, d$. Combined with the same computation as for A_n , we find that

$$\begin{aligned} \frac{\mathbb{E}(|U(2d, \psi, n)|)}{n^{2d}} &= \frac{n^d(n-1)^d}{n^{2d}} \frac{\binom{n-2}{2} \binom{n-4}{2} \dots \binom{n-2(d-1)}{2}}{\binom{n}{2} \binom{n}{2} \dots \binom{n}{2}} \\ &\rightarrow 1. \end{aligned}$$

Proceeding similarly as we did for A_n ,

$$\begin{aligned}\mathbb{E}(B_n) &= \mathbb{E}\left(\frac{|U(2d, \psi, n)|}{n^{2d}}\right) \prod_{j=1}^d \mathbb{E}(k^j(X^j, \tilde{X}^j)) \\ &\quad + \mathbb{E}\left(\frac{|R(2d, \psi, n)|}{n^{2d}}\right) \mathcal{O}(1) \\ &\rightarrow \zeta.\end{aligned}$$

Lastly, turning our attention to C_n , we define

$$U(d+1, \psi, n) := \left\{ (i_1, \dots, i_{d+1}) \in M_{d+1}(n) : (\psi^j(i_1), \psi^j(i_{j+1})) : j = 1, \dots, d \text{ are } 2 \times d \text{ distinct elements} \right\}$$

while defining $R(d+1, \psi, n) = M_{d+1}(n) \setminus U(d+1, \psi, n)$. Note that there are $n(n-1)^d$ tuples (i_1, \dots, i_{d+1}) of $M_{d+1}(n)$ for which $i_j \neq i_1$ for all $j = 2, \dots, d+1$. Hence,

$$\begin{aligned}\frac{|U(d+1, \psi, n)|}{n^{d+1}} &= \frac{n(n-1)^d}{n^{d+1}} \frac{\binom{n-2}{2} \binom{n-4}{2} \dots \binom{n-2(d-1)}{2}}{\binom{n}{2} \dots \binom{n}{2}} \\ &\rightarrow 1.\end{aligned}$$

Applying the same argument as before to $\mathbb{E}(C_n)$, we find that $\lim_n(C_n) \rightarrow \zeta$. It follows that $\mathbb{E}[\widehat{\text{dHSIC}}(\psi D)] = A_n + B_n - 2C_n \rightarrow 0$ and hence $\widehat{\text{dHSIC}}(\psi D) \rightarrow 0$ in probability. \square

It is now straightforward to prove the main result of this paper.

Theorem 3. (Consistency using a finite sample of permutations) Assume we are in Setting 1 with \mathbb{P}_X a distribution such that $\mathbb{P}_X \neq \mathbb{P}_{X^1} \times \dots \times \mathbb{P}_{X^d}$. Perform a permutation test on a sample of size n using B random permutation vectors ψ_1, \dots, ψ_B of length $d-1$, where $B \geq \frac{1}{\alpha} - 1$ for $\alpha \in (0, 1)$. Then

$$\lim_{n \rightarrow \infty} \mathbb{P}(p_B \leq \alpha) = 1.$$

Proof. Define $C := \text{dHSIC}(X^1, \dots, X^d) > 0$. Then

$$\begin{aligned}\mathbb{P}(p_B \leq \alpha) &\geq \mathbb{P}(\widehat{\text{dHSIC}}(D) > C/2, \widehat{\text{dHSIC}}(\psi_b D) < C/2 \text{ for } b = 1, \dots, B) \\ &\geq 1 - \left(\mathbb{P}(\widehat{\text{dHSIC}}(D) < C/2) + B \times \mathbb{P}(\widehat{\text{dHSIC}}(\psi_1 D) > C/2) \right) \\ &\rightarrow 1,\end{aligned}$$

as $n \rightarrow \infty$, which proves the theorem. The fact that the two probabilities in the last equation converge to 0 follows from Lemma 1 and the convergence of $\widehat{\text{dHSIC}}(D) \rightarrow C$ in probability. \square

3.1 | Comments on the proof strategy

Contrary to the proof above, the proof strategy proposed in Pfister et al. (2018) requires extending a proof of Romano (1989), which studies statistics of the form $\sqrt{n} \delta_{\mathcal{V}}(\mathbb{P}_X, \mathbb{P}_{X^1} \times \dots \times \mathbb{P}_{X^d})$ where $\delta_{\mathcal{V}}(P, Q) = \sup\{|P(V) - Q(V)| : V \in \mathcal{V}\}$ and \mathcal{V} can be an arbitrary Vapnik–Chervonenkis (VC) class. Since the VC class is arbitrary, there is no closed form expression for this statistic. While the proof of Romano (1989) may be extended from indicators of sets in a VC class, to functions in an RKHS, the fact that we have an analytic closed form of HSIC enables the simpler approach given in this paper.

In fact, Romano (1989) built on previous work by Hoeffding (1952), which describes a framework for asymptotic power of permutation tests. That work considers two conditions that together characterize asymptotic power of the permutation test. Translated to our setting, Condition 1 is that the random quantile $Q^{n,B}(\alpha) := \inf \left\{ t : 1 + \sum_{b=1}^B 1 \left\{ \widehat{\text{dHSIC}}(\psi_b D) \leq t \right\} \leq \alpha(B+1) \right\} \rightarrow \lambda$ in probability as $n \rightarrow \infty$, for some constant $\lambda \in \mathbb{R}$. Condition 2 is that there exists a function $H : \mathbb{R} \rightarrow [0, 1]$, continuous at λ , so that $\mathbb{P}(\widehat{\text{dHSIC}}(D) \leq t) \rightarrow H(t)$ for each t at which H is continuous, as $n \rightarrow \infty$. Under these conditions, it can be seen that the power $\mathbb{P}(p_B \leq \alpha) \geq \mathbb{P}(\widehat{\text{dHSIC}}(D) > Q^{n,B}(\alpha)) \rightarrow 1 - H(\lambda)$ (see Hoeffding, 1952). For simplicity, we chose not to introduce these conditions in the proof of Theorem 3. However, in the case of $\widehat{\text{dHSIC}}$, it is easy to see that Lemma 1 implies that Condition 1 is met with $Q_{n,B}(\alpha) \rightarrow 0$ (so that $\lambda = 0$ in Condition 1). Furthermore, using the convergence of $\widehat{\text{dHSIC}}(D)$ to its population limit, it is straightforward to verify that Condition 2 is met with the step-function $H(t) = 1 \{ \text{dHSIC}(X^1, \dots, X^d) \leq t \}$. Since for this H , it holds that $H(\lambda) = H(0) = 0$, corresponding to the fact that the test is consistent.

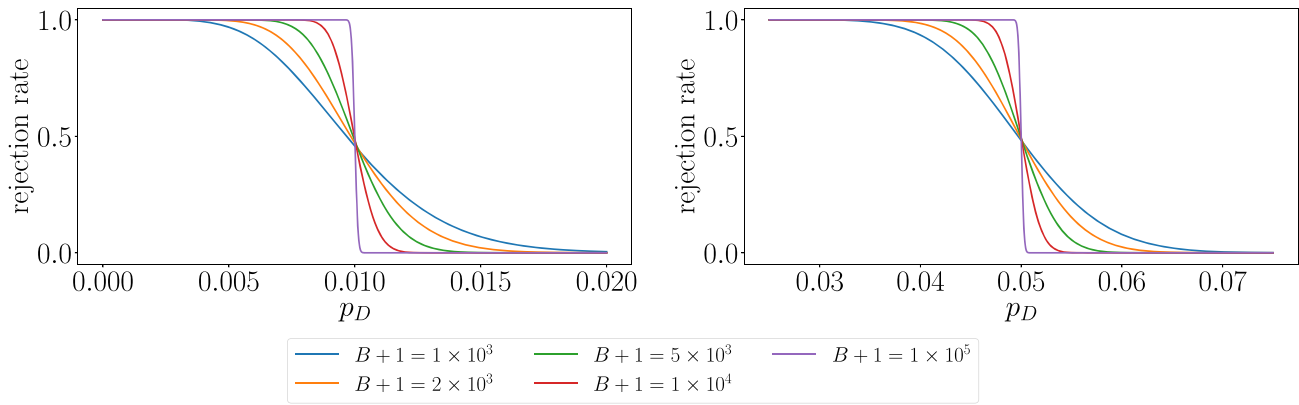


FIGURE 1 Say you are given (fixed) data D with associated p_D . When you enumerate all permutations, you reject if and only if $q \leq \alpha$. These plots plot $\mathbb{P}(p_B^{\text{cons}} \leq \alpha | p_D)$ for different values of B . In the left plot, $\alpha = 0.010$, and in the right plot, $\alpha = 0.05$

3.2 | Enumerating all permutations

Instead of sampling $B \times (d - 1)$ permutations, one may also consider a permutation test enumerating all permutation vectors. There are $(n!)^{d-1} - 1$ permutation vectors of length $d - 1$ so that not all coordinates are the identity permutation. One then computes the rank of $\widehat{\text{dHSIC}}(D)$ in the vector $(\widehat{\text{dHSIC}}(D), \widehat{\text{dHSIC}}(\psi_1 D), \dots, \widehat{\text{dHSIC}}(\psi_{(n!)^{d-1}-1} D))$, breaking ties at random and rejects H_0 if the resulting rank is not greater than $\alpha(n!)^{d-1}$. The resulting test has a valid level and is consistent, so it has the same guarantees as the test sampling a finite number of permutations. In practice, enumerating all possible permutation vectors is almost always unfeasible. See Hoeffding (1952) for a proof of the valid level of the test enumerating all permutations, and see Appendix C1 for a proof of its consistency.

4 | INFLUENCE OF THE NUMBER OF PERMUTATIONS ON THE POWER OF A PERMUTATION TEST

The previous results imply that if the kernel k is characteristic, then for every B so that $B \geq \frac{1}{\alpha(B+1)} - 1$ the dHSIC permutation test with B permutation vectors of length $d - 1$ is consistent and has a valid level. Since the permutation test has both these properties for each number of permutations, it is natural to ask how to select the number of permutations and how the number of permutations influences the power of the test. That is the topic of this section.

To answer these questions, we consider the permutation test of Definition 6, but instead break ties conservatively, by which we mean the following. Let ψ_b be i.i.d permutation vectors for $b = 1, \dots, B$. Then define R^{cons} by $R^{\text{cons}} := 1 + \sum_{b=1}^B \mathbf{1}\{\widehat{\text{dHSIC}}(\psi_b D) \geq \widehat{\text{dHSIC}}(D)\}$. Defining $p_B^{\text{cons}} := R^{\text{cons}}/(B+1)$, we see that $p_B^{\text{cons}} = \frac{Z+1}{B+1}$ where Z is a binomial random variable with success probability $p_D := \mathbb{P}(\widehat{\text{dHSIC}}(\psi D) \geq \widehat{\text{dHSIC}}(D) | D)$. Note, in particular, that $p_B^{\text{cons}} \rightarrow p_D$ as $B \rightarrow \infty$. In the case of ties between statistics on permuted data and the statistic on the original data, the test breaking ties conservatively are more conservative than the test breaking ties at random. However, when the data come from a continuous distribution and the dataset is large, typically few ties occur and the two tests perform similarly. In particular, Theorems 3 and 2 hold also when the ties are broken conservatively. We study the test breaking ties conservatively, since it simplifies the following sections.

4.1 | The effect of number of the permutations on rejection probabilities for a fixed dataset

In this section, we consider a fixed dataset D with an associated quantity p_D and study the effect of the number of permutations on the probability of rejecting H_0 . The rejection probabilities for a range of values of B for different values of p_D is given in Figure 1. We distinguish three cases based on the value of $p_D := \mathbb{P}(\widehat{\text{dHSIC}}(\psi D) \geq \widehat{\text{dHSIC}}(D) | D)$ of the given dataset D .

1. The data are such that $p_D < \alpha$: In this case, increasing the number of permutations increases the probability of rejecting the null hypothesis for this dataset.
2. The data are such that $p_D = \alpha$: In this case, the probability of rejecting the null hypothesis is approximately 1/2, since the mean and median of a binomial distribution are very close.
3. The data are such that $p_D > \alpha$: In this case, as the number of permutations increases, p_B^{cons} gets more and more likely to be close to p_D , and we are thus less likely to reject the null hypothesis. So using fewer permutations actually raises the probability of rejecting the null hypothesis based on this dataset.

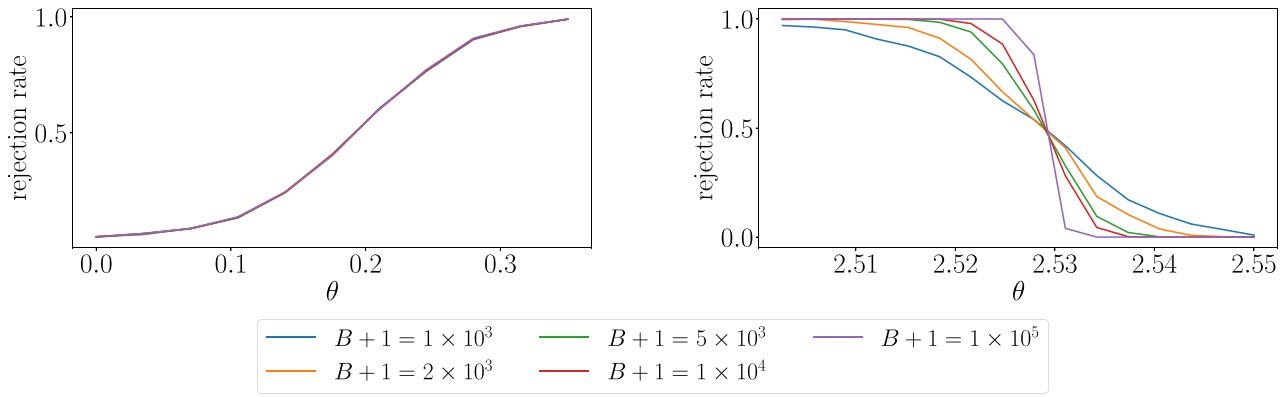


FIGURE 2 Left: The rejection rates of the permutation test with different values of B in Scenario 1. Using more permutations makes little difference to the power of test. Right: The rejection rates of the permutation test with different values of B in Scenario 2. Using more permutations makes one virtually always reject datasets with lower θ and virtually never with higher θ

4.2 | The effect of the number of permutations on the power

While the previous subsection studied the rejection rate for a fixed dataset, we now study the rejection rate for a fixed alternative hypothesis. Note the parameter p_D itself is a random quantity too (as it is a function of the data). Say, for the sake of simplicity, that for a fixed alternative hypothesis the quantity p_D has a density $f_{p_D}(p)$ on $[0, 1]$. The total probability of rejecting the null hypothesis when using a permutation test with B permutation vectors is

$$\mathbb{P}(p_B^{\text{cons}} \leq \alpha) = \int_{[0,1]} \mathbb{P}(Z \leq \alpha(B+1) - 1 | p_D = p) f_{p_D}(p) dp.$$

While we plotted the function $\mathbb{P}(Z \leq \alpha(B+1) - 1 | p_D = p)$ in Figure 1, the quantity $f_{p_D}(p)$ will depend on the data generating mechanism. It will be approximately uniform under the null hypothesis, but analytic descriptions of f_{p_D} are complicated for arbitrary distributions of X . We perform two simulation studies to illustrate the relationship between power and B . We study the case where $d = 2$ and $\alpha = 0.05$ and where X^1 and X^2 are five-dimensional random variables.

Scenario 1: Let

$$X^2 = \theta X^1 + \epsilon,$$

where $X^1, \epsilon \sim \mathcal{N}(0, I_5)$ independently where I_5 is the five-dimensional identity matrix. When $n = 100$, we find that the power is nearly identical for all numbers of permutations. This holds for all values of θ , as shown in Figure 2. An explanation is that the variance in the underlying p -value p_D is large when the sample size is 100, and as a result, f_{p_D} has a wide support, and the integral of the functions plotted in Figure 1 with density f_{p_D} all result in the same value.

Scenario 2: The previous scenario showed no difference in power between tests with different numbers of permutations. The explanation was the variance of p_D , or the width of the distribution f_{p_D} . It is not easy to find distributions of X^1 and X^2 such that f_{p_D} has a support only in the region where the curves in Figure 1 are separated (so near α) and this support is furthermore not symmetric around α . So in Scenario 2, for each value of θ , we choose a fixed (non-random) dataset. Namely,

$$X^2 = \sin(\theta X^1),$$

where $X_i^1 = i2\pi/100$ for $i = 1, \dots, 100$ (so again $n = 100$). In this case, as θ increases, the frequency of the oscillation increases and the sample looks less dependent. Indeed in Figure 2, we see that slow oscillations are more often rejected when using more permutations than when using fewer permutations. Similarly, fast oscillations are more often rejected when using fewer permutations than when using more permutations.

4.3 | How many permutations to use?

In practice, there is thus no clear guarantee that using more permutations leads to better power. But how does one then decide which value of B to use? There is no definitive answer, but we suggest here some relevant considerations.

In practice, one is not only interested in accepting or rejecting the null hypothesis at a certain significance level, but one wants to find a reliable estimate of p_D . So we recommend to choose B so large that p_B is likely to be sufficiently close to p_D for the problem at hand. We recall that, given the dataset, $p_B^{\text{cons}} = (Z + 1)/(B + 1)$ where $Z \sim \text{Binom}(B, p_D)$. Hence, accuracy of p_B^{cons} can be described simply through confidence intervals of the binomial distribution. That is,

$$\begin{aligned} \mathbb{P}(p_B^{\text{cons}} - \epsilon \leq p_D \leq p_B^{\text{cons}} + \epsilon) \\ = \mathbb{P}((p_D - \epsilon)(B + 1) - 1 \leq Z \leq (p_D + \epsilon)(B + 1) - 1). \end{aligned}$$

For a given ϵ , p_D , and a confidence level $1 - \lambda$, we can find B such that $p_B^{\text{cons}} \pm \epsilon$ is a confidence interval of level $1 - \lambda$. As we of course do not know p_D , we could choose B so large that for any p_D , the interval $p_B^{\text{cons}} \pm \epsilon$ is a $(1 - \lambda)$ -confidence interval. However, B will always be the highest for $p_D = 0.5$ as that value maximizes the variance of the binomial distribution, which means one requires very large numbers of permutations.

Noting that the accuracy of p_B is more important when p_D is small, and in particular not much larger than α , we propose a strategy to decide whether a value of B is sufficiently large, illustrating it on the example of $\alpha = 0.05$, $B = 2.3 \times 10^4$, and $C = 0.1$. We begin by selecting a $C \in (\alpha, 1)$ and consider the largest width of a 99% confidence interval in the case $p_D \in [0, C]$, and in the case $p_D \in [C, 1]$. In our example, the former occurs when $p_D = C$, and the resulting confidence interval is $p_B^{\text{cons}} \pm 0.005$, and the second is $p_B^{\text{cons}} \pm 0.01$, corresponding to $p_D = 0.5$. If the accuracy in both of these two cases is sufficient for the purpose at hand, we perform the test with B permutations; otherwise, we need to increase the number of permutations.

5 | CONCLUSION

We have studied kernel measures of dependence and how they are combined with permutation tests to perform hypothesis testing. Our main contribution is proving the consistency of the permutation test with statistic dHSIC with a characteristic kernel. Applying this result to the case $d = 2$, it follows in particular that the permutation tests of independence with the test statistics HSIC and distance covariance are also consistent for each alternative hypothesis. We further gave examples of how one may go about choosing a number of permutations in practice.

ACKNOWLEDGEMENT

David Rindt is supported by the UK Engineering and Physical Sciences Research Council (EPSRC).

ORCID

David Rindt  <https://orcid.org/0000-0002-5671-3910>

REFERENCES

- Berrett, T. B., Kontoyiannis, I., & Samworth, R. J. (2020). Optimal rates for independence testing via u -statistic permutation tests. *arXiv preprint arXiv:2001.05513*.
- Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 12, 723–773.
- Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., & Smola, A. (2008). A kernel statistical test of independence. *Advances in Neural Information Processing Systems*, 585–592.
- Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics*, 169–192.
- Kim, I., Balakrishnan, S., & Wasserman, L. (2020). Minimax optimality of permutation tests. *arXiv preprint arXiv:2003.13208*.
- Lyons, R. (2013). Distance covariance in metric spaces. *The Annals of Probability*, 41(5), 3284–3305.
- Pfister, N., Bühlmann, P., Schölkopf, B., & Peters, J. (2018). Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1), 5–31.
- Romano, J. P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *The Annals of Statistics*, 141–159.
- Schölkopf, B., & Smola, A. (2001). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. (1st ed.). Massachusetts: MIT press.
- Sejdinovic, D., Sriperumbudur, B., Gretton, A., & Fukumizu, K. (2012). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41, 2263–2291.
- Sriperumbudur, B. K., Fukumizu, K., & Lanckriet, G. R. G. (2011). Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(Jul), 2389–2410.
- Szabó, Z., & Sriperumbudur, B. K. (2017). Characteristic and universal tensor product kernels. *The Journal of Machine Learning Research*, 18(1), 8724–8752.
- Szekély, G., & Rizzo, M. (2009). Brownian distance covariance. *The Annals of Applied Statistics*, 3, 1236–1265.

How to cite this article: Rindt D, Sejdinovic D, Steinsaltz D. Consistency of permutation tests of independence using distance covariance, HSIC and dHSIC. *Stat*. 2021;10:e364. <https://doi.org/10.1002/sta4.364>

APPENDIX A

A CONSISTENCY OF THE TEST ENUMERATING ALL PERMUTATIONS

Theorem 4. (Consistency using all permutations) Let \mathbb{P}_X be any distribution such that $\mathbb{P}_X \neq \mathbb{P}_{X^1} \times \dots \times \mathbb{P}_{X^d}$. Perform the permutation test described in Section 3.2 and let p denote the resulting p -value. Then

$$\lim_{n \rightarrow \infty} \mathbb{P}(p \leq \alpha) = 1.$$

Proof. Note that $p \leq \mathbb{P} \left(\widehat{\text{dHSIC}}(\psi D) \geq \widehat{\text{dHSIC}}(D) | D \right)$, which would be the p -value when ties are broken conservatively. By Markov's inequality,

$$\begin{aligned} \mathbb{P}(p \geq \alpha) &\leq \mathbb{P} \left(\mathbb{P} \left(\widehat{\text{dHSIC}}(\psi D) \geq \widehat{\text{dHSIC}}(D) | D \right) \geq \alpha \right) \\ &\leq \frac{\mathbb{E} \left[\mathbb{P} \left(\widehat{\text{dHSIC}}(\psi D) \geq \widehat{\text{dHSIC}}(D) | D \right) \right]}{\alpha} \\ &= \frac{\mathbb{P} \left(\widehat{\text{dHSIC}}(\psi D) \geq \widehat{\text{dHSIC}}(D) \right)}{\alpha} \\ &\rightarrow 0 \end{aligned}$$

by Theorem 1 and the convergence of $\widehat{\text{dHSIC}}(D)$. □